MANUAL

INL/EXT-15-34123
Revision 6 draft
Printed March 2017

RAVEN User Manual

Cristian Rabiti, Andrea Alfonsi, Joshua Cogliati, Diego Mandelli, Robert Kinoshita, Sonat Sen, Congjian Wang, Paul W. Talbot, Daniel P. Maljovec, Jun Chen

Prepared by
Idaho National Laboratory
Idaho Falls, Idaho 83415

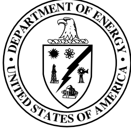
The Idaho National Laboratory is a multiprogram laboratory operated by
Battelle Energy Alliance for the United States Department of Energy
under DOE Idaho Operations Office. Contract DE-AC07-05ID14517.

Approved for unlimited release.

Issued by the Idaho National Laboratory, operated for the United States Department of Energy by Battelle Energy Alliance.

NOTICE: This report was prepared as an account of work sponsored by an agency of the United States Government. Neither the United States Government, nor any agency thereof, nor any of their employees, nor any of their contractors, subcontractors, or their employees, make any warranty, express or implied, or assume any legal liability or responsibility for the accuracy, completeness, or usefulness of any information, apparatus, product, or process disclosed, or represent that its use would not infringe privately owned rights. Reference herein to any specific commercial product, process, or service by trade name, trademark, manufacturer, or otherwise, does not necessarily constitute or imply its endorsement, recommendation, or favoring by the United States Government, any agency thereof, or any of their contractors or subcontractors. The views and opinions expressed herein do not necessarily state or reflect those of the United States Government, any agency thereof, or any of their contractors.

Printed in the United States of America. This report has been reproduced directly from the best available copy.



INL/EXT-15-34123
Revision 6 draft
Printed March 2017

RAVEN User Manual

Project Manager:
Cristian Rabiti
Principal Investigator and Technical Leader:
Andrea Alfonsi
Main Developers:
Andrea Alfonsi
Diego Mandelli
Joshua Cogliati
Congjian Wang
Paul W. Talbot
Daniel P. Maljovec
Robert Kinoshita
Former Developers:
Sonat Sen
Jun Chen
Contributors:
Alessandro Bandini (Post-Processor)
Ivan Rinaldi (documentation)
Claudia Picoco (new external code interface)
James B. Tompkins (new external code interface)
Matteo Donorio (new external code interface)
Fabio Giannetti (new external code interface)

Contents

1 Introduction . . . 15
2 Manual Formats . . . 16
3 Installation Overview . . . 17
4 RAVEN Dependencies Installation . . . 18
4.1 Preparing a Linux System for RAVEN . . . 18
4.1.1 Ubuntu . . . 19
4.1.1.1 Advanced Package Tool (APT) . . . 19
4.1.1.2 Miniconda . . . 19
4.1.1.3 Optional LateX installation . . . 19
4.1.2 Fedora . . . 20
4.1.2.1 Red Hat Package Manager (RPM) . . . 20
4.1.2.2 Miniconda . . . 20
4.1.2.3 Optional LateX installation . . . 20
4.2 Preparing an Apple Macintosh OSX System for RAVEN . . . 21
4.2.1 Installing XCode Command Line Tools . . . 21
4.2.2 Installing XQuartz . . . 21
4.2.3 Install RAVEN libraries . . . 21
4.3 Preparing a Windows System for RAVEN . . . 22
4.3.1 Prerequisites to use RAVEN on Windows . . . 22
4.3.2 Installation and Configuration of the MSYS2 environment . . . 22
4.3.3 Install Python Language and Package Support . . . 23
4.3.4 Compiler Installation and Configuration . . . 25
4.3.5 For More Information . . . 27
4.4 Manual Dependency Install . . . 27
4.5 How RAVEN finds Dependencies . . . 27
5 RAVEN Installation . . . 28
5.1 Submodule Git . . . 28
5.2 RAVEN Source Code Package . . . 28
5.3 RAVEN Compilation . . . 29
5.4 Troubleshooting the Installation . . . 29
5.5 In-use Testing . . . 30
6 Running RAVEN . . . 31
7 Raven Input Structure . . . 32
7.1 Comments . . . 32
7.2 Verbosity . . . 33
7.3 External Input Files . . . 34
8 RunInfo . . . 35
8.1 RunInfo: Input of Calculation Flow . . . 35
8.2 RunInfo: Input of Queue Modes . . . 38
8.3 RunInfo: Example Cluster Usage . . . 40

8.4 RunInfo: Advanced Users . . . 41
8.5 RunInfo: Examples . . . 44
9 Files . . . 45
10 VariableGroups . . . 46
11 Distributions . . . 48
11.1 1-Dimensional Probability Distributions . . . 48
11.1.1 1-Dimensional Continuous Distributions . . . 48
11.1.1.1 Beta Distribution . . . 49
11.1.1.2 Exponential Distribution . . . 50
11.1.1.3 Gamma Distribution . . . 51
11.1.1.4 Laplace Distribution . . . 52
11.1.1.5 Logistic Distribution . . . 52
11.1.1.6 LogNormal Distribution . . . 53
11.1.1.7 Normal Distribution . . . 54
11.1.1.8 Triangular Distribution . . . 55
11.1.1.9 Uniform Distribution . . . 56
11.1.1.10 Weibull Distribution . . . 57
11.1.1.11 Custom1D Distribution . . . 58
11.1.2 1-Dimensional Discrete Distributions. . . . 59
11.1.2.1 Bernoulli Distribution . . . 59
11.1.2.2 Binomial Distribution . . . 60
11.1.2.3 Geometric Distribution . . . 61
11.1.2.4 Poisson Distribution . . . 61
11.1.2.5 Categorical Distribution . . . 62
11.2 N-Dimensional Probability Distributions . . . 63
11.2.1 MultivariateNormal Distribution . . . 64
11.2.2 NDInverseWeight Distribution . . . 66
11.2.3 NDCartesianSpline Distribution . . . 67
12 Samplers . . . 70
12.1 Forward Samplers . . . 73
12.1.1 Monte Carlo . . . 74
12.1.2 Grid . . . 77
12.1.3 Sparse Grid Collocation . . . 82
12.1.4 Sobol . . . 87
12.1.5 Stratified . . . 90
12.1.6 Response Surface Design . . . 96
12.1.7 Factorial Design . . . 101
12.1.8 Ensemble Forward Sampling strategy . . . 106
12.1.9 Custom Sampling strategy . . . 109
12.2 Dynamic Event Tree (DET) Samplers . . . 111
12.2.1 Dynamic Event Tree . . . 111
12.2.2 Hybrid Dynamic Event Tree . . . 114

12.3 Adaptive Samplers 120
12.3.1 Limit Surface Search 121
12.3.2 Adaptive Dynamic Event Tree 126
12.3.3 Adaptive Hybrid Dynamic Event Tree 131
12.3.4 Adaptive Sparse Grid 138
12.3.5 Adaptive Sobol Decomposition 143
13 Optimizers 148
13.1 Gradient Based Optimizers 148
13.1.1 Simultaneous Perturbation Stochastic Approximation (SPSA) 150
13.1.2 Finite Difference Gradient Optimizer (FiniteDifferenceGradientOptimizer) 157
14 DataObjects 165
15 Databases 168
16 OutStream system 170
16.1 Printing system 170
16.1.1 **DataObjects** Printing 171
16.1.2 **ROM** Printing 172
16.2 Plotting system 173
16.2.1 Plot input structure 173
16.2.1.1 “Actions” input block 174
16.2.1.2 “plotSettings” input block 180
16.2.1.2.1 Specifying What Values to Plot 184
16.2.1.3 Predefined Plotting System: 2D/3D 184
16.2.2 2D & 3D Scatter plot 185
16.2.3 2D & 3D Line plot 186
16.2.4 2D & 3D Histogram plot 187
16.2.5 2D & 3D Stem plot 189
16.2.6 2D Step plot 189
16.2.7 2D Pseudocolor plot 190
16.2.8 2D Contour or filledContour plots 191
16.2.9 3D Surface Plot 191
16.2.10 3D Wireframe Plot 192
16.2.11 3D Tri-surface Plot 193
16.2.12 3D Contour or filledContour plots 194
16.2.13 DataMining plots 195
16.2.14 Example XML input 196
17 Models 198
17.1 Code 199
17.2 Dummy 201
17.3 ROM 203
17.3.1 NDspline 204
17.3.2 pickledROM 205
17.3.3 GaussPolynomialRom 206

17.3.4 HDMRRom . . . 209
17.3.5 MSR . . . 211
17.3.6 NDinvDistWeight . . . 214
17.3.7 SciKitLearn . . . 215
17.3.7.1 Linear Models . . . 215
17.3.7.1.1 Linear Model: Automatic Relevance Determination Regression . . . 215
17.3.7.1.2 Linear Model: Bayesian ridge regression . . . 217
17.3.7.1.3 Linear Model: Elastic Net . . . 218
17.3.7.1.4 Linear Model: Elastic Net CV . . . 219
17.3.7.1.5 Linear Model: Least Angle Regression model . . . 220
17.3.7.1.6 Linear Model: Cross-validated Least Angle Regression model . . . 221
17.3.7.1.7 Linear Model trained with L1 prior as regularizer (aka the Lasso) . . . 222
17.3.7.1.8 Lasso linear model with iterative fitting along a regularization path . . . 223
17.3.7.1.9 Lasso model fit with Least Angle Regression . . . 224
17.3.7.1.10 Cross-validated Lasso, using the LARS algorithm . . . 225
17.3.7.1.11 Lasso model fit with Lars using BIC or AIC for model selection . . . 226
17.3.7.1.12 Ordinary least squares Linear Regression . . . 228
17.3.7.1.13 Logistic Regression . . . 228
17.3.7.1.14 Multi-task Lasso model trained with L1/L2 mixed-norm as regularizer . . . 229
17.3.7.1.15 Multi-task Elastic Net model trained with L1/L2 mixed-norm as regularizer . . . 230
17.3.7.1.16 Orthogonal Mathching Pursuit model (OMP) . . . 231
17.3.7.1.17 Cross-validated Orthogonal Mathching Pursuit model (OMP) . . . 232
17.3.7.1.18 Passive Aggressive Classifier . . . 233
17.3.7.1.19 Passive Aggressive Regressor . . . 234
17.3.7.1.20 Perceptron . . . 236
17.3.7.1.21 Linear least squares with l2 regularization . . . 237
17.3.7.1.22 Classifier using Ridge regression . . . 238
17.3.7.1.23 Ridge classifier with built-in cross-validation . . . 239
17.3.7.1.24 Ridge regression with built-in cross-validation . . . 240
17.3.7.1.25 Linear classifiers (SVM, logistic regression, a.o.) with SGD training . . . 242
17.3.7.1.26 Linear model fitted by minimizing a regularized empirical loss with SGD . . . 244
17.3.7.2 Support Vector Machines . . . 246

17.3.7.2.1 Linear Support Vector Classifier 247
17.3.7.2.2 C-Support Vector Classification 248
17.3.7.2.3 Nu-Support Vector Classification 250
17.3.7.2.4 Support Vector Regression 251
17.3.7.3 Multi Class 252
17.3.7.3.1 One-vs-the-rest (OvR) multiclass/multilabel strategy . . 253
17.3.7.3.2 One-vs-one multiclass strategy 253
17.3.7.3.3 Error-Correcting Output-Code multiclass strategy 254
17.3.7.4 Naive Bayes 254
17.3.7.4.1 Gaussian Naive Bayes 255
17.3.7.4.2 Multinomial Naive Bayes 256
17.3.7.4.3 Bernoulli Naive Bayes 256
17.3.7.5 Neighbors 257
17.3.7.5.1 K Neighbors Classifier 258
17.3.7.5.2 Radius Neighbors Classifier 259
17.3.7.5.3 K Neighbors Regressor 261
17.3.7.5.4 Radius Neighbors Regressor 262
17.3.7.5.5 Nearest Centroid Classifier 263
17.3.7.6 Tree 264
17.3.7.6.1 Decision Tree Classifier 265
17.3.7.6.2 Decision Tree Regressor 266
17.3.7.6.3 Extra Tree Classifier 268
17.3.7.6.4 Extra Tree Regressor 269
17.3.7.7 Gaussian Process 270
17.3.8 ARMA 273
17.4 External Model 275
17.4.1 Method: `def _readMoreXML` 277
17.4.2 Method: `def initialize` 278
17.4.3 Method: `def createNewInput` 279
17.4.4 Method: `def run` 280
17.5 PostProcessor 280
17.5.1 BasicStatistics 281
17.5.2 ComparisonStatistics 286
17.5.3 ImportanceRank 288
17.5.4 SafestPoint 291
17.5.5 LimitSurface 293
17.5.6 LimitSurfaceIntegral 295
17.5.7 External 297
17.5.8 TopologicalDecomposition 299
17.5.9 DataMining 301
17.5.9.1 SciKitLearn 302
17.5.9.2 Gaussian mixture models 303

17.5.9.2.1 GMM classifier 303
17.5.9.2.2 Variational GMM Classifier (VBGMM) 304
17.5.9.3 Clustering 305
17.5.9.3.1 K-Means Clustering 307
17.5.9.3.2 Mini Batch K-Means 308
17.5.9.3.3 Affinity Propagation 310
17.5.9.3.4 Mean Shift 311
17.5.9.3.5 Spectral clustering 312
17.5.9.3.6 DBSCAN Clustering 314
17.5.9.3.7 Agglomerative Clustering 314
17.5.9.3.8 Clustering performance evaluation 316
17.5.9.4 Decomposing signals in components (matrix factorization problems) 316
17.5.9.4.1 Principal component analysis (PCA) 316
17.5.9.4.2 Truncated singular value decomposition 322
17.5.9.4.3 Fast ICA 322
17.5.9.5 Manifold learning 323
17.5.9.5.1 Isomap 324
17.5.9.5.2 Locally Linear Embedding 325
17.5.9.5.3 Spectral Embedding 327
17.5.9.5.4 Multi-dimensional Scaling (MDS) 328
17.5.9.6 Scipy 329
17.5.10 Interfaced 331
17.5.10.1 Data Format 333
17.5.10.2 Method: HStoPSOperator 333
17.5.10.3 Method: HistorySetSampling 335
17.5.10.4 Method: HistorySetSync 336
17.5.10.5 Method: HistorySetSnapShot 337
17.5.10.6 Method: HSPS 339
17.5.10.7 Method: TypicalHistoryFromHistorySet 339
17.5.10.8 Method: dataObjectLabelFilter 341
17.5.10.9 Method: HSPS 341
17.5.10.10 Method: Discrete Risk Measures 341
17.5.11 RavenOutput 346
17.5.12 ETImporter 349
17.5.13 Metric 351
17.5.14 CrossValidation 352
17.5.14.1 SciKitLearn 353
17.5.14.2 K-fold 353
17.5.14.3 Stratified k-fold 354
17.5.14.4 Label k-fold 354
17.5.14.5 Leave-One-Out - LOO 355

17.5.14.6 Leave-P-Out - LPO 355
17.5.14.7 Leave-One-Label-Out - LOLO 355
17.5.14.8 Leave-P-Label-Out 356
17.5.14.9 ShuffleSplit 356
17.5.14.10 Label-Shuffle-Split 357
17.6 EnsembleModel 359
17.7 HybridModel 362
18 Functions 368
19 Metrics 370
19.1 Minkowski 370
19.2 Dynamic Time Warping 371
19.3 SKL Metrics 372
19.4 CDFAreaDifference 376
19.5 PDFCommonArea 376
20 Steps 377
20.1 SingleRun 378
20.2 MultiRun 380
20.3 IOStep 383
20.4 RomTrainer 388
20.5 PostProcess 389
21 Existing Interfaces 391
21.1 Generic Interface 391
21.2 RAVEN Interface 394
21.2.1 ExternalXML and RAVEN interface 398
21.3 RELAP5 Interface 398
21.3.1 Sequence 398
21.3.2 batchSize and mode 398
21.3.3 RunInfo 398
21.3.4 Files 399
21.3.5 Models 400
21.3.6 Distributions 401
21.3.7 Samplers 401
21.3.8 Steps 403
21.3.9 Databases 405
21.3.10 Modified Version of the Institute of Nuclear Safety System Incorporated (Japan) 405
21.4 RELAP7 Interface 406
21.4.1 Files 406
21.4.2 Models 406
21.4.3 Distributions 407
21.4.4 Samplers 407
21.5 MooseBasedApp Interface 408

21.5.1 Files . . . 408
21.5.2 Models . . . 409
21.5.3 Distributions . . . 409
21.5.4 Samplers . . . 410
21.5.5 Steps . . . 412
21.5.6 Databases . . . 413
21.5.7 DataObjects . . . 414
21.5.8 OutStreams . . . 414
21.6 MooseVPP Interface . . . 415
21.7 OpenModelica Interface . . . 416
21.7.1 Files . . . 417
21.7.2 Models . . . 417
21.7.3 CSV Output . . . 418
21.8 Dymola Interface . . . 419
21.8.1 Files . . . 420
21.8.2 Models . . . 421
21.9 Mesh Generation Coupled Interfaces . . . 423
21.9.1 MooseBasedApp and Cubit Interface . . . 423
21.9.1.1 Files . . . 424
21.9.1.2 Models . . . 424
21.9.1.3 Distributions . . . 425
21.9.1.4 Samplers . . . 425
21.9.1.5 Steps,OutStreams,DataObjects . . . 426
21.9.1.6 File Cleanup . . . 426
21.9.2 MooseBasedApp and Bison Mesh Script Interface . . . 426
21.9.2.1 Files . . . 427
21.9.2.2 Models . . . 427
21.9.2.3 Distributions . . . 428
21.9.2.4 Samplers . . . 428
21.9.2.5 Steps,OutStreams,DataObjects . . . 429
21.9.2.6 File Cleanup . . . 429
21.10 Rattlesnake Interfaces . . . 430
21.10.1 Files . . . 430
21.10.1.1 Perturb Yak Multigroup Cross Section Libraries . . . 430
21.10.1.2 Perturb Instant format Cross Section Libraries . . . 432
21.10.2 Models . . . 433
21.10.3 Distributions . . . 433
21.10.3.1 Samplers . . . 433
21.10.4 Steps . . . 434
21.11 MAAP5 Interface . . . 435
21.11.1 RAVEN Input file . . . 435
21.11.1.1 Files . . . 435

21.11.1.2 Models 435
21.11.1.3 Other blocks 436
21.11.2 MAAP5 Input files 436
21.11.2.1 MAAP5 include file 437
21.11.2.2 MAAP5 input file 438
21.11.2.3 MAAP5 PLOTFIL blocks 439
21.12 MAMMOTH Interface 440
21.12.1 Files 440
21.12.2 Models 441
21.12.3 Distributions 441
21.12.3.1 Samplers 441
21.12.4 Steps 443
21.13 MELCOR Interface 444
21.13.1 Sequence 444
21.13.2 batchSize and mode 444
21.13.3 RunInfo 444
21.13.4 Files 445
21.13.5 Models 446
21.13.6 Distributions 446
21.13.7 Samplers 447
21.13.8 Steps 448
21.13.9 Databases 449
21.13.10 DataObjects 450
22 Advanced Users: How to couple a new code 452
22.1 Pre-requisites. 453
22.2 Code Interface Creation 456
22.2.1 Method: `generateCommand` 456
22.2.2 Method: `createNewInput` 458
22.2.3 Method: `getInputExtension` 459
22.2.4 Method: `finalizeCodeOutput` 459
22.2.5 Method: `checkForOutputFailure` 460
22.3 Tools for Developing Code Interfaces 460
22.3.1 File Objects 461
23 Advanced Users: How to create a RAVEN ExternalModel plugin 463
23.1 ExternalModel Plugin Input 464
23.2 ExternalModel Plugin Creation 465
23.2.1 Method: `run` 465
23.2.2 Method: `createNewInput` 466
23.2.3 Method: `_readMoreXML` 467
23.2.4 Method: `initialize` 468

Appendix

A Appendix: Example Primer 471
A.1 Example 1. 471
A.2 Example 2. 474
References 482

1 Introduction

RAVEN is a software framework able to perform parametric and stochastic analysis based on the response of complex system codes. The initial development was aimed at providing dynamic risk analysis capabilities to the thermohydraulic code RELAP-7, currently under development at Idaho National Laboratory (INL). Although the initial goal has been fully accomplished, RAVEN is now a multi-purpose stochastic and uncertainty quantification platform, capable of communicating with any system code.

In fact, the provided Application Programming Interfaces (APIs) allow RAVEN to interact with any code as long as all the parameters that need to be perturbed are accessible by input files or via python interfaces. RAVEN is capable of investigating system response and explore input space using various sampling schemes such as Monte Carlo, grid, or Latin hypercube. However, RAVEN strength lies in its system feature discovery capabilities such as: constructing limit surfaces, separating regions of the input space leading to system failure, and using dynamic supervised learning techniques.

The development of RAVEN started in 2012 when, within the Nuclear Energy Advanced Modeling and Simulation (NEAMS) program, the need to provide a modern risk evaluation framework arose. RAVEN's principal assignment is to provide the necessary software and algorithms in order to employ the concepts developed by the Risk Informed Safety Margin Characterization (RISMC) program. RISMC is one of the pathways defined within the Light Water Reactor Sustainability (LWRS) program.

In the RISMC approach, the goal is not just to identify the frequency of an event potentially leading to a system failure, but the proximity (or lack thereof) to key safety-related events. Hence, the approach is interested in identifying and increasing the safety margins related to those events. A safety margin is a numerical value quantifying the probability that a safety metric (e.g. peak pressure in a pipe) is exceeded under certain conditions.

Most of the capabilities, implemented having RELAP-7 as a principal focus, are easily deployable to other system codes. For this reason, several side activates have been employed (e.g. RELAP5-3D, any MOOSE-based App, etc.) or are currently ongoing for coupling RAVEN with several different software. The aim of this document is to detail the input requirements for RAVEN focusing on the input structure.

2 Manual Formats

In order to highlight some parts of the Manual having a particular meaning (e.g. input structure, examples, terminal commands, etc.), specific formats have been used. In this sections all the formats with a specific meaning are reported:

- ***Python Coding:***

```
class AClass():
  def aMethodImplementation(self):
    pass
```

- ***XML input example:***

```
<MainXMLBlock>
  ...
  <anXMLnode name='anObjectName' anAttribute='aValue'>
     <aSubNode>body</aSubNode>
  </anXMLnode>
  ...
</MainXMLBlock>
```

- ***Bash Commands:***

```
cd trunk/raven/
./raven_libs_script.sh
cd ../../
```

3 Installation Overview

The installation of the RAVEN code is a straightforward procedure; depending on the usage purpose and machine architecture, the installation process slightly differs.

In the following sections, all the different installation procedures are reported.

There are two main requirements to installing RAVEN, installing the RAVEN dependencies (Section 4) and installing RAVEN itself (Section 5). For any particular installation, only one of the raven dependency procedures and one of the raven installation paths needs to be taken.

For macOS (OSX) it is recommended that the dependencies be installed with Miniconda (Section 4.2.3). For Linux, it is recommended that the distribution package manager be used if possible (Section 4.1) or Miniconda (Section 4.2.3).

There are several different ways of getting RAVEN as described in Section 5. They vary depending on how easy they are to use and how easy they are to develop with (these tend to be inversely correlated).

4 RAVEN Dependencies Installation

RAVEN is built upon several freely available open-source software packages, which must be installed before it will function properly:

Package	Purpose
git	Source code control tool
g++	C++ language compiler suite (Needed to build support code in the RAVEN package)
libtool	Generic library management tool
python	Scripting language RAVEN is written in
swig	Simplified Wrapper and Interface Generator (Used to create Python interfaces to supporting C++ code)
hdf5	Library providing interface to HDF5 database file format used by RAVEN
h5py	Python interface to HDF5 database library
numpy	N-Dimensional array package for Python
scipy	Scientific computing package for Python
scikit-learn	Machine learning library for Python
matplotlib	Plotting library for Python

RAVEN is supported on three separate computing platforms: Linux, OSX (Apple Macintosh), and Microsoft Windows. Depending on which of these systems is used, the preparation of the system to run RAVEN varies.

4.1 Preparing a Linux System for RAVEN

The installation of RAVEN dependencies on a Linux system can be performed using two alternative methods:

- Native distribution's package manager
- Miniconda package manager

Using one of the above automates the process and automatically includes any needed dependencies of the requested packages. Below are instructions for doing so for two popular Linux distributions, Ubuntu and Fedora.

4.1.1 Ubuntu

4.1.1.1 Advanced Package Tool (APT)

The Ubuntu distribution of Linux makes use of the Advanced Package Tool (APT) to automate the installation of pre-configured software. Ubuntu 16.4 or newer provide all the packages needed for RAVEN. The following command uses the APT to add the needed packages and any needed dependencies not already on the system:

```
sudo apt-get install libtool git python-dev swig g++ \
python3-dev python-numpy python-sklearn python-h5py
```

4.1.1.2 Miniconda

The Minoconda package manager is a cross platform installation package specifically used for Python dependencies installation. The package manager can be downloaded and installed from `https://conda.io/miniconda.html`. After the installation of Minconda, the following command needs to be executed for the installation of the needed packages and any needed dependencies not already on the system:

```
conda create --name raven_libraries -y numpy=1.11.0 \
 h5py=2.6.0 scipy=0.17.1 scikit-learn=0.17.1 \
 matplotlib=1.5.1 python=2.7 hdf5 swig pylint lxml
```

4.1.1.3 Optional LateX installation

Optionally, if you want to be able to edit and rebuild the manual, you can install TEX Live and its related packages:

```
 sudo apt-get install texlive-latex-base \
 texlive-extra-utils texlive-latex-extra texlive-math-extra
```

Now go on to Section 5 for Raven installation.

4.1.2 Fedora

4.1.2.1 Red Hat Package Manager (RPM)

The Fedora distribution of Linux makes use of the ARed Hat Package Manager (RPM) to automate the installation of pre-configured software. The following command uses the RPM to add the needed packages and any needed dependencies not already on the system:

```
dnf install swig libtool gcc-c++ redhat-rpm-config python-devel \
  python3-devel numpy h5py scipy python-scikit-learn \
  python-matplotlib-qt4
```

Note: The 'dnf' command replaces 'yum' used on older versions of Fedora Linux.

4.1.2.2 Miniconda

The Minoconda package manager is a cross platform installation package specifically used for Python dependencies installation. The package manager can be downloaded and installed from `https://conda.io/miniconda.html`. After the installation of Minconda, the following command needs to be executed for the installation of the needed packages and any needed dependencies not already on the system:

```
 conda create --name raven_libraries -y numpy=1.11.0 \
  h5py=2.6.0 scipy=0.17.1 scikit-learn=0.17.1 \
  matplotlib=1.5.1 python=2.7 hdf5 swig pylint lxml
```

4.1.2.3 Optional LateX installation

In addition, if you would like to be able to build the documentation included in the RAVEN software distribution it is necessary to install TEX Live and its related packages:

```
dnf install texlive texlive-subfigure texlive-stmaryrd \
  texlive-titlesec texlive-preprint texlive-placeins \
  texlive-bigints texlive-relsize texlive-appendix
```

Now go on to Section 5 for Raven installation.

4.2 Preparing an Apple Macintosh OSX System for RAVEN

When using an Apple Macintosh computer, the above dependencies are met by following three steps: Installing the XCode command line tools from Apple, installing the XQuartz X-Window system server, and then installing and using the Miniconda package system to add the rest of the dependencies.

4.2.1 Installing XCode Command Line Tools

The XCode command line tools package from Apple Computer provides the C++ compilers and git source code control tools needed to obtain and build RAVEN. It is freely available from the Apple store. In order to obtain it the following command should be launched in an open terminal:

```
xcode-select --install
```

4.2.2 Installing XQuartz

XQuartz is an implementation of the X Server for the Mac OSX operating system. XQuartz is freely available on the web and can be downloaded from the link `https://dl.bintray.com/xquartz/downloads/XQuartz-2.7.9.dmg`.
After downloaded, install the package.

4.2.3 Install RAVEN libraries

For OSX, the simplest and most robust way to install the RAVEN dependencies is with Miniconda. Miniconda is a package manager for python packages and can be used to install the third party packages that RAVEN depends on. Minconda can be downloaded and installed from the following link `https://conda.io/miniconda.html`.
Once installed, open a Terminal and launch the following command:

```
conda create --name raven_libraries -y numpy=1.11.0 \
 h5py=2.6.0 scipy=0.17.1 scikit-learn=0.17.1 \
 matplotlib=1.5.1 python=2.7 hdf5 swig pylint lxml
```

This command will install all the libraries and dependencies needed for executing RAVEN in a Miniconda enviroment called “raven_libraries”.

Now go on to Section 5 for Raven installation.

4.3 Preparing a Windows System for RAVEN

Since RAVEN requires a UNIX-like shell to function, one must be installed for a Microsoft Windows system to run it. A freely available software package called MSYS2 is used to provide this functionality. More information about MSYS2 is available at `https://sourceforge.net/p/msys2/wiki/MSYS2%20introduction/`.

4.3.1 Prerequisites to use RAVEN on Windows

- A system running a 64-bit version of Microsoft Windows. Installation and operation has been verified on Windows 7, 10, and Windows Server 2012 R2 Standard. While there is also a 32-bit version of MSYS2 available, the RAVEN installation described here will not work with it.
- At least 9 Gigabytes of available disk space:
  - 0.5 GB for MSYS2, including supporting tools and git source code control
  - 1.5 GB for Python language and supporting packages
  - 1.5 GB for RAVEN and the MOOSE framework
  - 5.0 GB for the Visual Studio compiler needed to build RAVEN

4.3.2 Installation and Configuration of the MSYS2 environment

1. Obtain and run the latest basic 64-bit MSYS2 installer from `https://msys2.github.io/` (As of this writing it is named msys2-x86_64-20161025.exe and is approximately 67 Megabytes in size).
2. The page with the download also contains installation instructions. Perform the steps described there up to step 6 to install a minimal MSYS2 system and bring it up to date. Make sure that you install to path C:\msys64. This installation will create shortcuts in the Windows start menu that may be used to start UNIX-Like shells:
   - MSYS2 Shell
   - MinGW-w64 Win32 Shell
   - MinGW-w64 Win64 Shell

   When working with RAVEN, it is recommended to use ”MinGW-w64 Win64 Shell”, although any of them should work.
3. Use the MSYS2 package manager *pacman* to install a few tools that will be needed later. Enter the following command in an MSYS shell window:

```
USER@HOSTNAME MINGW64 ~
$ pacman -S git winpty make
```

The package manager will then download and install those packages (and their dependencies) from the MSYS2 repository.

4.3.3 Install Python Language and Package Support

1. Download the latest 64-bit installer for Windows Python 2.7 from `https://conda.io/miniconda.html` and install it.
2. The installer will ask whether Python should be installed for only the logged in user or for all users. Either option will work for RAVEN.
3. Locate and test the Python installation. Open a Windows command prompt and enter the command "*where python*", which attempts to locate a the Python language interpreter in the current system path. This looks like:

   ```
   C:\Users\USERID> where python
   C:\Users\USERID\AppData\Local\Continuum\Miniconda2\python.exe
   ```

4. Setup MSYS2 to find Python. MSYS2 has its own separate PATH which must also be adjusted so that Python and its associated tools may be found. This is done by converting the file system location of Python determined in the previous step to its MSYS2-compatible equivalent and using the result to setup MSYS2 so that it too can find it in the future.

   This is done by turning all backslashes ('\') in the path to be converted to forward slashes ('/'), and changing the drive letter from its '<letter>:' form to '/ <letter>'. In addition, any spaces in the path must be escaped using a backslash ('\') when converted.

   For example:

   ```
   C:\Users\USERID\AppData\Local\Continuum\Miniconda2
   ```

   becomes

   ```
   /c/Users/USERID/AppData/Local/Continuum/Miniconda2
   ```

   for MSYS2. Here is an example with spaces that need to be escaped:

   ```
   C:\Program Files\Common Files
   ```

   converted to MSYS2 form would become

```
/c/Program\ Files/Common\ Files
```

Three separate paths must be added to MSYS2 to enable all of the Python tools needed to be found. These are:

Path	Purpose
<Converted path from above>	Python executable
<Converted path from above>/Scripts	Conda (Needed to manage Python packages)
<Converted path from above>/Library/bin	Swig (Needed to build RAVEN)

These paths are added using shell commands that append new entries to the existing PATH variable without overwriting it. These commands take the following form:

```
export PATH=/c/Users/USERID/AppData/Local/Continuum/Miniconda2:$PATH
export PATH=/c/Users/USERID/AppData/Local/Continuum/Miniconda2/Scripts:$PATH
export PATH=/c/Users/USERID/AppData/Local/Continuum/Miniconda2/Library/bin:$PATH
```

To configure these needed paths in MSYS2 so that they persist, file " /.bashrc" will need to be edited. This may be done either using an MSYS2-based editor such as *vim* (VI-iMproved, which is included in the installation) or a Windows-based editor like *Wordpad* (included with Windows). Another excellent open source editor for Windows is *Notepad++* `https://notepad-plus-plus.org/`, which is also good for editing RAVEN input files.

5. Test Python in MSYS2. At this point open a new MSYS2 shell window and see if Python is now found in the PATH:
   Note: Due to the way that Python interacts with the MSYS2 shell, when using Python by itself in MSYS2 the *winpty* utility is provided. (If Python is run without winpty, it may appear to sit there and do nothing. Pressing <Ctrl>-C will interrupt it.)

```
USER@HOSTNAME MINGW64 ~
$ winpty python
Python 2.7.13 |Anaconda 4.0.0 (64-bit)| (default, Dec 19 2016, 13:29:36) [MSC v.1500 64 bit (AMD64)] on win32
Type "help", "copyright", "credits" or "license" for more information.
Anaconda is brought to you by Continuum Analytics.
Please check out: http://continuum.io/thanks and https://anaconda.org
>>>
>>> quit()
```

6. Install needed Python packages. RAVEN requires several Python packages to function properly. Now the *conda* command will be used to download and install them in an automated manner. The following asks *conda* to obtain the specified versions of the listed packages, as well as all of their dependencies.

```
conda install numpy=1.11.0 h5py=2.6.0 scipy=0.17.1 \
        scikit-learn=0.17.1 matplotlib=1.5.1 python=2.7 \
        hdf5 swig pylint lxml
```

4.3.4 Compiler Installation and Configuration

1. Download and install Visual Studio. A C++ language compiler that supports C++11 features is needed to perform this step. Microsoft's Visual Studio Community Edition is free and available from `https://www.visualstudio.com/downloads/`.

   The current version (as of this writing) is 2017. The 2015 and 2017 versions have been successfully used to build RAVEN. Professional and Enterprise versions of these will also work. If one of these is already present on your system, it is not necessary to obtain another one. Note that because C++11 language features are required, the "Microsoft Visual C++ Compiler for Python 2.7" often used for building Python add-ons will **not** work.

   After downloading and running the Visual Studio installer, it will ask what features to install. For building RAVEN, "Desktop development with C++" is needed at a minimum. Installation of other Visual Studio features should be fine.

2. Let the build system know where to find the compiler. When the build system attempts to search for an installed compiler, this process often fails with the error message "Unable to find vcvarsall.bat". This happens because Python version 2.7 has not been updated to automatically locate modern Visual Studio installations. To solve this it is necessary to help the Python build system find the C++ compiler on the system. The easiest way to do this is create a Windows batch (.BAT) file that will redirect the build system to the information it needs. First, locate the file VCVARSALL.BAT file installed as part of Visual Studio on your system. This location will usually be something like the following:

Visual Studio Version	Directory containing VCVARSALL.BAT
2015	C:\Program Files (x86)\Microsoft Visual Studio 14.0\VC
2017	C:\Program Files (x86)\Microsoft Visual Studio\2017\Community\VC\Auxiliary\Build

   Once the target file has been located it is necessary to create a couple of directories and one file. The first directory created must be named "VC" and should be created somewhere outside of the RAVEN source tree (such as your MinGW home directory):

   ```
   USER@HOSTNAME MINGW64 ~
   $ mkdir VC
   ```

   The next directory to be created must be inside the one just created. It is suggested to name it "target", because it is there that we will point the Python build system:

   ```
   USER@HOSTNAME MINGW64 ~
   $ cd VC

   USER@HOSTNAME MINGW64 ~/VC
   $ mkdir target
   ```

```
USER@HOSTNAME MINGW64 ~/VC
$ ls
target
```

The file to be created is named ”VCVARSALL.BAT”, and it must be written in the VC directory that was just made. The Python build system will be configured to find this file, which then redirects it to the actual file. Use a text editor (such as *vim* or *notepad* as described above) to create the file VCVARSALL.BAT:

```
USER@HOSTNAME MINGW64 ~/VC
$ vim VCVARSALL.BAT
```

or

```
USER@HOSTNAME MINGW64 ~/VC
$ notepad VCVARSALL.BAT
```

One line will need to be added to the new file VCVARSALL.BAT:

```
CALL "<Full_Path_To_VCVARSALL.BAT_File_Installed_by_Visual_Studio>" %1 %2 %3 %4 %5
```

For example, in the case of Visual Studio 2017 Community installed in the default location this would be:

```
CALL "C:\Program Files (x86)\Microsoft Visual Studio\2017\Community\VC\Auxiliary\Build\VCVARSALL.BAT" %1 %2 %3 %4 %5
```

Note the double quotes around the path and file name. These are necessary because there are spaces in some of the directory names that make up the full location of VCVARSALL.BAT.

After creating the new *VCVARSALL.BAT* in the directory *VC*, one more thing needs to be done to inform the Python build system where this file just created is. During the build process, an *environment variable* ”VS90COMNTOOLS” will be checked. The value of VS90COMNTOOLS will need to be set to the *target* directory just below the location of VCVARSALL.BAT file just created.

For example, if VCVARSALL.BAT was created in directory VC under your MinGW home directory, the variable VS90COMNTOOLS should point to ~/VC/target.

```
USER@HOSTNAME MINGW64 ~
$ export VS90COMNTOOLS=~/VC/target

USER@HOSTNAME MINGW64 ~
$ echo $VS90COMNTOOLS
/home/user/VC/target
```

4.3.5 For More Information

Note: An illustrated version of this procedure may be found on the Wiki page at `https://github.com/idaholab/raven/wiki/installationWindows`.

4.4 Manual Dependency Install

If other options don't work, the dependencies can be installed manually. This is sometimes tricky, pay attention to the order you install them. RAVEN uses the following packages (newer versions usually work):

1. numpy-1.11.0
2. h5py-2.6.0
3. scipy-0.17.1
4. sklearn-0.17.1
5. matplotlib-1.5.3
6. hdf5 1.8 or newer
7. swig 2.0 or newer
8. gcc 4.9 or newer (or other C++11 compiler)
9. TeX Live 2016 or newer (to build manuals)

Now go on to Section 5 for Raven installation.

4.5 How RAVEN finds Dependencies

RAVEN, when run from either `raven_framework` or in `run_tests` runs a script called `setup_raven_libs` which sets the up the dependencies if they are not already present.

If conda is available, it will try and activate a conda environment called `raven_libraries`

Otherwise, the `setup_raven_libs` bash script will try and source a script: `$HOME/.raven/environments/raven_libs_profile` or if that is not available: `/opt/raven_libs/environments/raven_libs_profile` These scripts can be created if alternative ways of finding the dependencies are needed.

5 RAVEN Installation

Once the RAVEN dependencies have been installed (See Section 4), the rest of RAVEN can be installed.

There are two different ways to get RAVEN. There are trade offs between how easy it is to setup and how easy it is to develop with the method. If you want to do development on RAVEN or keep up with RAVEN changes go to Section 5.1, which describes obtaining the software directly from the online repository. Otherwise, RAVEN may be installed from a stand alone source code package (Section 5.2).

5.1 Submodule Git

This install method uses git to obtain the software and uses submodules to get MOOSE. This can be used for RAVEN development.

First RAVEN needs to be cloned, and then the submodules initialized.

```
git clone https://github.com/idaholab/raven.git
cd raven
git submodule update --init moose
```

Next follow the compilation and testing instructions in Section 5.3.

To update the software, the git pull and submodule update commands can be used:

```
git pull
git submodule update
```

5.2 RAVEN Source Code Package

Untar the source install (if there is more than one version of the source tarball, the full filename will need to be used instead of *):

```
tar -xvzf raven_framework_*_source.tar.gz
cd raven
```

Next follow the compilation and testing instructions in Section 5.3.

5.3 RAVEN Compilation

The RAVEN modules should be compiled:

```
#change into the raven directory if needed.
make
```

Then the testing should be done:

```
./run_tests
```

The output should describe why any tests failed.

At the end, there should be a line that looks similar to the output below:

```
8 passed, 19 skipped, 0 pending, 0 failed
```

Normally there are skipped tests because either some of the codes are not available, or some of the test are not currently working. The output will explain why each is skipped.

If all the tests pass, you are ready to read about Running RAVEN in Section 6.

If the tests did not pass, check Section 5.4 on troubleshooting.

5.4 Troubleshooting the Installation

Often the problems result from one or more of the libraries being incorrect or missing. In the raven directory, the command:

```
./run_tests --library_report
```

can be used to check if all the libraries are available, and which ones are being used. If amsc, distribution1D or interpolationND are missing, then the RAVEN modules need to be compiled or recompiled. Otherwise, the RAVEN dependencies need to be fixed.

Note, that when using RAVEN remotely in a graphical session with X11 forwarded to the client, some tests may depend on live X11 forwarding to the remote client. If the user is not using X11 forwarding then RAVEN will work fine and not use X11. However, when the user has forwarded their X11 environment to a ssh client, the connection may timeout. The standard timeout of X for an untrusted connection is 20 minutes. The full test suite including those involving graphical output can take longer than the aforementioned timeout. One way to alleviate this is to login to the remote host of RAVEN using a trusted connection by using the -Y flag:

```
ssh -Y hostname
```

This should only be done when the user is using a secure connection to a known host though, as there are security concerns to the client machine when allowing a remote computer access to its graphical user interface.

5.5 In-use Testing

In use testing can be done by re-running the installation tests as described in Section 5.3.

6 Running RAVEN

The RAVEN code is a blend of C++, C, and Python software. The entry point resides on the Python side and is accessible via a command line interface. After following the instructions in the previous Section, RAVEN is ready to be used. The `raven_framework` script is in the raven folder. To run RAVEN, open a terminal and use the following command (replace `<inputFileName.xml>` with your RAVEN input file):

```
raven_framework <inputFileName.xml>
```

Alternatively, the `Driver.py` script can be directly used. The RAVEN driver is contained in the folder "`raven/framework`." In this case, the command is:

```
python raven/framework/Driver.py <inputFileName.xml>
```

7 Raven Input Structure

The RAVEN code does not have a fixed calculation flow, since all of its basic objects can be combined in order to create a user-defined calculation flow. Thus, its input (XML format) is organized in different XML blocks, each with a different functionality. The main input blocks are as follows:

- **<Simulation>**: The root node containing the entire input, all of the following blocks fit inside the *Simulation* block.
- **<RunInfo>**: Specifies the calculation settings (number of parallel simulations, etc.).
- **<Files>**: Specifies the files to be used in the calculation.
- **<Distributions>**: Defines distributions needed for describing parameters, etc.
- **<Samplers>**: Sets up the strategies used for exploring an uncertain domain.
- **<DataObjects>**: Specifies internal data objects used by RAVEN.
- **<Databases>**: Lists the HDF5 databases used as input/output to a RAVEN run.
- **<OutStreams>**: Visualization and Printing system block.
- **<Models>**: Specifies codes, ROMs, post-processing analysis, etc.
- **<Functions>**: Details interfaces to external user-defined functions and modules. the user will be building and/or running.
- **<Steps>**: Combines other blocks to detail a step in the RAVEN workflow including I/O and computations to be performed.

Each of these blocks are explained in dedicated sections in the following chapters.

7.1 Comments

Comments may be included in the RAVEN input using standard XML comments, using `<!--` and `-->` as shown in the example below.

```
<Simulation>
  ...
  <!-- An Example Comment -->
  <Samplers>
  ...
```

Comments may be placed anywhere *except* before the **<Simulation>** node or after the **</Simulation>** node. Comments outside the root node will cause errors in maintaining input file compatability. Additionally, comments must completely surround any nodes they comment out. Comments are intended to completely remove blocks of code, or to add readability. For instance, the following is INCORRECT usage:

```
<!--<Assembler> -->
<!--</Assembler> -->
```

and the following is compatible usage for a code block:

```
<!--<Samplers>
  <Monte Carlo name='mc'>
    ...
  </Monte Carlo>
  ...
</Samplers> -->
```

7.2 Verbosity

Each block within RAVEN also makes use of a **verbosity** system, which allows a user to control the level of output to the user interface. These settings are declared globally as attributes in the **<Simulation>** node, and locally in each block. The verbosity levels are

- **'silent'** - Only simulation-breaking errors are displayed.
- **'quiet'** - Errors as well as warnings are displayed.
- **'all'** (default) - Errors, warnings, and messages are displayed.
- **'debug'** - For developers. All errors, warnings, messages, and debug messages are displayed.

Examples of verbosity usage are included in many examples throughout this manual.

At the **<Simulation>** node, the following global variables can be set:

- **verbosity**, optional string, determines the global verbosity level. Defaults to **'all'**.
- **printTimeStamps**, optional boolean, determines whether time stamps will be added to printed messages. Defaults to true.
- **color**, optional boolean, determines whether ANSI color tags will be used in printed messages. Defaults to false.

7.3 External Input Files

The **<ExternalXML>** node defines external input file (XML format) that can be used to replace any XML nodes under **<Simulation>** in the RAVEN input file. This node allows a user to load any external input file that contains the required XML nodes into the RAVEN input file. Each **<ExternalXML>** node has the following attributes:

- **node**, ***required string attribute***, user-defined XML node of RAVEN input file.
- **xmlToLoad**, ***required string attribute***, file name with its absolute or relative path. Note: if a relative path is specified, it must be relative with respect to the RAVEN input file.

For example, if the file `Models.xml` contain the required RAVEN input XML node **<Models>**, the RAVEN input file might appear as:

```
<Simulation>
  ...
  <Steps>
    ...
  </Steps>
  ...
  <ExternalXML node='Models'
     xmlToLoad='external_input/Models.xml'/>
  ...
</Simulation>
```

Another example, if the file `MultiRun.xml` contain the required RAVEN input XML node **<MultiRun>** under node **<Steps>**, the RAVEN input file might appear as:

```
<Simulation>
  ...
  <Steps>
    ...
    <ExternalXML node='MultiRun'
       xmlToLoad='external_input/MultiRun.xml'/>
    ...
  </Steps>
  ...
</Simulation>
```

8 RunInfo

In the **RunInfo** block, the user specifies how the overall computation should be run. This block accepts several input settings that define how to drive the calculation and set up, when needed, particular settings for the machine the code needs to run on (e.g. queueing system, if not PBS, etc.). In the following subsections, we explain all the keywords and how to use them in detail.

8.1 RunInfo: Input of Calculation Flow

This sub-section contains the information regarding the XML nodes used to define the settings of the calculation flow that is being performed through RAVEN:

- **<WorkingDir>**, ***string, required field***, specifies the absolute or relative (with respect to the location where the xml file is located) path to a directory that will store all the results of the calculations and where RAVEN looks for the files specified in the block **<Files>**. If `runRelative='True'` is used as an attribute, then it will be relative to where raven is run.
Default: None

- **<RemoteRunCommand>**, ***string, optional field***, specifies the absolute or relative (with respect to the framework directory) path to a command that can be used on a remote machine to execute a command. The command is passed in as the environmental variable COMMAND.

  Default: raven_qsub_command.sh

- **<NodeParameter>**, ***string, optional field***, specifies the flag used to specify a node file for the MPIExec command. This will be followed by a file with the nodes that a single batch will run on.
Default: -f

- **<MPIExec>**, ***string, optional field***, specifies the command used to run mpi. This will be followed by the **<NodeParameter>** and then the node file and then the code command.
Default: mpiexec

- **<batchSize>**, ***integer, optional field***, specifies the number of parallel runs executed simultaneously (e.g., the number of driven code instances, e.g. RELAP5-3D, that RAVEN will spawn at the same time). Each parallel run will use `NumThreads` $*$ `NumMPI` cores.
Default: 1

- **<maxQueueSize>**, ***integer, optional field***, specifies the number of parallel runs that can be staged for running simultaneously. The RAVEN architecture is inherently multithreaded

where a job queue is continuously monitored by a job handling thread. New jobs are added to this queue as they become available from the main thread of execution. Since the main thread is also responsible for collecting the results of previously finished jobs, it is possible that faster jobs may complete before the main thread can replenish the queue. By increasing this value, you are allowing RAVEN to consume more memory in order to stage more jobs, placing them in a pending job queue, with the benefit that slower job collection times will be masked as the job handler will flush the complete jobs and run whatever is available on the pending queue. With smaller values, RAVEN will consume less memory staging jobs, but there is potential that the job processing thread may be starved of jobs and waste parallel cycles as the code degrades to serially waiting for the main thread to complete collecting finished jobs. Where **<batchSize>** represents the number of jobs running, **<maxQueueSize>** represents the total number of jobs running plus the queued jobs. Values of **<maxQueueSize>** less than **<batchSize>** will be ignored. By default, **<maxQueueSize>** will be equal to **<batchSize>**.

- **<Sequence>**, ***comma separated string, required field***, is an ordered list of the step names that RAVEN will run (see Section 20).

- **<JobName>**, ***string, optional field***, specifies the name to use for the job when submitting to a pbs queue. Acceptable characters include alphanumerics as well as "-" and "_". If more than 15 characters are provided, RAVEN will truncate it using a hyphen between the first 10 and last 4 character, i.e., "1234567890abcdefgh" will be truncated to "1234567890-efgh".
  Default: raven_qsub

- **<printInput>**, ***string, optional field***, if provided, indicates RAVEN should print out a duplicate of the input file. If the provided text is **'false'**, or the node is not provided, then no duplicate will be printed. If the node is provided but no name specified, it will use the default name. Otherwise, the file will be written in the working directory as `name_provided.xml`.
  Default: duplicated_input.xml

- **<NumThreads>**, ***integer, optional field***, can be used to specify the number of threads RAVEN should associate when running the driven software. For example, if RAVEN is driving a code named "FOO," and this code has multi-threading support, this block is used to specify how many threads each instance of FOO should use (e.g. "`FOO --n-threads=N`" where `N` is the number of threads).
  Default: 1 (or None when the driven code does not have multi-threading support)

- **<NumMPI>**, ***integer, optional field***, can be used to specify the number of MPI CPUs RAVEN should associate when running the driven software. For example, if RAVEN is driving a code named "FOO," and this code has MPI support, this block specifies how many MPI CPUs each instance of FOO should use (e.g. "`mpiexec FOO -np N`" where `N` is the number of CPUs).
  Default: 1 (or None when the driven code does not have MPI support)

- **<totalNumCoresUsed>**, ***integer, optional field***, is the global number of CPUs RAVEN is going to use for performing the calculation. When the driven code has MPI and/or multi-threading support and the user specifies NumThreads > 1 and NumMPI > 1, then totalNumCoresUsed is set according to the following formula:
totalNumCoresUsed = NumThreads $*$ NumMPI $*$ batchSize.
Default: 1

- **<internalParallel>**, ***boolean, optional field***, is a boolean flag that controls the type of parallel implementation needs to be used for Internal Objects (e.g. ROMs, External Models, PostProcessors, etc.). If this flag is set to:

  - **False**, the internal parallelism is employed using multi-threading (i.e. 1 processor, multiple threads equal to the **<batchSize>**).
    Note: This "parallelism mode" runs multiple instances of the Model in a single processor. If the evaluation of the model is memory intensive (i.e. it uses a lot of memory) or computational intensive (i.e. a lot of computation operations evolving in a $CPUt \approx> 0.1\frac{sec}{evaluation}$) the single processor might get over-loaded determining a degradation of performance. In such cases, the internal parallelism needs to be used (see the following);
  - **True**, the internal parallelism is employed using a internally-developed multi-processor approach (i.e. **<batchSize>** processors, 1 single thread). This approach works for both Shared Memory Systems (e.g. PC, laptops, workstations, etc.) and Distributed Memory Machines (e.g. High Performance Computing Systems, etc.).
    Note: This "parallelism mode" runs multiple instances of the Model in multiple processors. Since the parallelism is employed in Python, some overhead is present. This "mode" needs to be used when:
    * the Model evaluation is memory intensive (i.e. the multi-threading approach will cause the over-load of a single processor);
    * the Model evaluation is computation intensive (i.e. $CPUt \approx> 0.1\frac{sec}{evaluation}$).

  Default: False

- **<precommand>**, ***string, optional field***, specifies a command that needs to be inserted before the actual command that is used to run the external model (e.g., mpiexec -n 8 precommand ./externalModel.exe (...)). Note that the precommand as well as the postcommand are ONLY applied to execution commands flagged as "parallel" within the code interface.
Default: None

- **<postcommand>**, ***string, optional field***, specifies a command that needs to be appended after the actual command that is used to run the external model (e.g., mpiexec -n 8 ./externalModel.exe (...) postcommand). Note that the postcommand as well as the precommand are ONLY applied to execution commands flagged as "parallel"

within the code interface.
Default: None

- **<clusterParameters>**, ***string, optional field***, specifies extra parameters to be used with the cluster submission command. For example, if qsub is used to submit a command, then these parameters will be used as extra parameters with the qsub command. This can be repeated multiple times as needed and they will all be passed to the cluster submission command.
  Default: None

- **<MaxLogFileSize>**, ***integer, optional field***. specifies the maximum size of the log file in bytes. Every time RAVEN drives a code/software, it creates a logfile of the code's screen output.
  Default: ∞

  *(**Note:** This flag is not implemtend yet.)*

- **<deleteOutExtension>**, ***comma separated string, optional field***, specifies, if a run of an external model has not failed, which output files should be deleted by their extension (e.g., **<deleteOutExtension>**`txt,pdf`**</deleteOutExtension>** will delete all generated txt and pdf files). **Note:** This flag is only active for Models of type "Code".
  Default: None

- **<delSucLogFiles>**, ***boolean, optional field***, when True and the run of an external model has not failed (return code = 0), deletes the associated log files. **Note:** This flag is only active for Models of type "Code".
  Default: False

8.2 RunInfo: Input of Queue Modes

In this sub-section, all of the keywords (XML nodes) for setting the queue system are reported.

- **<mode>**, ***string, optional field***, can specify which kind of protocol the parallel enviroment should use. RAVEN currently supports one pre-defined "mode":
  - **mpi**: this "mode" uses **<MPIExec>** command (default: `mpiexec`) to distribute the running program; more information regarding this protocol can be found in [1]. Mode "MPI" can either generate a `qsub` command or can execute on selected nodes. In order to make the "mpi" mode generate a `qsub` command, an additional keyword (xml sub-node) needs to be specified:
    * If RAVEN is executed in the HEAD node of an HPC system using [2], the user needs to input a sub-node, **<runQSUB>**, right after the specification of the mpi mode (i.e.

<mode>mpi<runQSUB/></mode>). If the keyword is provided, RAVEN generates a qsub command, instantiates itself, and submits itself to the queue system.

* If the user decides to execute RAVEN from an "interactive node" (a certain number of nodes that have been reserved in interactive PBS mode), RAVEN, using the "mpi" system, is going to utilize the reserved resources (CPUs and nodes) to distribute the jobs, but, will not generate a qsub command.

When the user decides to run in "mpi" mode without making RAVEN generate a qsub command, different options are available:

* If the user decides to run on the local machine (either in local desktop/workstation or a remote machine), no additional keywords are needed (i.e. <mode>mpi</mode>).
* If the user is running on multiple nodes, the node ids have to be specified:
  · the node ids can be specified in an external text file (node ids separated by blank space). This file needs to be provided in the XML node <mode>, introducing a sub-node named <nodefile> (e.g. <mode>mpi<nodefile>/tmp/nodes</nodefile></mode>).
  · the node ids can be contained in an enviromental variable (node ids separated by blank space). This variable needs to be provided in the <mode> XML node, introducing a sub-node named <nodefileenv> (e.g. <mode>mpi<nodefileenv>NODEFILE</nodefileenv></mode>>).
  · If none of the above options are used, RAVEN will attempt to find the nodes' information in the enviroment variable PBS_NODEFILE.
* The cores needed can be specified manually with the <coresneeded>. This is directly used in the qsub command select statement.
* The max memory needed can be specified with the <memory> XML node. This will be used in the qsub command select statement.
* The placement can be specified with the <place> XML node. This will be used in the qsub place statement.
* There is a "mpilegacy" mode. This probably will be removed in the future. In this mode exec can be forced to run on one shared memory node with the <NoSplitNode>. If this is present, the splitting apart of the batches will put each batch on one shared memory node. Without <NoSplitNode>, they can be split across nodes. There is an option maxOnNode which puts at most maxOnNode number of mpi processes on one node. <NoSplitNode> can cause processes to not be placed, so <NoSplitNode> should not be used unless needed. If limiting the number of mpi processes on one node is desired without forcing them to only run on one node, <LimitNode> can be used. Both <NoSplitNode> and <LimitNode> can have a noOverlap which prevents multiple batches from running on a single node.

In addition, this flag activates the remote (PBS) execution of internal Models (e.g. ROMs, ExternalModels, PostProcessors, etc.). If this node is not present, the inter-

nal Models are run using a multi-threading approach (i.e. master processor, multiple parallel threads)

- **<CustomMode>**, ***xml node, optional field***, is an xml node where “advanced” users can implement newer “modes.” Please refer to sub-section 8.4 for advanced users.
- **<queueingSoftware>**, ***string, optional field***. RAVEN has support for the PBS queueing system. If the platform provides a different queueing system, the user can specify its name here (e.g., PBS PROFESSIONAL, etc.).
  Default: PBS PROFESSIONAL
- **<expectedTime>**, ***colum separated string, optional field (mpi or custom mode)***, specifies the time the whole calculation is expected to last. The syntax of this node is *hours:minutes:seconds* (e.g. `40:10:30` equals 40 hours, 10 minutes, 30 seconds). After this period of time, the HPC system will automatically stop the simulation (even if the simulation is not completed). It is preferable to rationally overestimate the needed time.
  Default: 10:00:00 (10 hours.)

8.3 RunInfo: Example Cluster Usage

For this example, we have a PBSPro cluster, and there are thousands of node, and each node has 4 processors that share memory. There are a couple different ways this can be used. One way is to use interactive mode and have a RunInfo block:

```
<RunInfo>
    <WorkingDir>.</WorkingDir>
    <Sequence>FirstMRun</Sequence>
    <batchSize>3</batchSize>
    <NumThreads>4</NumThreads>
    <mode>mpi</mode>
    <NumMPI>2</NumMPI>
</RunInfo>
```

Then the commands can be used:

```
#Note: select=NumMPI*batchSize, ncpus=NumThreads
qsub -l select=6:ncpus=4:mpiprocs=1 -l walltime=10:00:00 -I
#wait for processes to be allocated and interactive shell to start

#Switch to the correct directory
cd $PBS_O_WORKDIR

#Load the module with the raven libraries
```

```
module load raven-devel-gcc

#Start Raven
python  ../../framework/Driver.py test_mpi.xml
```

Alternatively, RAVEN can be asked to submit the qsub directory. With this, the RunInfo is:

```xml
<RunInfo>
    <WorkingDir>.</WorkingDir>
    <Sequence>FirstMQRun</Sequence>
    <batchSize>3</batchSize>
    <NumThreads>4</NumThreads>
    <mode>
      mpi
      <runQSUB/>
    </mode>
    <NumMPI>2</NumMPI>
    <expectedTime>10:00:00</expectedTime>
</RunInfo>
```

In this case, the command run from the cluster submit node:

```
python ../../framework/Driver.py test_mpiqsub_local.xml
```

8.4 RunInfo: Advanced Users

This sub-section addresses some customizations of the running environment that are possible in RAVEN. Firstly, all the keywords reported in the previous sections can be pre-defined by the user in an auxiliary XML input file. Every time RAVEN gets instantiated (i.e. the code is run), it looks for an optional file, named "`default_runinfo.XML`" contained in the "`\home\username\.raven\`" directory (i.e. "`\home\username\.raven\default_runinfo.XML`"). This file (same syntax as the RunInfo block defined in the general input file) will be used for defining default values for the data in the RunInfo block. In addition to the keywords defined in the previous sections, in the **<RunInfo>** node, an additional keyword can be defined:

- **<DefaultInputFile>**, ***string, optional field***. In this block, the user can change the default xml input file RAVEN is going to look for if none have been provided as a command-line argument.
  Default: "test.xml".

As already mentioned, this file is read to define default data for the RunInfo block. This means that all the keywords defined here will be overridden by any values specified in the actual RAVEN input file.
In section 8.2, it is explained how RAVEN can handle the queue and parallel systems. If the currently available "modes" are not suitable for the user's system (workstation, HPC system, etc.), it is possible to define a custom "mode" modifying the **<RunInfo>** block as follows:

```
<RunInfo>
    ...
    <CustomMode file="newMode.py" class="NewMode">
       aNewMode
    </CustomMode>
    <mode>aNewMode</mode>
    ...
</RunInfo>
```

The file field can use %BASE_WORKING_DIR% and %FRAMEWORK_DIR% to specify the location of the file with respect to the base working directory or the framework directory.

The python file should define a class that inherits from `Simulation.SimulationMode` of the RAVEN framework and overrides the necessary functions. Generally, `modifySimulation` will be overridden to change the precommand or postcommand parts which will be added before and after the executable command. An example Python class is given below with the functions that can and should be overridden:

```
import Simulation
class NewMode(Simulation.SimulationMode):
  def remoteRunCommand(self, runInfoDict):
    # If it returns a dictionary, then run the command in args
    # Example: {"args":["ssh","remotehost","raven_framework"]}
    # Note that this command needs to be able to tell when it
    # is running remotely, and then return None at that point
    return None

  def modifyInfo(self, runInfoDict):
    # modifyInfo is called after the runInfoDict has been
    # setup and allows the mode to change any parameters that
    # need changing. This typically modifies the precommand and
    # the postcommand that are put before/after the command.
    # In order to change them, return a dictionary with new values.
    # Those new values will be used.
    return {}

  def XMLread(self,XMLNode):
```

```
    # XMLread is called with the mode node, and can be used to
    # get extra parameters needed for the simulation mode.
    pass
```

RAVEN's Job Handler module controls the creation and execution of individual code runs. Essentially, the SimulationMode class may be used when it is necessary to customize that behavior. First, it allows providing a remote command for running RAVEN. This first method can be used if for example RAVEN needs to be run on a different machine such as a head node of a computer cluster. In such a case, a remoteRunCommand function can be created that causes RAVEN to be instantiated on the cluster head node (in cases where that is different than the computer where the user is currently working). Secondly, (and usually easier when this is sufficient) the SimulationMode class allows modifying the various run info parameters before the code is run.

For modification of the run info parameters, generally the two most important are precommand and postcommand. They are placed in front and back before running the code. So for example if precommand is 'mpiexec -n 3' and postcommand is '–number-threads=4' and the code command is 'runIt' then the full command would be: 'mpiexec -n 3 runIt –number-threads=4' The precommand and postcommand are used for any run type that is 'parallel', but not for 'serial' codes. They can be modified by overriding the `modifyInfo` method and returning a new dictionary with new values. The `runInfoDict` in the simulation is passed in.

To help with these commands, there are several variables that are substituted in before running the command. These are:

%INDEX% Contains the zero-based index in list of running jobs. Note that this is stable for the life of the job. After the job finishes, this is reused. An example use would be if there were four cpus and the batch size was four, the %INDEX% could be used to determine which cpu to run on.

%INDEX1% Contains the one-based index in the list of running jobs, same as %INDEX%+1

%CURRENT_ID% zero-based id for the job handler. This starts as 0, and increases for each job the job handler starts.

%CURRENT_ID1% one-based id for the job handler, same as %CURRENT_ID%+1

%SCRIPT_DIR% Expands to the full path of the script directory (raven/scripts)

%FRAMEWORK_DIR% Expands to the full path of the framework directory (raven/framework)

%WORKING_DIR% Expands to the working directory where the input is

%BASE_WORKING_DIR% Expands to the base working directory given in RunInfo. This will likely be a parent of WORKING_DIR

%METHOD% Expands to the environmental variable $METHOD

%NUM_CPUS% Expands to the number of cpus to use per single batch. This is NumThreads in the XML file.

The final joining of the commands and substituting the variables is done in the JobHandler class.

8.5 RunInfo: Examples

Here we present a few examples using different components of the RunInfo node:

```
<RunInfo>
    <WorkingDir>externalModel</WorkingDir>
    <Sequence>MonteCarlo</Sequence>
    <batchSize>100</batchSize>
    <NumThreads>4</NumThreads>
    <mode>mpi</mode>
    <NumMPI>2</NumMPI>
</RunInfo>

<Files>
    <Input name='lorentzAttractor.py'
       type=''>lorentzAttractor.py</Input>
</Files>
```

This examples specifies the working directory (`WorkingDir`) where the necessary file (`Files`) is located and to run a series of 100 (`batchSize`) Monte-Carlo calculations (`Sequence`). MPI mode (`mode`) is used along with 4 threads (`NumThreads`) and 2 MPI processes per run (`NumMPI`).

9 Files

The <Files> block defines any files that might be needed within the RAVEN run. This could include inputs to the Model, pickled ROM files, or CSV files for postprocessors, to name a few. Each entry in the <Files> block is a tag with the file type. Files given through the input XML at this point are all <Input> type. Each <Input> node has the following attributes:

- **name**, ***required string attribute***, user-defined name of the file. This does not need to be the actual filename; this is the name by which RAVEN will identify the file. **Note:** As with other objects, this name can be used to refer to this specific entity from other input blocks in the XML.
- **type**, ***optional string attribute***, a type label for this file. While RAVEN does not directly make use of file types, they are available in the CodeInterface as identifiers. If not provided, the type will be stored as python `None` type.
- **perturbable**, ***optional boolean attribute***, flag to indicate whether a file can be perturbed or not. RAVEN does not directly use this attribute, but it is availabe in the CodeInterface. If not provided, defaults to True.
- **subDirectory**, ***optional string attribute***, sub-directory that should be created in the perturbation process. The file specified in the body of the XML node should be located in the **subDirectory** under the **workingDir** specified in the <RunInfo> XML block (i.e. $workingDir/subDirectory$). If specified, the file will be placed in the sub-directory. For example, in a *MultiRun* step, the file will be copied into $workingDir/stepName/\%counter\%/subDirectory$, where $workingDir$ is the working directory specified in the *RunInfo* XML block, $stepName$ is the name of the step, $/\%counter\%$ is the realization identifier (e.g. 1,2, etc.) and $subDirectory$ is the sub-directory here specified. If not provided, defaults to an empty string.

For example, if the files `templateInput.i, materials.i, history.i, mesh.e` are required to run a Model, the <Files> block might appear as:

```
  ...
  <Files>
    <Input name='main' type='maininput'>templateInput.i</Input>
    <Input name='mat'  type='mtlinput' >materials.i</Input>
    <Input name='hist' type='histinput'>history.i</Input>
    <Input name='mesh' type='mesh'
       perturbable='false'>mesh.e</Input>
    <Input name='fileInSubDir' type=''
       subDirectory="theSubDirectory">theFileInTheSubDir.inp</Input>
  </Files>
    ...
</Simulation>
```

10 VariableGroups

The <VariableGroups> block is an optional input for the convenience of the user. It allows the possibility of creating a collection of variables instead of re-listing all the variables in places throughout the input file, such as DataObjects, ROMs, and ExternalModels. Each entry in the <VariableGroups> block has a distinct name and list of each constituent variable in the group. Alternatively, set operations can be used to construct variable groups from other variable groups. In this case, the dependent groups and the base group on which operations should be performed must be listed. The following types of set operations are included in RAVEN:

- +, Union, the combination of all variables in the **'base'** set and listed set,
- −, Complement, the relative complement of the listed set in the **'base'** set,
- ^, Intersection, the variables common to both the **'base'** and listed set,
- %, Symmetric Difference, the variables in only either the **'base'** or listed set, but not both.

Multiple operations can be performed by separating them with commas in the text of the group node. In the event the listed set is a single variable, it will be treated like a set with a single entry.

When using the variable groups in a node, they can be listed alone or as part of a comma-separated list. The variable group name will only be substituted in the text of nodes, not attributes or tags.

Each <Group> node has the following attributes:

- name, ***required string attribute***, user-defined name of the group. This is the identifier that will be used elsewhere in the RAVEN input.
- dependencies, ***optional comma-separated string attribute***, the other variable groups on which this group is dependent for construction. If listed, all entries in the text of this node must be proceeded by one of the set operators above. Defaults to an empty string.
- base, ***optional string attribute***, the starting set for constructing a dependent variable group. This attribute is required if any dependencies are listed. Set operations are performed by performing the chosen operation on the variable group listed in this attribute along with the listed variable group. No default.

An example of constructing and using variable groups is listed here. The variable groups **'x_odd'**, **'x_even'**, **'x_first'**, and **'y_group'** are constructed independently, and the remainder are examples of other operations.

```
...
<VariableGroups>
  <Group name="x_odd"  >x1,x3,x5</Group>
```

```
    <Group name="x_even" >x2,x4,x6</Group>
    <Group name="x_first">x1,x2,x3</Group>
    <Group name="y_group">y1,y2</Group>
    <Group name="add_remove" dependencies="x_first"
      base="x_first">-x1,+ x4,+x5</Group>
    <Group name="union"      dependencies="x_odd,x_even"
      base="x_odd">+x_even</Group>
    <Group name="complement" dependencies="x_odd,x_first"
      base="x_odd">-x_first</Group>
    <Group name="intersect"  dependencies="x_even,x_first"
      base="x_even">^x_first</Group>
    <Group name="sym_diff"   dependencies="x_odd,x_first"
      base="x_odd">% x_first</Group>
  </VariableGroups>
  ...
  <DataObjects>
    <PointSet name="dataset">
      <Input>union</Input>
      <Output>y_group</Output>
    </PointSet>
  </DataObjects>
  ...
</Simulation>
```

11 Distributions

RAVEN provides support for several probability distributions. Currently, the user can choose among several 1-dimensional distributions and N-dimensional ones, either custom or multidimensional normal.

The user will specify the probability distributions, that need to be used during the simulation, within the **<Distributions>** XML block:

```
<Simulation>
   ...
  <Distributions>
    <!-- All the necessary distributions will be listed here -->
  </Distributions>
  ...
</Simulation>
```

In the next two sub-sections, the input requirements for all of the distributions are reported.

11.1 1-Dimensional Probability Distributions

This sub-section is organized in two different parts: 1) continuous 1-D distributions and 2) discrete 1-D distributions. These two paragraphs cover all the requirements for using the different distribution entities.

11.1.1 1-Dimensional Continuous Distributions

In this paragraph all the 1-D distributions currently available in RAVEN are reported.

Firstly, all the probability distributions functions in the code can be truncated by using the following keywords:

```
<Distributions>
   ...
   <aDistributionType>
      ...
      <lowerBound>aFloatValue</lowerBound>
      <upperBound>aFloatValue</upperBound>
      ...
   </aDistributionType>
```

```
</Distributions>
```

Each distribution has a pre-defined, default support (domain) based on its definition, however these domains can be shifted/stretched using the appropriate **<low>** and **<high>** parameters where applicable, and/or truncated using the nodes in the example above, namely **<lowerBound>** and **<upperBound>**. For example, the Normal distribution domain is $[-\infty, +\infty]$, and thus cannot be shifted or stretched, as it is already unbounded, but can be truncated. RAVEN currently provides support for 13 1-Dimensional distributions. In the following paragraphs, all the input requirements are reported and commented.

11.1.1.1 Beta Distribution

The **Beta** distribution is parameterized by two positive shape parameters, denoted by α and β, that appear as exponents of the random variable. Its default support (domain) is $x \in [0, 1]$. The distribution domain can be changed, specifying new boundaries, to fit the user's needs. The user can specify a **Beta** distribution in two ways. The standard is to provide the parameters **<low>**, **<high>**, **<alpha>**, and **<beta>**. Alternatively, to approximate a normal distribution that falls to 0 at the endpoints, the user may provide the parameters **<low>**, **<high>**, and **<peakFactor>**. The peak factor is a value between 0 and 1 that determines the peakedness of the distribution. At 0 it is dome-like ($\alpha = \beta = 4$) and at 1 it is very strongly peaked around the mean ($\alpha = \beta = 100$). A reasonable approximation to a Gaussian normal is a peak factor of 0.5.

The specifications of this distribution must be defined within a **<Beta>** XML block. This XML node accepts one attribute:

- **name**, ***required string attribute***, user-defined name of this distribution. **Note:** As with other objects, this identifier can be used to reference this specific entity from other input blocks in the XML.

This distribution can be initialized with the following children nodes:

- Standard initialization:
  - **<alpha>**, ***float, conditional required parameter***, first shape parameter. If specified, **<beta>** must also be inputted and **<peakFactor>** can not be specified.
  - **<beta>**, ***float, conditional required parameter***, second shape parameter. If specified, **<alpha>** must also be inputted and **<peakFactor>** can not be specified.
  - **<low>**, ***float, optional parameter***, lower domain boundary.
    Default: 0.0
  - **<high>**, ***float, optional parameter***, upper domain, boundary.
    Default: 1.0

- Alternative initialization:

  - **<peakFactor>**, ***float, optional parameter***, alternative to specifying **<alpha>** and **<beta>**. Acceptable values range from 0 to 1.
  - **<low>**, ***float, optional parameter***, lower domain boundary.
    Default: 0.0
  - **<high>**, ***float, optional parameter***, upper domain, boundary.
    Default: 1.0

Example:

```xml
<Distributions>
  ...
  <Beta name='aUserDefinedName'>
     <low>aFloatValue</low>
     <high>aFloatValue</high>
     <alpha>aFloatValue</alpha>
     <beta>aFloatValue</beta>
  </Beta>
  <Beta name='aUserDefinedName2'>
     <low>aFloatValue</low>
     <high>aFloatValue</high>
     <peakFactor>aFloatValue</peakFactor>
  </Beta>
  ...
</Distributions>
```

11.1.1.2 Exponential Distribution

The **Exponential** distribution has a default support of $x \in [0, +\infty)$.

The specifications of this distribution must be defined within an **<Exponential>** XML block. This XML node accepts one attribute:

- **name**, ***required string attribute***, user-defined name of this distribution. **Note:** As with other objects, this identifier can be used to reference this specific entity from other input blocks in the XML.

This distribution can be initialized with the following child node:

- **<lambda>**, ***float, required parameter***, rate parameter.

- <low>, ***float, optional parameter***, lower domain boundary.
  Default: 0.0

Example:

```
<Distributions>
  ...
  <Exponential name='aUserDefinedName'>
    <lambda>aFloatValue</lambda>
    <low>aFloatValue</low>
  </Exponential>
  ...
</Distributions>
```

11.1.1.3 Gamma Distribution

The **Gamma** distribution is a two-parameter family of continuous probability distributions. The common exponential distribution and χ-squared distribution are special cases of the gamma distribution. Its default support is $x \in [0, +\infty]$.

The specifications of this distribution must be defined within a <Gamma> XML block. This XML node accepts one attribute:

- name, ***required string attribute***, user-defined name of this distribution. **Note:** As with other objects, this identifier can be used to reference this specific entity from other input blocks in the XML.

This distribution can be initialized with the following children nodes:

- <alpha>, ***float, required parameter***, shape parameter.
- <beta>, ***float, optional parameter***, 1/scale or the inverse scale parameter.
  Default: 1.0
- <low>, ***float, optional parameter***, lower domain boundary.
  Default: 0.0

Example:

```
<Distributions>
  ...
  <Gamma name='aUserDefinedName'>
    <alpha>aFloatValue</alpha>
```

```
    <beta>aFloatValue</beta>
    <low>aFloatValue</low>
  </Gamma>
  ...
</Distributions>
```

11.1.1.4 Laplace Distribution

The **Laplace** distribution is a two-parameter continuous probability distribution. It is the distribution of the differences between two independent random variables with identical exponential distributions. Its default support is $x \in (-\infty, +\infty)$.

The specifications of this distribution must be defined within a <Laplace> XML block. This XML node accepts one attribute:

- name, ***required string attribute***, user-defined name of this distribution. **Note:** As with other objects, this identifier can be used to reference this specific entity from other input blocks in the XML.

This distribution can be initialized with the following children nodes:

- <location>, ***float, required parameter***, determines the location or shift of the distribution.
- <scale>, ***float, required parameter***, must be greater than 0, and determines how spread out the distribution is.

Example:

```
<Distributions>
  ...
  <Laplace name='aUserDefinedName'>
    <location>aFloatValue</location>
    <scale>aFloatValue</scale>
  </Laplace>
  ...
</Distributions>
```

11.1.1.5 Logistic Distribution

The **Logistic** distribution is similar to the normal distribution with a CDF that is an instance of a logistic function ($Cdf(x) = \frac{1}{1+e^{-\frac{(x-location)}{scale})}}$). It resembles the normal distribution in shape but has

heavier tails (higher kurtosis). Its default support is $x \in [-\infty, +\infty]$.

The specifications of this distribution must be defined within a **<Logistic>** XML block. This XML node accepts one attribute:

- **name**, ***required string attribute***, user-defined name of this distribution. **Note:** As with other objects, this identifier can be used to reference this specific entity from other input blocks in the XML.

This distribution can be initialized with the following children nodes:

- **<location>**, ***float, required parameter***, the distribution mean.
- **<scale>**, ***float, required parameter***, scale parameter that is proportional to the standard deviation ($\sigma^2 = \frac{1}{3}\pi^2 scale^2$).

Example:

```
<Distributions>
  ...
  <Logistic name='aUserDefinedName'>
    <location>aFloatValue</location>
    <scale>aFloatValue</scale>
  </Logistic>
  ...
</Distributions>
```

11.1.1.6 LogNormal Distribution

The **LogNormal** distribution is a distribution with the logarithm of the random variable being normally distributed. Its default support is $x \in [0, +\infty]$.

The specifications of this distribution must be defined within a **<LogNormal>** XML block. This XML node accepts one attribute:

- **name**, ***required string attribute***, user-defined name of this distribution. **Note:** As with other objects, this identifier can be used to reference this specific entity from other input blocks in the XML.

This distribution can be initialized with the following children nodes:

- **<mean>**, ***float, required parameter***, the log of the distribution mean or expected value.
- **<sigma>**, ***float, required parameter***, standard deviation.

- <low>, ***float, optional parameter***, lower domain boundary.
  Default: 0.0

Note: The <mean> and <sigma> listed above are NOT the mean and standard deviation of the distribution; they are the mean and standard deviation of the log of the distribution. Using the following notation:

- μ_ℓ: the μ parameter of the lognormal distribution, which RAVEN expects in the <mean> node;
- σ_ℓ: the σ parameter of the lognormal distribution, which RAVEN expects in the <sigma> node;
- M: the user-desired mean value of the distribution;
- S: the user-desired standard deviation of the distribution;

a conversion is defined to translate from mean M and standard deviation S into the parameters RAVEN expects:

$$\mu_\ell = \log\left(\frac{M}{\sqrt{1+\frac{S^2}{M^2}}}\right), \tag{1}$$

$$\sigma_\ell = \sqrt{\log 1 + \frac{S^2}{M^2}}. \tag{2}$$

Example:

```
<Distributions>
  ...
  <LogNormal name='aUserDefinedName'>
    <mean>aFloatValue</mean>
    <sigma>aFloatValue</sigma>
    <low>aFloatValue</low>
  </LogNormal>
  ...
</Distributions>
```

11.1.1.7 Normal Distribution

The **Normal** distribution is an extremely useful continuous distribution. Its utility is due to the central limit theorem, which states that, under mild conditions, the mean of many random variables

independently drawn from the same distribution is distributed approximately normally, irrespective of the form of the original distribution. Its default support is $x \in [-\infty, +\infty]$.

The specifications of this distribution must be defined within a **<Normal>** XML block. This XML node accepts one attribute:

- **name**, ***required string attribute***, user-defined name of this distribution. **Note:** As with other objects, this identifier can be used to reference this specific entity from other input blocks in the XML.

This distribution can be initialized with the following children nodes:

- **<mean>**, ***float, required parameter***, the distribution mean or expected value.
- **<sigma>**, ***float, required parameter***, the standard deviation.

Example:

```
<Distributions>
  ...
  <Normal name='aUserDefinedName'>
    <mean>aFloatValue</mean>
    <sigma>aFloatValue</sigma>
  </Normal>
  ...
</Distributions>
```

11.1.1.8 Triangular Distribution

The **Triangular** distribution is a continuous distribution that has a triangular shape for its PDF. Like the uniform distribution, upper and lower limits are "known," but a "best guess," of the mode or center point is also added. It has been recommended as a "proxy" for the beta distribution. Its default support is $x \in [min, max]$.

The specifications of this distribution must be defined within a **<Triangular>** XML block. This XML node accepts one attribute:

- **name**, ***required string attribute***, user-defined name of this distribution. **Note:** As with other objects, this identifier can be used to reference this specific entity from other input blocks in the XML.

This distribution can be initialized with the following children nodes:

- **<apex>**, ***float, required parameter***, peak location

- `<min>`, ***float, required parameter***, domain lower boundary.
- `<max>`, ***float, required parameter***, domain upper boundary.

Example:

```
<Distributions>
  ...
  <Triangular name='aUserDefinedName'>
    <apex>aFloatValue</apex>
    <min>aFloatValue</min>
    <max>aFloatValue</max>
  </Triangular>
  ...
</Distributions>
```

11.1.1.9 Uniform Distribution

The **Uniform** distribution is a continuous distribution with a rectangular-shaped PDF. It is often used where the distribution is only vaguely known, but upper and lower limits are known. Its default support is $x \in [lower, upper]$.

The specifications of this distribution must be defined within a `<Uniform>` XML block. This XML node accepts one attribute:

- `name`, ***required string attribute***, user-defined name of this distribution. **Note:** As with other objects, this identifier can be used to reference this specific entity from other input blocks in the XML.

This distribution can be initialized with the following children nodes:

- `<lowerBound>`, ***float, required parameter***, domain lower boundary.
- `<upperBound>`, ***float, required parameter***, domain upper boundary.

Note: Since the Uniform distribution is a rectangular-shaped PDF, the truncation does not have any effect; this is the reason why the children nodes are the ones generally used for truncated distributions. Example:

```
<Distributions>
  ...
  <Uniform name='aUserDefinedName'>
    <lowerBound>aFloatValue</lowerBound>
```

```
    <upperBound>aFloatValue</upperBound>
  </Uniform>
  ...
</Distributions>
```

11.1.1.10 Weibull Distribution

The **Weibull** distribution is a continuous distribution that is often used in the field of failure analysis; in particular, it can mimic distributions where the failure rate varies over time. If the failure rate is:

- constant over time, then $k = 1$, suggests that items are failing from random events;
- decreases over time, then $k < 1$, suggesting "infant mortality";
- increases over time, then $k > 1$, suggesting "wear out" - more likely to fail as time goes by.

Its default support is $x \in [0, +\infty)$.

The specifications of this distribution must be defined within a **<Weibull>** XML block. This XML node accepts one attribute:

- **name**, ***required string attribute***, user-defined name of this distribution. **Note:** As with other objects, this identifier can be used to reference this specific entity from other input blocks in the XML.

This distribution can be initialized with the following children nodes:

- **<k>**, ***float, required parameter***, shape parameter.
- **<lambda>**, ***float, required parameter***, scale parameter.
- **<low>**, ***float, optional parameter***, lower domain boundary.
  Default: 0.0

Example:

```
<Distributions>
  ...
  <Weibull name='aUserDefinedName'>
    <lambda>aFloatValue</lambda>
    <k>aFloatValue</k>
    <low>aFloatValue</low>
  </Weibull>
  ...
</Distributions>
```

11.1.1.11 Custom1D Distribution

The **Custom1D** distribution is a custom continuous distribution that can be initialized from a dataObject generated by RAVEN. This distribution cannot be initialized from a dataObject directly but through a .csv file. This file must contain the values of either cdf or pdf of the random variable sampled along the range of the desired random variable. In the distribution block of the RAVEN input file, the user needs to specify which file (including its working directory) needs to be used to initialize the distribution. In addition, the user is required to specify which type (cdf or pdf) or values are contained in the file and also the IDs of both the random variable and cdf/pdf. Thus the csv file contains a set of points that samples the function $pdf(x)$ or $cdf(x)$ for several values of the stochastic variable x. The user needs to specify which variable IDs correspond to x and $pdf(x)$ (or $cdf(x)$). The distribution create a fourth order spline interpolation from the provided input points. Note that the support of this distribution is set between the minimum and maximum values of the random variable which are specified in the distribution input file.

Refer to the test example ($tests/framework/test_distributionCustom1D.xml$) for more clarification.

The specifications of this distribution must be defined within a **<Custom1D>** XML block. This XML node accepts one attribute:

- **name**, ***required string attribute***, user-defined name of this distribution. **Note:** As with other objects, this identifier can be used to reference this specific entity from other input blocks in the XML.

This distribution can be initialized with the following children nodes:

- **<dataFilename>**, ***string, required parameter***, file name to be used to initialize the distribution.
- **<workingDir>**, ***string, optional parameter***, relative working directory that contains the input file.
- **<functionType>**, ***string, required parameter***, type of initialization values specified in the input file (pdf or cdf).
- **<variableID>**, ***string, required parameter***, ID of the variable contained in the input file.
- **<functionID>**, ***string, required parameter***, ID of the function associated to the variableID contained in the input file.

Example:

```
<Distributions>
  ...
```

```xml
    <Custom1D name="pdf_custom">
      <dataFilename>PointSetFile2_dump.csv</dataFilename>
      <functionID>pdf_values</functionID>
      <variableID>x</variableID>
      <functionType>pdf</functionType>
      <workingDir>custom1D/</workingDir>
    </Custom1D>
    <Custom1D name="cdf_custom">
      <dataFilename>PointSetFile3_dump.csv</dataFilename>
      <functionID>cdf_values</functionID>
      <variableID>x</variableID>
      <functionType>cdf</functionType>
      <workingDir>custom1D/</workingDir>
    </Custom1D>
  ...
</Distributions>
```

The example above initializes two distributions from two .csv files. For example, the first distribution retrieves the pdf values, located in the column with label pdf_values, for several locations of the variable located in the column with label x in the file $PointSetFile2_dump.csv$.

11.1.2 1-Dimensional Discrete Distributions.

RAVEN currently supports 3 discrete distributions. In the following paragraphs, the input requirements are reported.

11.1.2.1 Bernoulli Distribution

The **Bernoulli** distribution is a discrete distribution of the outcome of a single trial with only two results, 0 (failure) or 1 (success), with a probability of success *p*. It is the simplest building block on which other discrete distributions of sequences of independent Bernoulli trials can be based. Basically, it is the binomial distribution (k = 1, *p*) with only one trial. Its default support is $k \in 0, 1$.

The specifications of this distribution must be defined within a **<Bernoulli>** XML block. This XML node accepts one attribute:

- **name**, ***required string attribute***, user-defined name of this distribution. **Note:** As with other objects, this identifier can be used to reference this specific entity from other input blocks in the XML.

This distribution can be initialized with the following child node:

- <p>, ***float, required parameter***, probability of success.

Example:

```
<Distributions>
  ...
  <Bernoulli name='aUserDefinedName'>
    <p>aFloatValue</p>
  </Bernoulli>
  ...
</Distributions>
```

11.1.2.2 Binomial Distribution

The **Binomial** distribution is the discrete probability distribution of the number of successes in a sequence of n independent yes/no experiments, each of which yields success with probability p. Its default support is $k \in 0, 1, 2, ..., n$.

The specifications of this distribution must be defined within a **<Binomial>** XML block. This XML node accepts one attribute:

- **name**, ***required string attribute***, user-defined name of this distribution. **Note:** As with other objects, this identifier can be used to reference this specific entity from other input blocks in the XML.

This distribution can be initialized with the following children nodes:

- <p>, ***float, required parameter***, probability of success.
- <n>, ***integer, required parameter***, number of experiments.

Example:

```
<Distributions>
  ...
  <Binomial name='aUserDefinedName'>
    <n>aIntegerValue</n>
    <p>aFloatValue</p>
  </Binomial>
  ...
</Distributions>
```

11.1.2.3 Geometric Distribution

The **Geometric** distribution is a one-parameter discrete probability distribution. The distribution uses the probability p that trial will be successful. The geometric distribution gives the probability of observing k trials before the first success. Its support is $k \in 0, 1, 2, ..., n$.

The specifications of this distribution must be defined within a **<Geometric>** XML block.

This XML node accepts one attribute:

- **name**, ***required string attribute***, user-defined name of this distribution. **Note:** As with other objects, this identifier can be used to reference this specific entity from other input blocks in the XML.

This distribution can be initialized with the following children nodes:

- **<p>**, ***float, required parameter***, the success fraction for the trials.

Example:

```
<Distributions>
  ...
  <Geometric name='aUserDefinedName'>
    <p>aFloatValue</p>
  </Geometric>
  ...
</Distributions>
```

11.1.2.4 Poisson Distribution

The **Poisson** distribution is a discrete probability distribution that expresses the probability of a given number of events occurring in a fixed interval of time and/or space if these events occur with a known average rate and independently of the time since the last event. Its default support is $k \in 1, 2, 3, 4, ...$.

The specifications of this distribution must be defined within a **<Poisson>** XML block. This XML node accepts one attribute:

- **name**, ***required string attribute***, user-defined name of this distribution. **Note:** As with other objects, this identifier can be used to reference this specific entity from other input blocks in the XML.

This distribution can be initialized with the following child node:

- <mu>, ***float, required parameter***, mean rate of events/time.

Example:

```
<Distributions>
  ...
  <Poisson name='aUserDefinedName'>
    <mu>aFloatValue</mu>
  </Poisson>
  ...
</Distributions>
```

11.1.2.5 Categorical Distribution

The **Categorical** distribution is a discrete distribution that describes the result of a random variable that can have K possible outcomes. The probability of each outcome is separately specified. The possible outcomes must be only numerical values (either integer or float numbers). No string can be assigned to any outcome. There is not necessarily an underlying ordering of these outcomes, but labels are assigned in describing the distribution (in the range 1 to K). The specifications of this distribution must be defined within a <Categorical> XML block. This XML node accepts one attribute:

- name, ***required string attribute***, user-defined name of this distribution. **Note:** As with other objects, this identifier can be used to reference this specific entity from other input blocks in the XML.

This distribution can be initialized with the following child node:

- <state>, ***float, required parameter***, probability for outcome 1
  - outcome, ***float, required parameter***, outcome value.
- <state>, ***float, required parameter***, probability for outcome 2
  - outcome, ***float, required parameter***, outcome value.
- ...
- <state>, ***float, required parameter***, probability for outcome K
  - outcome, ***float, required parameter***, outcome value.

Example:

```
<Distributions>
  ...
    <Categorical name='testCategorical'>
        <state outcome="10">0.1</state>
        <state outcome="20">0.2</state>
        <state outcome="50">0.15</state>
        <state outcome="60">0.4</state>
        <state outcome="90">0.15</state>
    </Categorical>
  ...
</Distributions>
```

11.2 N-Dimensional Probability Distributions

The group of N-Dimensional distributions allow the user to model stochastic dependences between parameters. Thus instead of using N distributions for N parameters, the user can define a single distribution lying in a N-Dimensional space. The following N-Dimensional Probability Distributions are available within RAVEN:

- MultivariateNormal: Multivariate normal distribution (see Section 11.2.1)
- NDInverseWeight: ND Inverse Weight interpolation distribution (see Section 11.2.2)
- NDCartesianSpline: ND spline interpolation distribution (see Section 11.2.3)

For NDInverseWeight and NDCartesianSpline distributions, the user provides the sampled values of either CDF or PDF of the distribution. The sampled values can be scattered distributed (for NDInverseWeight) or over a cartesian grid (for NDCartesianSpline).

The user could specify, for each N-Dimensional distribution, the parameters of the random number generator function:

- **<initialGridDisc>**, ***positive integer, optional field***, user-defined initial grid discretization. This parameter specifies the number of discretizations that need to be performed, initially, for each Dimension in order to find N-Dimensional coordinate that corresponds to the CDF represented by a random number (0-1);
- **<tolerance>**, ***float, optional field***, user-defined tolerance in order to find the N-D coordinates corresponding to a random number. This tolerance is expressed in terms of CDF.

in the **<samplerInit>** block defined in sampler block **<samplerInit>** (see Section 12).

11.2.1 MultivariateNormal Distribution

the multivariate normal distribution or multivariate Gaussian distribution, is a generalization of the one-dimensional (univariate) normal distribution to higher dimensions. The multivariate normal distribution is often used to describe, at least approximately, any set of (possibly) correlated real-valued random variables each of which clusters around a mean value. The multivariate normal distribution of a k-dimensional random vector $\mathbf{x} = [x_1, x_2, \ldots, x_k]$ can be written in the following notation: $\mathbf{x} \ \sim \ \mathcal{N}(\boldsymbol\mu, \, \boldsymbol\Sigma)$ with with k-dimensional mean vector

$$\boldsymbol\mu = [E[x_1], E[x_2], \ldots, E[x_k]]$$

and $k \times k$ covariance matrix

$$\boldsymbol\Sigma = [Cov[x_i, x_j]], i = 1, 2, \ldots, k; j = 1, 2, \ldots, k$$

The probability distribution function for this distribution is the following:

$$f_{\mathbf x}(x_1,\ldots,x_k) = \frac{1}{\sqrt{(2\pi)^k|\boldsymbol\Sigma|}} \exp\left(-\frac{1}{2}({\mathbf x}-{\boldsymbol\mu})^\mathrm{T}{\boldsymbol\Sigma}^{-1}({\mathbf x}-{\boldsymbol\mu})\right),$$

The specifications of this distribution must be defined within the xml block **<MultivariateNormal>**. This XML node needs to contain the attributes:

- **name**, ***required string attribute***, user-defined identifier of this multivariate normal distribution. **Note:** As with other objects, this is the name that can be used to refer to this specific entity from other input XML blocks.
- **method**, ***required string attribute***, defines which method is used to generate the multivariate normal distribution. The only allowable methods are **'spline'** and **'pca'**.

In RAVEN the MultivariateNormal distribution can be initialized through the following keywords:

- **<mu>**, list of mean values of each dimension
- **<covariance>**, list of element values in the covariance matrix. There are two types of **<covariance>**, based on the **type**:
  - **type**, ***string, optional field***, specifies the type of covariance, the default **type** is **'abs'**. Possible values for **type** are **'abs'** and **'rel'**. **Note:** **'abs'** indicates the covariance is a normal covariance matrix, while **'rel'** indicates the covariance is a relative covariance matrix. In addition, method **'pca'** can be combined with both types, and method **'spline'** only accept the type **'abs'**
- **<transformation>**, ***XML node, optional field***, option to enable input parameter transformation using principal component analysis (PCA) approach. If this node is provided, PCA will be used to compute the principal components of input covariance matrix. The

subnode <rank> is used to indicate the number of principal components that will be used for the input transformation. The content will specify one attribute:

- <rank>, ***positive integer, required field***, user-defined dimensionality reduction.

Example:

```
<Distributions>
  ...
    <MultivariateNormal name='MultivariateNormal_test'
       method='spline'>
        <mu>0.0 60.0</mu>
        <covariance>
        1.0 0.7
        0.7 1.0
        </covariance>
    </MultivariateNormal>
    <MultivariateNormal name='MultivariateNormal_abs'
       method='pca'>
        <mu>0.0 60.0</mu>
        <covariance type='abs'>
        1.0 0.7
        0.7 1.0
        </covariance>
    </MultivariateNormal>
    <MultivariateNormal name='MultivariateNormal_rel'
       method='pca'>
        <mu>0.0 60.0</mu>
        <covariance type='rel'>
        1.0 0.7
        0.7 1.0
        </covariance>
    </MultivariateNormal>
  ...
</Distributions>
```

In the following, we defined a distribution with a transformation node using PCA method. The number of principal components is defined in <rank>. In this distribution, PCA is employed to restruct the multivariate normal distribution. In addition, the size of uncorrelated variables is also determined by <rank>.

```
<Distributions>
  ...
```

```xml
    <MultivariateNormal name='MultivariateNormal_test'
      method='pca'>
        <mu>0.0 10.0 20.0</mu>
        <covariance type="abs">
            1.0    0.7    -0.2
            0.7    1.0    0.4
            -0.2   0.4    1.0
        </covariance>
        <transformation>
          <rank>2</rank>
        </transformation>
    </MultivariateNormal>
  ...
</Distributions>
```

11.2.2 NDInverseWeight Distribution

The NDInverseWeight distribution creates a N-Dimensional distribution given a set of points scattered distributed. These points sample the PDF of the original distribution. Distribution values (PDF or CDF) are calculated using the inverse weight interpolation scheme.

The specifications of this distribution must be defined within a **<NDInverseWeight>** XML block. This XML node accepts the following attributes:

- **name**, ***required string attribute***, user-defined name of this distribution. **Note:** As with other objects, this identifier can be used to reference this specific entity from other input blocks in the XML.

In RAVEN the NDInverseWeight distribution can be initialized through the following nodes:

- **<p>**, ***float, required parameter***, power parameter. Greater values of p assign greater influence to values closest to the interpolated point.
- **<data_filename>**, ***string, required parameter***, name of the data file containing scattered values (file type '.txt').
  - **type**, ***required string attribute***, indicates if the data in indicated file is PDF or CDF.
- **<working_dir>**, ***string, required parameter***, folder location of the data file

Example:

```xml
<Distributions>
  ...
  <NDInverseWeight name='...'>
        <p>...</p>
        <dataFilename type='...'>...</dataFilename>
        <workingDir>...</workingDir>
  </NDInverseWeight>
  ...
</Distributions>
```

Each data entry contained in data_filename is listed row by row and must be listed as follows:

- number of dimensions
- number of sampled points
- ND coordinate of each sampled point
- value of each sampled point

As an example, the following shows the data entries contained in data_filename for a 3-dimensional data set that contained two sampled CDF values: ([0.0,0.0,0.0], 0.1) and ([1.0, 1.0,0.0], 0.8)

Example scattered data file:

```
3
2
0.0
0.0
0.0
1.0
1.0
0.0
0.1
0.8
```

11.2.3 NDCartesianSpline Distribution

The NDCartesianSpline distribution creates a N-Dimensional distribution given a set of points regularly distributed on a cartesian grid. These points sample the PDF of the original distribution. Distribution values (PDF or CDF) are calculated using the ND spline interpolation scheme.

The specifications of this distribution must be defined within a **<NDCartesianSpline>** XML block. This XML node accepts the following attributes:

- **name**, ***required string attribute***, user-defined name of this distribution. **Note:** As with other objects, this identifier can be used to reference this specific entity from other input blocks in the XML.

In RAVEN the NDCartesianSpline distribution can be initialized through the following nodes:

- **<data_filename>**, ***string, required parameter***, name of the data file containing scattered values (file type '.txt').
  - **type**, ***required string attribute***, indicates if the data in indicated file is PDF or CDF.
- **<working_dir>**, ***string, required parameter***, folder location of the data file

Example:

```
<Distributions>
  ...
  <NDCartesianSpline name='...'>
        <dataFilename type='...'>...</dataFilename>
        <workingDir></workingDir>
  </NDCartesianSpline>
  ...
</Distributions>
```

Each data entry contained in data _filename is listed row by row and must be listed as follows:

- number of dimensions
- number of discretization for each dimension
- discretization values for each dimension
- value of each sampled point

As an example, the following shows the data entries contained in data _filename for a 2-dimensional CDF data set on the following grid (x, y):

- first dimension (x): -0.5, 0.5
- first dimension (y): 1.0 2.0 3.0

Example scattered data file:

```
2
2
3
-0.5
0.5
1.0
2.0
3.0
CDF value of (-0.5,1.0)
CDF value of (+0.5,1.0)
CDF value of (-0.5,2.0)
CDF value of (+0.5,2.0)
CDF value of (-0.5,3.0)
CDF value of (+0.5,3.0)
```

12 Samplers

The sampler is probably the most important entity in the RAVEN framework. It performs the driving of the specific sampling strategy and, hence, determines the effectiveness of the analysis, from both an accuracy and computational point of view. The samplers, that are available in RAVEN, can be categorized into three main classes:

- **Forward** (see Section 12.1)
- **Dynamic Event Tree (DET)** (see Section 12.2)
- **Adaptive** (see Section 12.3)

Before analyzing each sampler in detail, it is important to mention that each type has a similar syntax to input the variables to be "sampled". In the example below, the variable **'variableName'** is going to be sampled by the Sampler **'whatever'** using the distribution named **'aDistribution'**.

```
<Simulation>
  ...
  <Samplers>
    ...
    <WhatEverSampler name='whatever'>
      ...
     <variable name='variableName'>
        ...
        <distribution>aDistribution</distribution>
        ...
     </variable>
      ...
    </WhatEverSampler>
    ...
  </Samplers>
  ...
</Simulation>
```

As reported in section 21, the variable naming syntax, for external driven codes, depends on the way the "code interface" has been implemented. For example, if the code has an input structure like the one reported below (YAML), the variable name might be **'I-Level|II-Level|variable'**. In this way, the relative code interface (and input parser) will know which variable needs to be perturbed and the "recipe" to access it. As reported in 21, its syntax is chosen by the developer of

the “code interface” and is implemented in the interface only (no modifications are needed in the RAVEN code).

Example YAML based Input:

```
[I-Level]
  [./II-Level]
    variable = xxx
  [../]
[]
```

Example XML block to define the variables and associated distributions:

```
<variable name='I-Level|II-Level|variable'>
  <distribution>exampleDistribution</distribution>
</variable>
```

If the variable is associated to a multi-dimensional ND distribution, it is needed to specify which dimension of the ND distribution is associated to such variable. An example is shown below: the variable “variableX” is associated to the third dimension of the ND distribution “ND-distribution”.

```
<variable name='variableX'>
     <distribution dim='3'>NDdistribution</distribution>
</variable>
```

For most codes, it is prudent that there are no redundant inputs; however there are cases where this is not reality. For example, if there is a variable **'inner_radius'** and a variable **'outer_radius'**, there may be a third variable **'thickness'** that is actually derived from the previous two, as **'thickness'** = **'outer_radius'** - **'inner_radius'**. RAVEN supports this type of redundant input through a Function entity. In this case, instead of a **<distribution>** node in the **<variable>** block, there is a **<function>** node, specifying the name of the function (defined in the **<Functions>** block). In order to work properly, this function must have a method named “evaluate” that returns a single python float object. In this way, multiple variables can be associated with the same function. For example,

```
...
<Functions>
  <External name='torus_calcs' file='torus_calcs.py'>
    <variable>outer_radius</variable>
    <variable>inner_radius</variable>
  </External>
<Functions>
...
```

```xml
<Samplers>
  <WhatEverSampler name='myExampleSampler'>
    <variable name='inner_radius'>
      <distribution>inner_dist</distribution>
    </variable>
    <variable name='outer_radius'>
      <distribution>outer_dist</distribution>
    </variable>
    <variable name='thickness'>
      <function>torus_calcs</function>
    </variable>
  </WhatEverSampler>
</Samplers>
```

The corresponding function file **'torus_calcs.py'** needs the following method:

```python
def evaluate(self):
  return self.outer_radius - self.inner_radius
```

The **'thickness'** parameter will still be treated as an input for the sake of csv printing and DataObjects storage.
Note: It is important to notice that if the user use variables with no-Python compatible names (e.g. parenthesis, etc.), the **<alias>** system needs to be used to alias the variables.

In the sampler class a special node exists: the **<sampler_init>** node. This node contains specific parameters that characterize each particular sampler. In addition, **<sampler_init>** might contain the information regarding the random generator function for each N-Dimensional distribution (specified in the **<dist_init>** node):

- initial_grid_disc
- tolerance

An example of **<dist_init>** node is provided below:

```xml
<distInit>
    <distribution name= 'ND_dist_name'>
         <initialGridDisc>5</initialGridDisc>
          <tolerance>0.2</tolerance>
     </distribution>
  </distInit>
```

In the **<sampler_init>** node it is possible to add also the subnode **<globalGrid>**. The **<globalGrid>** can be used in two cases:

- 1D distributions: an identical grid that is associated to several distributions
- ND distribution: a grid associated to a single ND distribution. This is the case when a stratified sampling is performed on the CDF of an ND distribution: the **<globalGrid>** is shared among the variables associated to the Nd distribution

12.1 Forward Samplers

The Forward sampler category collects all the strategies that perform the sampling of the input space without exploiting, through dynamic learning approaches, the information made available from the outcomes of calculations previously performed (adaptive sampling) and the common system evolution (patterns) that different sampled calculations can generate in the phase space (dynamic event tree). In the RAVEN framework, several different "Forward" samplers are available:

- **Monte Carlo (MC)**
- **Stratified**
- **Grid Based**
- **Sparse Grid Collocation**
- **Sobol Decomposition**
- **Response Surface Design of Experiment**
- **Factorial Design of Experiment**
- **Ensemble Forward Sampling strategy**
- **Custom Sampling strategy**

From a practical point of view, these sampling strategies represent different ways to explore the input space. In the following paragraphs, the input requirements and a small explanation of the different sampling methodologies are reported.

12.1.1 Monte Carlo

The **Monte-Carlo** sampling approach is one of the most well-known and widely used approaches to perform exploration of the input space. The main idea behind MonteCarlo sampling is to randomly perturb the input space according to uniform or parameter-based probability density functions.

The specifications of this sampler must be defined within a **<MonteCarlo>** XML block. This XML node accepts the following attributes:

- **name**, *required string attribute*, user-defined name of this Sampler. N.B. As for the other objects, this is the name that can be used to refer to this specific entity from other input blocks (xml);

In the **MonteCarlo** input block, the user needs to specify the variables need to be sampled. As already mentioned, these variables are inputted within consecutive xml blocks called **<variable>**. In addition, the settings for this sampler need to be specified in the **<samplerInit>** XML block:

- **<samplerInit>**, ***XML node, required parameter***. In this xml-node,the following xml sub-nodes need to be specified:
  - **<limit>**, ***integer,required field***, number of MonteCarlo samples needs to be generated;
  - **<initialSeed>**, ***integer, optional field***, initial seeding of random number generator
  - **<reseedEachIteration>**, ***boolean/string(case insensitive), optional field***, perform a re-seeding for each sample generated (True values = True, yes, y, t).
    Default: False;
  - **<distInit>**, ***integer, optional field***, in this node the user specifies the initialization of the random number generator function for each N-Dimensional Probability Distributions (see Section 11.2).
  - **<samplingType>**, ***string, optional field***, sub-type of sampling
    Default: None. the user can choose to perform a Monte-Carlo sampling where the location of the samples in the input space is uniformly distributed and not generated accordingly to the specific set of distributions. This can be specificed in the **<samplingType>** with the kewyword "uniform". This option works only if all the distributions have an upper and lower bound specified (i.e., **<lowerBound>** and **<upperBound>**). Allowed fields for this node are "None" and "uniform".

- **<variable>**, ***XML node, required parameter*** will specify one attribute:
  - name, ***required string attribute***, user-defined name of this variable.

  This **<variable>** recognizes the following child nodes:

- **<distribution>**, ***string, required field***, name of the distribution that is associated to this variable. Its name needs to be contained in the **<Distributions>** block explained in Section 11. In addition, if NDDistribution is used, the attribute **dim** is required. **Note:** Alternatively, this node must be omitted if the **<function>** node is supplied.
- **<function>**, ***string, required field***, name of the function that defines the calculation of this variable from other distributed variables. Its name needs to be contained in the **<Functions>** block explained in Section 18. This function must implement a method named "evaluate". **Note:** Alternatively, this node must be ommitted if the **<distribution>** node is supplied.

- **<constant>**, ***XML node, optional parameter*** the user is able to input variables that need to be kept constant. For doing this, as many **<constant>** nodes as needed can be input, where the body of the node contains the constant value that is going to be injected as an additional variable

If the input parameters are correlated, the **MonteCarlo** sampling approach can be also used if the user specified a multivariate distributions inside the **<Distributions>** (see Section 11.2). Furthermore, if the covariance matrix is provided and the input parameters is assumed to have the multivariate normal distribution, one can also use **MonteCarlo** approach to sample the input parameters in the transformed space (aka subspace, reduced space). If this is the case, the user needs to provide additional information, i.e. the **<transformation>** under **<MultivariateNormal>** of **<Distributions>** (more information can be found in Section 11.2). In addition, the node **<variablesTransformation>** is also required for **MonteCarlo** sampling. This node is used to tranform the variables specified by **<latentVariables>** in the transformed space of input into variables spefified by **<manifestVariables>** in the input space. The variables listed in **<latentVariables>** should be predefined in **<variable>**, and the variables listed in **<manifestVariables>** are used by the **<Models>**.

- **<variablesTransformation>**, ***optional field***. this XML node accepts one attribute:
  - **distribution**, ***required string attribute***, the name for the distribution defined in the XML node **<Distributions>**. This attribute indicates the values of **<manifestVariables>** are drawn from **distribution**.

  In addition, this XML node also accepts three childen nodes:

  - **<latentVariables>**, ***comma separated string, required field***, user-defined latent variables that are used for the variables transformation. All the variables listed under this node should be also mentioned in **<variable>**.
  - **<manifestVariables>**, ***comma separated string, required field***, user-defined manifest variables that can be used by the **model**.

- **<manifestVariablesIndex>**, ***comma separated string, optional field***, user-defined manifest variables indices paired with **<manifestVariables>**. These indices indicate the position of manifest variables associated with multivariate normal distribution defined in the XML node **<Distributions>**. The indices should be postive integer. If not provided, the code will use the positions of manifest variables listed in **<manifestVariables>** as the indices.
- **<method>**, ***string, required field***, the method that is used for the variables transformation. The currently available method is '**pca**'.

Assembler Objects These objects are either required or optional depending on the functionality of the MonteCarlo Sampler. The objects must be listed with a rigorous syntax that, except for the XML node tag, is common among all the objects. Each of these nodes must contain 2 attributes that are used to identify them within the simulation framework:

- **class**, ***required string attribute***, the main "class" of the listed object. For example, it can be **'Models'**, **'Functions'**, etc.
- **type**, ***required string attribute***, the object identifier or sub-type. For example, it can be **'ROM'**, **'External'**, etc.

The **MonteCarlo** approach requires or optionally accepts the following object types:

- **<Restart>**, ***string, optional field***, the body of this XML node must contain the name of an appropriate **DataObject** defined in the **<DataObjects>** block (see Section 14). It is used as a "restart" tool, where it accepts pre-existing solutions in the PointSet instead of recalculating solutions.

The following node is an additional option when a restart DataObject is provided:

- **<restartTolerance>**, ***float, optional field***, the body of this XML node must contain a valid floating point value. If a **<Restart>** node is supplied for this **<Sampler>**, this node offers a way to determine how strictly matching points are determined. Given a point in the input space, if that point is within a relative Euclidean distance (equal to the tolerance) of a restart point, the nearest restart point will be used.
  Default: 1e-15

Example:

```xml
<Samplers>
  ...
  <MonteCarlo name='MCname'>
    <samplerInit>
      <limit>10</limit>
      <initialSeed>200286</initialSeed>
      <reseedEachIteration>false</reseedEachIteration>
      <distInit>
        <distribution name= 'ND_InverseWeight_P'>
          <initialGridDisc>10</initialGridDisc>
          <tolerance>0.2</tolerance>
        </distribution>
      </distInit>
    </samplerInit>
    <variable name='var1'>
      <distribution>aDistributionNameDefinedInDistributionBlock
      </distribution>
    </variable>
    <Restart class='DataObject' type='PointSet'>data</Restart>
  </MonteCarlo>
  ...
</Samplers>
  ...
  <PointSet name="data">
    <Input>var1</Input>
    <Output>ans</Output>
  </PointSet>
  ...
```

12.1.2 Grid

The **Grid** sampling approach is probably the simplest exploration approach that can be employed to explore an uncertain domain. The idea is to construct an N-dimensional grid where each dimension is represented by one uncertain variable. This approach performs the sampling at each node of the grid. The sampling of the grid consists in evaluating the answer of the system under all possible combinations among the different variables' values with respect to a predefined discretization metric. In RAVEN two discretization metrics are available: 1) cumulative distribution function, and 2) value. Thus, the grid meshing can be input via probability or variable values. Regarding the N-dimensional distributions, the user can specify for each dimension the type of grid to be used (i.e., value or CDF). Note the discretization of the CDF, only for the grid sampler, is performed on

the marginal distribution for the specific variable considered.

The specifications of this sampler must be defined within a **<Grid>** XML block. This XML node accepts one attribute:

- **name**, ***required string attribute***, user-defined name of this sampler. **Note:** As with other objects, this identifier can be used to reference this specific entity from other input blocks in the XML.

In the **<Grid>** input block, the user needs to specify the variables to sample. As already mentioned, these variables are specified within consecutive **<variable>** XML blocks:

- **<variable>**, ***XML node, required parameter*** will specify one attribute:
  - **name**, ***required string attribute***, user-defined name of this variable.

  This **<variable>** recognizes the following child nodes:
  - **<distribution>**, ***string, required field***, name of the distribution that is associated to this variable. Its name needs to be contained in the **<Distributions>** block explained in Section 11. In addition, if NDDistribution is used, the attribute **dim** is required. **Note:** Alternatively, this node must be omitted if the **<function>** node is supplied.
  - **<function>**, ***string, required field***, name of the function that defines the calculation of this variable from other distributed variables. Its name needs to be contained in the **<Functions>** block explained in Section 18. This function must implement a method named "evaluate". **Note:** Alternatively, this node must be ommitted if the **<distribution>** node is supplied.
  - **<grid>**, ***space separated floats, required field***, the content of this XML node depends on the definition of the associated attributes:
    * **type**, ***required string attribute***, user-defined discretization metric type: 1) **'CDF'**, the grid will be specified based on cumulative distribution function probability thresholds, and 2) **'value'**, the grid will be provided using variable values.
    * **construction**, ***required string attribute***, how the grid needs to be constructed, independent of its type (i.e. **'CDF'** or **'value'**).

    Based on the **construction** type, the content of the **<grid>** XML node and the requirements for other attributes change:
    * **construction='equal'**. The grid is going to be constructed equally-spaced (**type='value'**) or equally probable (**type='CDF'**). This construction type requires the definition of additional attributes:
      · **steps**, ***required integer attribute***, number of equally spaced/probable discretization steps.

This construction type requires that the content of the <grid> node represents the lower and upper bounds (either in probability or value). Two values need to be specified; the lowest one will be considered as the $lowerBound$, the largest, the $upperBound$. The lower and upper bounds are checked against the associated <distribution> bounds. If one or both of them falls outside the distribution's bounds, the code will raise an error. The $stepSize$ is determined as follows:
$stepSize = (upperBound - lowerBound)/steps$

  * `construction='custom'`. The grid will be directly specified by the user. No additional attributes are needed. This construction type requires that the <grid> node contains the actual mesh bins. For example, if the grid `type` is `'CDF'`, in the body of <grid>, the user will specify the CDF probability thresholds (nodalization in probability). All the bins are checked against the associated <distribution> bounds. If one or more of them falls outside the distribution's bounds, the code will raise an error.

Note: The <grid> node is only required if a <distribution> node is supplied. In the case of a <function> node, no grid information is requested.

- <constant>, ***XML node, optional parameter*** the user is able to input variables that need to be kept constant. For doing this, as many <constant> nodes as needed can be input, where the body of the node contains the constant value that is going to be injected as an additional variable

If the input parameters are correlated, the **Grid** sampling approach can be also used if the user specified a multivariate distributions inside the <Distributions> (see Section 11.2). Furthermore, if the covariance matrix is provided and the input parameters is assumed to have the multivariate normal distribution, one can also use **Grid** approach to sample the input parameters in the transformed space (aka subspace, reduced space). This means one creates the grids of variables listed by <latentVariables> in the transformed space. If this is the case, the user needs to provide additional information, i.e. the <transformation> under <MultivariateNormal> of <Distributions> (more information can be found in Section 11.2). In addition, the node <variablesTransformation> is also required for **Grid** sampling. This node is used to tranform the variables specified by <latentVariables> in the transformed space of input into variables spefified by <manifestVariables> in the input space. The variables listed in <latentVariables> should be predefined in <variable>, and the variables listed in <manifestVariables> are used by the <Models>.

- <variablesTransformation>, ***optional field***. this XML node accepts one attribute:

  - `distribution`, ***required string attribute***, the name for the distribution defined in the XML node <Distributions>. This attribute indicates the values of <manifestVariables> are drawn from `distribution`.

  In addition, this XML node also accepts three childen nodes:

- <latentVariables>, ***comma separated string, required field***, user-defined latent variables that are used for the variables transformation. All the variables listed under this node should be also mentioned in <variable>.
- <manifestVariables>, ***comma separated string, required field***, user-defined manifest variables that can be used by the model.
- <manifestVariablesIndex>, ***comma separated string, optional field***, user-defined manifest variables indices paired with <manifestVariables>. These indices indicate the position of manifest variables associated with multivariate normal distribution defined in the XML node <Distributions>. The indices should be postive integer. If not provided, the code will use the positions of manifest variables listed in <manifestVariables> as the indices.
- <method>, ***string, required field***, the method that is used for the variables transformation. The currently available method is '**pca**'.

Assembler Objects These objects are either required or optional depending on the functionality of the Grid Sampler. The objects must be listed with a rigorous syntax that, except for the XML node tag, is common among all the objects. Each of these nodes must contain 2 attributes that are used to identify them within the simulation framework:

- class, ***required string attribute***, the main "class" of the listed object. For example, it can be 'Models', 'Functions', etc.
- type, ***required string attribute***, the object identifier or sub-type. For example, it can be 'ROM', 'External', etc.

The **Grid** approach requires or optionally accepts the following object types:

- <Restart>, ***string, optional field***, the body of this XML node must contain the name of an appropriate **DataObject** defined in the <DataObjects> block (see Section 14). It is used as a "restart" tool, where it accepts pre-existing solutions in the PointSet instead of recalculating solutions.

The following node is an additional option when a restart DataObject is provided:

- <restartTolerance>, ***float, optional field***, the body of this XML node must contain a valid floating point value. If a <Restart> node is supplied for this <Sampler>, this node offers a way to determine how strictly matching points are determined. Given a point in the input space, if that point is within a relative Euclidean distance (equal to the tolerance) of a restart point, the nearest restart point will be used.
  Default: 1e-15

Example:

```
<Samplers>
  ...
  <Grid name='Gridname'>
    <variable name='var1'>
      <distribution>aDistributionNameDefinedInDistributionBlock1
      </distribution>
      <grid type='value' construction='equal' steps='100' >0.2
         10</grid>
    </variable>
    <variable name='var2'>
      <distribution>aDistributionNameDefinedInDistributionBlock2
      </distribution>
      <grid type='CDF' construction='equal' steps='5' >0.2
         0.8</grid>
    </variable>
    <variable name='var3'>
      <distribution>aDistributionNameDefinedInDistributionBlock3
      </distribution>
      <grid type='value' construction='equal' steps='100' >0.2
         21.0</grid>
    </variable>
    <variable name='var4'>
      <distribution>aDistributionNameDefinedInDistributionBlock4
      </distribution>
      <grid type='CDF' construction='equal' steps='5' >0.2
         1.0</grid>
    </variable>
    <variable name='var5'>
      <distribution>aDistributionNameDefinedInDistributionBlock5
      </distribution>
      <grid type='value' construction='custom'>0.2 0.5
         10.0</grid>
    </variable>
    <variable name='var6'>
      <distribution>aDistributionNameDefinedInDistributionBlock6
      </distribution>
      <grid type='CDF' construction='custom'>0.2 0.5 1.0</grid>
    </variable>
    <Restart class='DataObject' type='PointSet'>data</Restart>
    <restartTolerance>1e-6</restartTolerance>
  </Grid>
```

```
  ...
</Samplers>
  ...
  <PointSet name="data">
    <Input>var1,var2,var3,var4,var5,var6</Input>
    <Output>ans</Output>
  </PointSet>
  ...
```

Note: A restart example is included here but is not necessary in general.

12.1.3 Sparse Grid Collocation

Sparse Grid Collocation builds on generic **Grid** sampling by selecting evaluation points based on characteristic quadratures as part of stochastic collocation for generalized polynomial chaos uncertainty quantification. In collocation you construct an N-dimensional grid, with each uncertain variable providing an axis. Along each axis, the points of evaluation correspond to quadrature points necessary to integrate polynomials (see 17.3.3). In the simplest (and most naive) case, a N-Dimensional tensor product of all possible combinations of points from each dimension's quadrature is constructed as sampling points. The number of necessary samples can be reduced by employing Smolyak-like sparse grid algorithms, which use reduced combinations of polynomial orders to reduce the necessary sampling space. The specifications of this sampler must be defined within a **<SparseGridCollocation>** XML block. .

- **name**, ***required string attribute***, user-defined name of this sampler. **Note:** As with other objects, this identifier can be used to reference this specific entity from other input blocks in the XML.
- **parallel**, ***optional string attribute***, option to disable parallel construction of the sparse grid. Because of increasing computational expense with increasing input space dimension, RAVEN will default to parallel construction of the sparse grid.
- **outfile**, ***optional string attribute***, option to allow the generated sparse grid points and weights to be printed to a file with the given name.
  Default: True

In the **<SparseGridCollocation>** input block, the user needs to specify the variables to sample. As already mentioned, these variables are specified within consecutive **<variable>** XML blocks:

- **<variable>**, ***XML node, required parameter*** will specify one attribute:

- `name`, ***required string attribute***, user-defined name of this variable.

In the variable node, the following xml-node needs to be specified:

- `<distribution>`, ***string, required field***, name of the distribution that is associated to this variable. Its name needs to be contained in the `<Distributions>` block explained in Section 11. In addition, if NDDistribution is used, the attribute `dim` is required. **Note:** Alternatively, this node must be omitted if the `<function>` node is supplied.
- `<function>`, ***string, required field***, name of the function that defines the calculation of this variable from other distributed variables. Its name needs to be contained in the `<Functions>` block explained in Section 18. This function must implement a method named "evaluate". **Note:** Alternatively, this node must be ommitted if the `<distribution>` node is supplied.

- `<constant>`, ***XML node, optional parameter*** the user is able to input variables that need to be kept constant. For doing this, as many `<constant>` nodes as needed can be input, where the body of the node contains the constant value that is going to be injected as an additional variable

Because of the tight coupling between the Sampler and the ROM in stochastic collocation for generalized polynomial chaos, the Sampler needs access to the ROM via the assembler do determine the polynomials, quadratures, and importance weights to use in each dimension (see 17.3.3).

Assembler Objects These objects are either required or optional depending on the functionality of the SparseGridCollocation Sampler. The objects must be listed with a rigorous syntax that, except for the XML node tag, is common among all the objects. Each of these nodes must contain 2 attributes that are used to identify them within the simulation framework:

- `class`, ***required string attribute***, the main "class" of the listed object. For example, it can be `'Models'`, `'Functions'`, etc.
- `type`, ***required string attribute***, the object identifier or sub-type. For example, it can be `'ROM'`, `'External'`, etc.

The **SparseGridCollocation** approach requires or optionally accepts the following object types:

- `<ROM>`, ***string, required field***, the body of this XML node must contain the name of an appropriate ROM defined in the `<Models>` block (see Section 17.3).
- `<Restart>`, ***string, optional field***, the body of this XML node must contain the name of an appropriate **DataObject** defined in the `<DataObjects>` block (see Section 14). It is

used as a "restart" tool, where it accepts pre-existing solutions in the PointSet instead of recalculating solutions.

The following node is an additional option when a restart DataObject is provided:

- **<restartTolerance>**, ***float, optional field***, the body of this XML node must contain a valid floating point value. If a **<Restart>** node is supplied for this **<Sampler>**, this node offers a way to determine how strictly matching points are determined. Given a point in the input space, if that point is within a relative Euclidean distance (equal to the tolerance) of a restart point, the nearest restart point will be used.
  Default: 1e-15

```
Example:
<Samplers>
  ...
  <SparseGridCollocation name="mySG" parallel="0">
    <variable name="x1">
      <distribution>myDist1</distribution>
    </variable>
    <variable name="x2">
      <distribution>myDist2</distribution>
    </variable>
    <ROM class = 'Models' type = 'ROM' >SCROM</ROM>
    <Restart class = 'DataObjects' type = 'PointSet' >solns</Restart>
  </SparseGridCollocation>
  ...
</Samplers>
  ...
  <PointSet name="solns">
    <Input>x1,x2</Input>
    <Output>y</Output>
  </PointSet>
  ...
```

In general, **SparseGridCollocation** requires uncorrelated input parameters. If the input parameters are correlated, one can transform the correlated parameters into uncorrelated parameters; the **SparseGridCollocation** can also be used with the uncorrelated parameters in the transformed space. Like in the **Grid** sampler, if the covariance matrix is provided and the input parameters are assumed to have the multivariate normal distribution, the **SparseGridCollocation** can be used. This means one creates the sparse grids of variables listed by **<latentVariables>** in the transformed space. If this is the case, the user needs to provide additional information, i.e. the **<transformation>** under **<MultivariateNormal>** of **<Distributions>** (more information can be found in Section 11.2). In addition, the node **<variablesTransformation>** is also required for **SparseGridCollocation** sampler. This node is used to tranform the variables specified by **<latentVariables>** in the transformed space of input into variables spefified by

<manifestVariables> in the input space. The variables listed in <latentVariables> should be predefined in <variable>, and the variables listed in <manifestVariables> are used by the <Models>.

- <variablesTransformation>, *optional field*. this XML node accepts one attribute:
  - distribution, ***required string attribute***, the name for the distribution defined in the XML node <Distributions>. This attribute indicates the values of <manifestVariables> are drawn from distribution.

  In addition, this XML node also accepts three childen nodes:
  - <latentVariables>, ***comma separated string, required field***, user-defined latent variables that are used for the variables transformation. All the variables listed under this node should be also mentioned in <variable>.
  - <manifestVariables>, ***comma separated string, required field***, user-defined manifest variables that can be used by the model.
  - <manifestVariablesIndex>, ***comma separated string, optional field***, user-defined manifest variables indices paired with <manifestVariables>. These indices indicate the position of manifest variables associated with multivariate normal distribution defined in the XML node <Distributions>. The indices should be postive integer. If not provided, the code will use the positions of manifest variables listed in <manifestVariables> as the indices.
  - <method>, ***string, required field***, the method that is used for the variables transformation. The currently available method is '**pca**'.

```
...
<Models>
    ...
    <ExternalModel ModuleToLoad="lorentzAttractor_noK"
       name="PythonModule" subType="">
        <variables>sigma,rho,beta,x,y,z,time,z0,y0,z0</variables>
    </ExternalModel>
    <ROM name="SCROM" subType="GaussPolynomialRom">
        <Target>and</Target>
        <Features>x1,y1,z1</Features>
        <IndexSet>TensorProduct</IndexSet>
        <PolynomialOrder>1</PolynomialOrder>
    </ROM>
    ...
</Models>
```

```xml
<Distributions>
    ...
    <MultivariateNormal name='MVNDist' method='pca'>
        <transformation>
            <rank>3</rank>
        </transformation>
        <mu>0.0 1.0 2.0</mu>
        <covariance type="abs">
            1.0         0.6        -0.4
            0.6         1.0        0.2
            -0.4        0.2        0.8
        </covariance>
    </MultivariateNormal>
    ...
</Distributions>

<Samplers>
  ...
  <SparseGridCollocation name='SC'>
        <variable name='x0'>
            <distribution dim='1'>MVNDist</distribution>
        </variable>
        <variable name='y0'>
            <distribution dim='2'>MVNDist</distribution>
        </variable>
        <variable name='z0'>
            <distribution dim='3'>MVNDist</distribution>
        </variable>
        <variablesTransformation model="PythonModule">
            <latentVariables>x1,y1,z1</latentVariables>
            <manifestVariables>x0,y0,z0</manifestVariables>
            <method>pca</method>
        </variablesTransformation>
        <ROM class = 'Models' type = 'ROM' >SCROM</ROM>
        <Restart class="DataObjects"
          type="PointSet">solns</Restart>
  </SparseGridCollocation>
  ...
</Samplers>
  ...
  <PointSet name="solns">
    <Input>x0,y0,z0</Input>
```

```
    <Output>ans</Output>
  </PointSet>
  ...
```

12.1.4 Sobol

The **Sobol** sampler uses high-density model reduction (HDMR) a.k.a. Sobol decomposition to approximate a function as the sum of increasing-complexity interactions. At its lowest level (order 1), it treats the function as a sum of the reference case plus a functional of each input dimesion separately. At order 2, it adds functionals to consider the pairing of each dimension with each other dimension. The benefit to this approach is considering several functions of small input cardinality instead of a single function with large input cardinality. This allows reduced order models like generalized polynomial chaos (see 17.3.3) to approximate the functionals accurately with few computations runs. This Sobol sampler uses the associated HDMRRom (see 17.3.4) to determine at what points the input space need be evaluated. Since Sobol sampler relies on SparseGridCollocation, it is also compatible with multivariate normal distribution objects. The `<Sobol>` node supports the following attributes:

- `name`, ***required string attribute***, user-defined name of this sampler. **Note:** As with other objects, this identifier can be used to reference this specific entity from other input blocks in the XML.
- `parallel`, ***optional string attribute***, option to disable parallel construction of the sparse grid. Because of increasing computational expense with increasing input space dimension, RAVEN will default to parallel construction of the sparse grid.
  Default: True

In the `<Sobol>` input block, the user needs to specify the variables to sample. As already mentioned, these variables are specified within consecutive `<variable>` XML blocks:

- `<variable>`, ***XML node, required parameter*** will specify one attribute:
  - `name`, ***required string attribute***, user-defined name of this variable.

  In the variable node, the following xml-node needs to be specified:

  - `<distribution>`, ***string, required field***, name of the distribution that is associated to this variable. Its name needs to be contained in the `<Distributions>` block explained in Section 11. In addition, if NDDistribution is used, the attribute `dim` is required. **Note:** Alternatively, this node must be omitted if the `<function>` node is supplied.

- <function>, ***string, required field***, name of the function that defines the calculation of this variable from other distributed variables. Its name needs to be contained in the <Functions> block explained in Section 18. This function must implement a method named "evaluate". **Note:** Alternatively, this node must be ommitted if the <distribution> node is supplied.

- <constant>, ***XML node, optional parameter*** the user is able to input variables that need to be kept constant. For doing this, as many <constant> nodes as needed can be input, where the body of the node contains the constant value that is going to be injected as an additional variable

Like the **SparseGridCollocation**, if multivariate normal distribution is provided, the following node need to be specified:

- <variablesTransformation>, ***optional field***. this XML node accepts one attribute:
  - distribution, ***required string attribute***, the name for the distribution defined in the XML node <Distributions>. This attribute indicates the values of <manifestVariables> are drawn from distribution.

  In addition, this XML node also accepts three childen nodes:
  - <latentVariables>, ***comma separated string, required field***, user-defined latent variables that are used for the variables transformation. All the variables listed under this node should be also mentioned in <variable>.
  - <manifestVariables>, ***comma separated string, required field***, user-defined manifest variables that can be used by the model.
  - <manifestVariablesIndex>, ***comma separated string, optional field***, user-defined manifest variables indices paired with <manifestVariables>. These indices indicate the position of manifest variables associated with multivariate normal distribution defined in the XML node <Distributions>. The indices should be postive integer. If not provided, the code will use the positions of manifest variables listed in <manifestVariables> as the indices.
  - <method>, ***string, required field***, the method that is used for the variables transformation. The currently available method is '**pca**'.

Because of the tight coupling between the Sobol sampler and the HDMRRom, the Sampler needs access to the ROM via the assembler do determine the polynomials, quadratures, Sobol order, and importance weights to use in each dimension (see 17.3.4).

Assembler Objects These objects are either required or optional depending on the functionality of the Sobol Sampler. The objects must be listed with a rigorous syntax that, except for the

XML node tag, is common among all the objects. Each of these nodes must contain 2 attributes that are used to identify them within the simulation framework:

- **class**, ***required string attribute***, the main "class" of the listed object. For example, it can be **'Models'**, **'Functions'**, etc.
- **type**, ***required string attribute***, the object identifier or sub-type. For example, it can be **'ROM'**, **'External'**, etc.

The **Sobol** approach requires or optionally accepts the following object types:

- **<ROM>**, ***string, required field***, the body of this XML node must contain the name of an appropriate ROM defined in the **<Models>** block (see Section 17.3).

- **<Restart>**, ***string, optional field***, the body of this XML node must contain the name of an appropriate **DataObject** defined in the **<DataObjects>** block (see Section 14). It is used as a "restart" tool, where it accepts pre-existing solutions in the PointSet instead of recalculating solutions.

The following node is an additional option when a restart DataObject is provided:

- **<restartTolerance>**, ***float, optional field***, the body of this XML node must contain a valid floating point value. If a **<Restart>** node is supplied for this **<Sampler>**, this node offers a way to determine how strictly matching points are determined. Given a point in the input space, if that point is within a relative Euclidean distance (equal to the tolerance) of a restart point, the nearest restart point will be used.
  Default: 1e-15

```
Example:
<Samplers>
  ...
  <Sobol name="mySobol" parallel="0">
    <variable name="x1">
      <distribution>myDist1</distribution>
    </variable>
    <variable name="x2">
      <distribution>myDist2</distribution>
    </variable>
    <ROM class = 'Models' type = 'ROM' >myHDMR</ROM>
    <Restart class="DataObjects" type="PointSet">solns</Restart>
  </Sobol>
  ...
```

```
</Samplers>
  ...
  <PointSet name="solns">
    <Input>x1,y2</Input>
    <Output>ans</Output>
  </PointSet>
  ...
```

12.1.5 Stratified

The **Stratified** sampling approach is a method for the exploration of the input space that consists of dividing the uncertain domain into subgroups before sampling. In the "stratified" sampling, these subgroups must be:

- mutually exclusive: every element in the population must be assigned to only one stratum (subgroup);
- collectively exhaustive: no population element can be excluded.

Then simple random sampling or systematic sampling is applied within each stratum. It is worthwhile to note that the well-known Latin hypercube sampling represents a specialized version of the stratified approach, when the domain strata are constructed in equally-probable CDF bins.

The specifications of this sampler must be defined within a `<Stratified>` XML block. This XML node accepts one attribute:

- `name`, ***required string attribute***, user-defined name of this sampler. **Note:** As with other objects, this identifier can be used to reference this specific entity from other input blocks in the XML.

In the `<Stratified>` input block, the user needs to specify the variables to sample. As already mentioned, these variables are specified within consecutive `<variable>` XML blocks:

- `<variable>`, ***XML node, required parameter*** will specify one attribute:
  - `name`, ***required string attribute***, user-defined name of this variable.

  This `<variable>` recognizes the following child nodes:

  - `<distribution>`, ***string, required field***, name of the distribution that is associated to this variable. Its name needs to be contained in the `<Distributions>` block

explained in Section 11. In addition, if NDDistribution is used, the attribute `dim` is required. **Note:** Alternatively, this node must be omitted if the `<function>` node is supplied.

  - `<function>`, ***string, required field***, name of the function that defines the calculation of this variable from other distributed variables. Its name needs to be contained in the `<Functions>` block explained in Section 18. This function must implement a method named "evaluate". **Note:** Alternatively, this node must be ommitted if the `<distribution>` node is supplied.
  - `<grid>`, ***space separated floats, required field***, the content of this XML node depends on the definition of the associated attributes:
    * `type`, ***required string attribute***, user-defined discretization metric type: 1) `'CDF'`, the grid will be specified based on cumulative distribution function probability thresholds, and 2) `'value'`, the grid will be provided using variable values.
    * `construction`, ***required string attribute***, how the grid needs to be constructed, independent of its type (i.e. `'CDF'` or `'value'`).

    Based on the `construction` type, the content of the `<grid>` XML node and the requirements for other attributes change:

    * `construction='equal'`. The grid is going to be constructed equally-spaced (`type='value'`) or equally probable (`type='CDF'`). This construction type requires the definition of additional attributes:
      · `steps`, ***required integer attribute***, number of equally spaced/probable discretization steps.

      This construction type requires that the content of the `<grid>` node represents the lower and upper bounds (either in probability or value). Two values need to be specified; the lowest one will be considered as the $lowerBound$, the largest, the $upperBound$. The lower and upper bounds are checked against the associated `<distribution>` bounds. If one or both of them falls outside the distribution's bounds, the code will raise an error. The $stepSize$ is determined as follows:
      $stepSize = (upperBound - lowerBound)/steps$
    * `construction='custom'`. The grid will be directly specified by the user. No additional attributes are needed. This construction type requires that the `<grid>` node contains the actual mesh bins. For example, if the grid `type` is `'CDF'`, in the body of `<grid>`, the user will specify the CDF probability thresholds (nodalization in probability). All the bins are checked against the associated `<distribution>` bounds. If one or more of them falls outside the distribution's bounds, the code will raise an error.

    Note: The `<grid>` node is only required if a `<distribution>` node is supplied. In the case of a `<function>` node, no grid information is requested.

- `<constant>`, ***XML node, optional parameter*** the user is able to input variables that need to be kept constant. For doing this, as many `<constant>` nodes as needed can be input,

where the body of the node contains the constant value that is going to be injected as an additional variable

In addition, the settings for this sampler need to be specified in the `<samplerInit>` XML block:

- `<samplerInit>`, ***XML node, required parameter***. In this xml-node,the following xml sub-nodes need to be specified:
  - `<initialSeed>`, ***integer, optional field***, initial seeding of random number generator
  - `<distInit>`, ***integer, optional field***, in this node the user specifies the initialization of the random number generator function for each N-Dimensional Probability Distributions (see Section 11.2).

As one can see, the input specifications for the **Stratified** sampler are similar to that of the **Grid** sampler. It is important to mention again that for each zone (grid mesh) only a point, randomly selected, is picked and not all the nodal combinations (like in the **Grid** sampling).

Assembler Objects These objects are either required or optional depending on the functionality of the Stratified Sampler. The objects must be listed with a rigorous syntax that, except for the XML node tag, is common among all the objects. Each of these nodes must contain 2 attributes that are used to identify them within the simulation framework:

- `class`, ***required string attribute***, the main "class" of the listed object. For example, it can be `'Models'`, `'Functions'`, etc.
- `type`, ***required string attribute***, the object identifier or sub-type. For example, it can be `'ROM'`, `'External'`, etc.

The **Stratified** approach requires or optionally accepts the following object types:

- `<Restart>`, ***string, optional field***, the body of this XML node must contain the name of an appropriate **DataObject** defined in the `<DataObjects>` block (see Section 14). It is used as a "restart" tool, where it accepts pre-existing solutions in the PointSet instead of recalculating solutions.

The following node is an additional option when a restart DataObject is provided:

- `<restartTolerance>`, ***float, optional field***, the body of this XML node must contain a valid floating point value. If a `<Restart>` node is supplied for this `<Sampler>`, this node offers a way to determine how strictly matching points are determined. Given a point

in the input space, if that point is within a relative Euclidean distance (equal to the tolerance) of a restart point, the nearest restart point will be used.
Default: 1e-15

Example:

```
<Samplers>
  ...
  <Stratified name='StratifiedName'>
    <variable name='var1'>
      <distribution>aDistributionNameDefinedInDistributionBlock1
      </distribution>
      <grid type='CDF' construction='equal' steps='5' >0.2
         0.8</grid>
    </variable>
    <variable name='var2'>
      <distribution>aDistributionNameDefinedInDistributionBlock2
      </distribution>
      <grid type='value' construction='equal' steps='100' >0.2
         21.0</grid>
    </variable>
    <variable name='var3'>
      <distribution>aDistributionNameDefinedInDistributionBlock3
      </distribution>
      <grid type='CDF' construction='custom'>0.2 0.5 1.0</grid>
    </variable>
  </Stratified>
  ...
</Samplers>
```

For N-dimensional (ND) distributions, there are two different approahes to perform the stratified sampling. In the first approach, the subgroups is determined by the joint CDF of given multivariate distributions. If this approach is used, the sampling is performed on a grid on a CDF, while the user is required to specify the same CDF grid for all the dimensions of the ND distribution. This is possible by defining a **<globalGrid>** node and associate such **<globalGrid>** to each variable belonging to the ND distribution as follows.

```
<Samplers>
  ...
  <Stratified name='StratifiedName'>
        <variable name='x0'>
            <distribution
               dim='1'>ND_InverseWeight_P</distribution>
```

```
            <grid type='globalGrid'>name_grid1</grid>
        </variable>
        <variable name='y0,z0'>
            <distribution
               dim='2'>ND_InverseWeight_P</distribution>
            <grid type='globalGrid'>name_grid1</grid>
        </variable>
        <globalGrid>
            <grid name='name_grid1' type='CDF'
               construction='custom'>0.1 1.0 0.2</grid>
        </globalGrid>
  </Stratified>
  ...
</Samplers>
...
```

The second approach is different than the first approach. Like in the **Grid** sampling, if the covariance matrix is provided and the input parameters is assumed to have the multivariate normal distribution, one can also use **Stratified** approach to sample the input parameters in the transformed space (aka subspace, reduced space). This means one creates the grids of variables listed by **<latentVariables>** in the transformed space. If this is the case, the user needs to provide additional information, i.e. the **<transformation>** under **<MultivariateNormal>** of **<Distributions>** (more information can be found in Section 11.2). In addition, the node **<variablesTransformation>** is also required for **Stratified** sampler. This node is used to tranform the variables specified by **<latentVariables>** in the transformed space of input into variables spefified by **<manifestVariables>** in the input space. The variables listed in **<latentVariables>** should be predefined in **<variable>**, and the variables listed in **<manifestVariables>** are used by the **<Models>**. In addition, **<globalGrid>** will be not used for approach.

- **<variablesTransformation>**, ***optional field***. this XML node accepts one attribute:
  - **distribution**, ***required string attribute***, the name for the distribution defined in the XML node **<Distributions>**. This attribute indicates the values of **<manifestVariables>** are drawn from **distribution**.

  In addition, this XML node also accepts three childen nodes:

  - **<latentVariables>**, ***comma separated string, required field***, user-defined latent variables that are used for the variables transformation. All the variables listed under this node should be also mentioned in **<variable>**.
  - **<manifestVariables>**, ***comma separated string, required field***, user-defined manifest variables that can be used by the **model**.

- <manifestVariablesIndex>, ***comma separated string, optional field***, user-defined manifest variables indices paired with <manifestVariables>. These indices indicate the position of manifest variables associated with multivariate normal distribution defined in the XML node <Distributions>. The indices should be postive integer. If not provided, the code will use the positions of manifest variables listed in <manifestVariables> as the indices.
- <method>, ***string, required field***, the method that is used for the variables transformation. The currently available method is '**pca**'.

```
...
<Models>
    ...
    <ExternalModel ModuleToLoad="lorentzAttractor_noK"
       name="PythonModule" subType="">
        <variables>sigma,rho,beta,x,y,z,time,z0,y0,z0</variables>
    </ExternalModel>
    ...
</Models>

<Distributions>
    ...
    <MultivariateNormal name='MVNDist' method='pca'>
        <transformation>
            <rank>3</rank>
        </transformation>
        <mu>0.0 1.0 2.0</mu>
        <covariance type="abs">
            1.0        0.6       -0.4
            0.6        1.0       0.2
            -0.4       0.2       0.8
        </covariance>
    </MultivariateNormal>
    ...
</Distributions>

<Samplers>
  ...
  <Stratified name='StratifiedName'>
        <variable name='x0'>
            <distribution dim='1'>MVNDist</distribution>
            <grid type='CDF' construction='equal' steps='3'>0.1
               0.9</grid>
```

```
        </variable>
        <variable name='y0'>
            <distribution dim='2'>MVNDist</distribution>
            <grid type='value' construction='equal'
               steps='3'>0.1 0.9</grid>
        </variable>
        <variable name='z0'>
            <distribution dim='3'>MVNDist</distribution>
            <grid type='CDF' construction='equal' steps='3'>0.2
               0.8</grid>
        </variable>
        <variablesTransformation model="PythonModule">
            <latentVariables>x1,y1,z1</latentVariables>
            <manifestVariables>x0,y0,z0</manifestVariables>
            <method>pca</method>
        </variablesTransformation>
  </Stratified>
  ...
</Samplers>
...
```

12.1.6 Response Surface Design

The **Response Surface Design**, or Response Surface Modeling (RSM), approach is one of the most common Design of Experiment (DOE) methodologies currently in use. It explores the relationships between several explanatory variables and one or more response variables. The main idea of RSM is to use a sequence of designed experiments to obtain an optimal response. RAVEN currently employs two different algorithms that can be classified within this family of methods:

- **Box-Behnken**: This methodology aims to achieve the following goals:
  - Each factor, or independent variable, is placed at one of three equally spaced values, usually coded as -1, 0, +1. (At least three levels are needed for the following goal);
  - The design should be sufficient to fit a quadratic model, that is, one squared term per factor and the products of any two factors;
  - The ratio of the number of experimental points to the number of coefficients in the quadratic model should be reasonable (in fact, their designs keep it in the range of 1.5 to 2.6);
  - The estimation variance should more or less depend only on the distance from the center (this is achieved exactly for the designs with 4 and 7 factors), and should not vary too much inside the smallest (hyper)cube containing the experimental points.

Each design can be thought of as a combination of a two-level (full or fractional) factorial design with an incomplete block design. In each block, a certain number of factors are put through all combinations for the factorial design, while the other factors are kept at the central values.

- **Central Composite**: This design consists of three distinct sets of experimental runs:
  - A factorial (perhaps fractional) design in the factors are studied, each having two levels;
  - A set of center points, experimental runs whose values of each factor are the medians of the values used in the factorial portion. This point is often replicated in order to improve the precision of the experiment;
  - A set of axial points, experimental runs identical to the centre points except for one factor, which will take on values both below and above the median of the two factorial levels, and typically both outside their range. All factors are varied in this way.

  This methodology is useful for building a second order (quadratic) model for the response variable without needing to use a complete three-level factorial experiment.

All the parameters, needed for setting up the algorithms reported above, must be defined within a **<ResponseSurfaceDesign>** block. This XML node accepts one attribute:

- **name**, ***required string attribute***, user-defined name of this sampler. **Note:** As with other objects, this identifier can be used to reference this specific entity from other input blocks in the XML.

In the **<ResponseSurfaceDesign>** input block, the user needs to specify the variables to sample. As already mentioned, these variables are specified within consecutive **<variable>** XML blocks:

- **<variable>**, ***XML node, required parameter*** will specify one attribute:
  - **name**, ***required string attribute***, user-defined name of this variable.

  This **<variable>** recognizes the following child nodes:
  - **<distribution>**, ***string, required field***, name of the distribution that is associated to this variable. Its name needs to be contained in the **<Distributions>** block explained in Section 11. In addition, if NDDistribution is used, the attribute **dim** is required. **Note:** Alternatively, this node must be omitted if the **<function>** node is supplied.

- **<function>**, ***string, required field***, name of the function that defines the calculation of this variable from other distributed variables. Its name needs to be contained in the **<Functions>** block explained in Section 18. This function must implement a method named "evaluate". **Note:** Alternatively, this node must be ommitted if the **<distribution>** node is supplied.
- **<grid>**, ***space separated floats, required field***, the content of this XML node depends on the definition of the associated attributes:
  - `type`, ***required string attribute***, user-defined discretization metric type: 1) **'CDF'**, the grid will be specified based on cumulative distribution function probability thresholds, and 2) **'value'**, the grid will be provided using variable values.
  - `construction`, ***required string attribute***, how the grid needs to be constructed, independent of its type (i.e. **'CDF'** or **'value'**).

  Based on the `construction` type, the content of the **<grid>** XML node and the requirements for other attributes change. In this case, only the following is available:

  - `construction`**='custom'**. The grid will be directly specified by the user. This construction type requires that the **<grid>** node contains the actual mesh bins. For example, if the grid `type` is **'CDF'**, in the body of **<grid>**, the user will specify the CDF probability thresholds (nodalization in probability). All the bins are checked against the associated **<distribution>** bounds. If one or more of them falls outside the distribution's bounds, the code will raise an error. No additional attributes are needed.

  Note: The **<grid>** node is only required if a **<distribution>** node is supplied. In the case of a **<function>** node, no grid information is requested. **Note:** Only the construction "custom" is available. In the **<grid>** body only the lower and upper bounds can be inputted (2 numbers only).

- **<constant>**, ***XML node, optional parameter*** the user is able to input variables that need to be kept constant. For doing this, as many **<constant>** nodes as needed can be input, where the body of the node contains the constant value that is going to be injected as an additional variable

- **<ResponseSurfaceDesignSettings>**, ***required***, In this sub-node, the user needs to specify different settings depending on the algorithm being used:

  - **<algorithmType>**, ***string, required field***, this XML node will contain the name of the algorithm to be used. Based on the chosen algorithm, other nodes need to be defined:
    - **<algorithmType>**`BoxBehnken`**<algorithmType/>**. If Box-Behnken is specified, the following additional node is recognized:
      - **<ncenters>**, ***integer, optional field***, the number of center points to include in the box. If this parameter is not specified, then a pre-determined number of

points are automatically included.
Default: Automatic Generation.

Note: In order to employ the "Box-Behnken" design, at least 3 variables must be used.

* <algorithmType>CentralComposite<algorithmType/>. If Central Composite is specified, the following additional nodes will be recognized:
  - <centers>, ***comma separated integers, optional field***, the number of center points to be included. This block needs to contain 2 integers values separated by a comma. The first entry represents the number of centers to be added for the factorial block; the second one is the one for the star block.
    Default: 4,4.
  - <alpha>, ***string, optional field***, in this node, the user decides how an α factor needs to be determined. Two options are available:

    orthogonal for orthogonal design.

    rotatable for rotatable design.

    Default: orthogonal.
  - <face>, ***string, optional field***, in this node, the user defines how faces should be constructed. Three options are available:

    circumscribed for circumscribed facing

    inscribed for inscribed facing

    faced for faced facing.

    Default: circumscribed.

Note: In order to employ the "Central Composite" design, at least 2 variables must be used.

Furthermore, if the covariance matrix is provided and the input parameters are assumed to have a multivariate normal distribution, one can use **ResponseSurfaceDesign** approach to sample the input parameters in the transformed space (aka subspace, reduced space). In this case, the user needs to provide additional information, i.e. the <transformation> under <MultivariateNormal> of <Distributions> (more information can be found in Section 11.2). In addition, the node <variablesTransformation> is also required for **ResponseSurfaceDesign** sampling. This node is used to tranform the variables specified by <latentVariables> in the transformed space of input into variables spefified by <manifestVariables> in the input space. The variables listed in <latentVariables> should be predefined in <variable>, and the variables listed in <manifestVariables> are used by the <Models>.

- <variablesTransformation>, ***optional field***. this XML node accepts one attribute:

- `distribution`, ***required string attribute***, the name for the distribution defined in the XML node `<Distributions>`. This attribute indicates the values of `<manifestVariables>` are drawn from `distribution`.

In addition, this XML node also accepts three childen nodes:

- `<latentVariables>`, ***comma separated string, required field***, user-defined latent variables that are used for the variables transformation. All the variables listed under this node should be also mentioned in `<variable>`.
- `<manifestVariables>`, ***comma separated string, required field***, user-defined manifest variables that can be used by the `model`.
- `<manifestVariablesIndex>`, ***comma separated string, optional field***, user-defined manifest variables indices paired with `<manifestVariables>`. These indices indicate the position of manifest variables associated with multivariate normal distribution defined in the XML node `<Distributions>`. The indices should be postive integer. If not provided, the code will use the positions of manifest variables listed in `<manifestVariables>` as the indices.
- `<method>`, ***string, required field***, the method that is used for the variables transformation. The currently available method is '**pca**'.

Example:

```
<Samplers>
  ...
    <ResponseSurfaceDesign name='BoxBehnkenRespDesign'>
        <ResponseSurfaceDesignSettings>
            <algorithmType>BoxBehnken</algorithmType>
            <ncenters>1</ncenters>
        </ResponseSurfaceDesignSettings>
        <variable name='var1' >
            <distribution >Gauss1</distribution>
          <grid type='CDF' construction='custom'  >0.2
             0.8</grid>
        </variable>
        <!-- N.B. at least 3 variables need to inputted
                in order to employ this algorithm
         -->
    </ResponseSurfaceDesign>
    <ResponseSurfaceDesign name='CentralCompositeRespDesign'>
        <ResponseSurfaceDesignSettings>
            <algorithmType>CentralComposite</algorithmType>
            <centers>1, 2</centers>
```

```
            <alpha>orthogonal</alpha>
            <face>circumscribed</face>
        </ResponseSurfaceDesignSettings>
        <variable name='var4' >
            <distribution >Gauss1</distribution>
            <grid type='CDF' construction='custom'  >0.2
              0.8</grid>
        </variable>
        <!-- N.B. at least 2 variables need to inputted
                in order to employ this algorithm
         -->
    </ResponseSurfaceDesign>
    <ResponseSurfaceDesign name='transformedSpaceSampling'>
        <ResponseSurfaceDesignSettings>
            <algorithmType>BoxBehnken</algorithmType>
            <ncenters>1</ncenters>
        </ResponseSurfaceDesignSettings>
        <variable name='var1' >
            <distribution >Gauss1</distribution>
           <grid type='CDF' construction='custom'  >0.2
             0.8</grid>
        </variable>
        ...
        <variablesTransformation model="givenModel">
          <latentVariables>var1,...</latentVariables>
          <manifestVariables>...</manifestVariables>
          <method>pca</method>
        </variablesTransformation>
    </ResponseSurfaceDesign>
  ...
</Samplers>
```

12.1.7 Factorial Design

The **Factorial Design** method is an important method to determine the effects of multiple variables on a response. A factorial design can reduce the number of samples one has to perform by studying multiple factors simultaneously. Additionally, it can be used to find both main effects (from each independent factor) and interaction effects (when both factors must be used to explain the outcome). A factorial design tests all possible conditions. Because factorial designs can lead to a large number of trials, which can become expensive and time-consuming, they are best used for small numbers of variables with only a few domain discretizations (1 to 3). Factorial designs work

well when interactions between variables are strong and important and where every variable contributes significantly. RAVEN currently employs three different algorithms that can be classified within this family of techniques:

- **General Full Factorial** explores the input space by investigating all possible combinations of a set of factors (variables).
- **2-Level Fractional-Factorial** consists of a carefully chosen subset (fraction) of the experimental runs of a full factorial design. The subset is chosen so as to exploit the sparsity-of-effects principle exposing information about the most important features of the problem studied, while using a fraction of the effort of a full factorial design in terms of experimental runs and resources.
- **Plackett-Burman** identifies the most important factors early in the experimentation phase when complete knowledge about the system is usually unavailable. It is an efficient screening method for identifying the active factors (variables) using as few samples as possible. In Plackett-Burman designs, main effects have a complicated confounding relationship with two-factor interactions. Therefore, these designs should be used to study main effects when it can be assumed that two-way interactions are negligible.

All the parameters needed for setting up the algorithms reported above must be defined within a **<FactorialDesign>** block. This XML node accepts one attribute:

- **name**, ***required string attribute***, user-defined name of this sampler. **Note:** As with other objects, this identifier can be used to reference this specific entity from other input blocks in the XML.

In the **<FactorialDesign>** input block, the user needs to specify the variables to sample. As already mentioned, these variables are specified within consecutive **<variable>** XML blocks:

- **<variable>**, ***XML node, required parameter*** will specify one attribute:
  - **name**, ***required string attribute***, user-defined name of this variable.

  This **<variable>** recognizes the following child nodes:

  - **<distribution>**, ***string, required field***, name of the distribution that is associated to this variable. Its name needs to be contained in the **<Distributions>** block explained in Section 11. In addition, if NDDistribution is used, the attribute **dim** is required. **Note:** Alternatively, this node must be omitted if the **<function>** node is supplied.

- <function>, ***string, required field***, name of the function that defines the calculation of this variable from other distributed variables. Its name needs to be contained in the <Functions> block explained in Section 18. This function must implement a method named "evaluate". **Note:** Alternatively, this node must be ommitted if the <distribution> node is supplied.
- <grid>, ***space separated floats, required field***, the content of this XML node depends on the definition of the associated attributes:
  * type, ***required string attribute***, user-defined discretization metric type: 1) 'CDF', the grid will be specified based on cumulative distribution function probability thresholds, and 2) 'value', the grid will be provided using variable values.
  * construction, ***required string attribute***, how the grid needs to be constructed, independent of its type (i.e. 'CDF' or 'value').

  Based on the construction type, the content of the <grid> XML node and the requirements for other attributes change:

  * construction='equal'. The grid is going to be constructed equally-spaced (type='value') or equally probable (type='CDF'). This construction type requires the definition of additional attributes:
    - steps, ***required integer attribute***, number of equally spaced/probable discretization steps.

    This construction type requires that the content of the <grid> node represents the lower and upper bounds (either in probability or value). Two values need to be specified; the lowest one will be considered as the $lowerBound$, the largest, the $upperBound$. The lower and upper bounds are checked against the associated <distribution> bounds. If one or both of them falls outside the distribution's bounds, the code will raise an error. The $stepSize$ is determined as follows:
    $stepSize = (upperBound - lowerBound)/steps$
  * construction='custom'. The grid will be directly specified by the user. No additional attributes are needed. This construction type requires that the <grid> node contains the actual mesh bins. For example, if the grid type is 'CDF', in the body of <grid>, the user will specify the CDF probability thresholds (nodalization in probability). All the bins are checked against the associated <distribution> bounds. If one or more of them falls outside the distribution's bounds, the code will raise an error.

  Note: The <grid> node is only required if a <distribution> node is supplied. In the case of a <function> node, no grid information is requested.

- <constant>, ***XML node, optional parameter*** the user is able to input variables that need to be kept constant. For doing this, as many <constant> nodes as needed can be input, where the body of the node contains the constant value that is going to be injected as an additional variable

The main **<FactorialDesign>** block needs to contain an additional sub-node called **<FactorialSettings>**. In this sub-node, the user needs to specify different settings depending on the algorithm being used:

- **<algorithmType>**, ***string, required field***, specifies the algorithm to be used. Based on the chosen algorithm, other nodes may be defined:
  - **<algorithmType>**full**<algorithmType/>**. Full factorial design. If full is specified, no additional nodes are necessary.
    Note: The full factorial design does not have any limitations on the number of discretization bins that can be used in the **<grid>** XML node for each **<variable>** specified.
  - **<algorithmType>**2levelFract**<algorithmType/>**. Two-level Fractional-Factorial design. If
    2levelFract is specified, the following additional nodes must be specified:
    * **<gen>**, ***space separated strings, required field***, specifies the confounding mapping. For instance, in this block the user defines the decisions on a fraction of the full-factorial by allowing some of the factor main effects to be compounded with other factor interaction effects. This is done by defining an alias structure that defines, symbolically, these interactions. These alias structures are written like "C = AB" or "I = ABC", or "AB = CD", etc. These define how a column is related to the others.
    * **<genMap>**, ***space separated strings, required field***, defines the mapping between the **<gen>** symbolic aliases and the variables that have been inputted in the **<FactorialDesign>** main block.

    Note: The Two-levels Fractional-Factorial design is limited to 2 discretization bins in the **<grid>** node for each **<variable>**.
  - **<algorithmType>**pb**<algorithmType/>**. Plackett-Burman design. If pb is specified, no additional nodes are necessary.
    Note: The Plackett-Burman design does not have any limitations on the number of discretization bins allowed in the **<grid>** node for each **<variable>**.

Example:

```xml
<Samplers>
  ...
  <FactorialDesign name='fullFactorial'>
    <FactorialSettings>
      <algorithmType>full</algorithmType>
    </FactorialSettings>
    <variable name='var1' >
```

```
      <distribution>aDistributionNameDefinedInDistributionBlock1
      </distribution>
      <grid type='value' construction='custom' >0.02 0.03
         0.5</grid>
    </variable>
    <variable name='var2' >
      <distribution>aDistributionNameDefinedInDistributionBlock2
      </distribution>
      <grid type='CDF' construction='custom'>0.5 0.7 1.0</grid>
    </variable>
  </FactorialDesign>
  <FactorialDesign name='2levelFractFactorial'>
    <FactorialSettings>
      <algorithmType>2levelFract</algorithmType>
      <gen>a,b,ab</gen>
      <genMap>var1,var2,var3</genMap>
    </FactorialSettings>
    <variable name='var1' >
      <distribution>aDistributionNameDefinedInDistributionBlock3
      </distribution>
      <grid type='value' construction='custom' >0.02 0.5</grid>
    </variable>
    <variable name='var2' >
      <distribution>aDistributionNameDefinedInDistributionBlock
      </distribution>
      <grid type='CDF' construction='custom'>0.5 1.0</grid>
    </variable>
    <variable name='var3'>
      <distribution>aDistributionNameDefinedInDistributionBlock5
      </distribution>
      <grid type='value' upperBound='4' construction='equal'
         steps='1'>0.5</grid>
    </variable>
  </FactorialDesign>
  <FactorialDesign name='pbFactorial'>
    <FactorialSettings>
      <algorithmType>pb</algorithmType>
    </FactorialSettings>
    <variable name='var1' >
      <distribution>aDistributionNameDefinedInDistributionBlock6
      </distribution>
      <grid type='value' construction='custom' >0.02 0.5</grid>
```

```
    </variable>
    <variable name='VarGauss2' >
      <distribution>aDistributionNameDefinedInDistributionBlock7
      </distribution>
      <grid type='CDF' construction='custom'>0.5 1.0</grid>
    </variable>
  </FactorialDesign>
  ...
</Samplers>
```

12.1.8 Ensemble Forward Sampling strategy

The **Ensemble Forward** sampling approach allows the user to combine multiple Forward sampling strategies into one single strategy. For example, it can happen that a variable is more suitable for a particular sampling strategy (e.g. a stochastic event modeled with a Monte Carlo approach) and a second variable is more suitable for another sampling method (e.g. because part of a parametric space modeled with a Grid-based approach). The specifications of this sampler must be defined within a **<EnsembleForward>** XML block. This XML node accepts the following attributes:

- **name**, *required string attribute*, user-defined name of this Sampler. N.B. As for the other objects, this is the name that can be used to refer to this specific entity from other input blocks (xml);

In the **EnsembleForward** input block, the user needs to specify the sampling strategies that he wants to combine together.
Currently, only the following strategies can be combined:

- **<MonteCarlo>**
- **<Grid>**
- **<Stratiefied>**
- **<FactorialDesign>**
- **<ResponseSurfaceDesign>**
- **<CustomSampler>**

For each of the above samplers, the input specifications can be found in the relative sections.

Example:

```xml
<Samplers>
  ...
    <EnsembleForward name="testEnsembleForward">
        <MonteCarlo name = "theMC">
            <samplerInit> <limit>4</limit> </samplerInit>
            <variable name="sigma">
                <distribution>norm</distribution>
            </variable>
        </MonteCarlo>
        <Grid name = "theGrid">
            <variable name="x0">
                <distribution>unif</distribution>
                <grid construction="custom" type="value">0.02
                   0.5 0.6</grid>
            </variable>
        </Grid>
        <Stratified name = "theStratified">
            <variable name="z0">
                <distribution>tri</distribution>
                <grid construction="equal" steps="2"
                   type="CDF">0.2 0.8</grid>
            </variable>
            <variable name="y0">
                <distribution>unif</distribution>
                <grid construction="equal" steps="2"
                   type="value">0.5 0.8</grid>
            </variable>
        </Stratified>
        <ResponseSurfaceDesign name = "theRSD">
            <ResponseSurfaceDesignSettings>
                <algorithmType>CentralComposite</algorithmType>
                <centers>1,2</centers>
                <alpha>orthogonal</alpha>
                <face>circumscribed</face>
            </ResponseSurfaceDesignSettings>
            <variable name="rho">
                <distribution>unif</distribution>
                <grid construction="custom" type="CDF">0.0
                   1.0</grid>
            </variable>
            <variable name="beta">
                <distribution>tri</distribution>
```

```
                <grid construction="custom" type="value">0.1
                  1.5</grid>
            </variable>
        </ResponseSurfaceDesign>
    </EnsembleForward>
  ...
</Samplers>
```

Care should be used when using deterministic random seeds for EnsembleForward sampling. The EnsembleForward sample will ignore any seeds set in any of its subset samplers; however, the global random seed can be set by adding a **<samplerInit>** block with the **<initialSeed>** block therein, with an integer value providing the seed. For example,

```
<Samplers>
  ...
  <EnsembleForward name='testEnsembleForward'>
    <samplerInit>
      <initialSeed>42</initialSeed>
    </samplerInit>
  ...
   </EnsembleForward>
  ...
</Samplers>
```

Because RAVEN has a single global random number generator, this will set the seed for the full calculation when the Step containing a run using this ForwardSampler is begun.

Note also variables that are defined from functions, as well as constants, need to be defined outside the samplers of the ensemble sampler. An example is shown below.

Example:

```
<Samplers>
  <EnsembleForward name='testEnsembleForward'>
    <variable name='x3'>
        <function>funct1</function>
    </variable>
    <variable name='x4,x5'>
        <function>funct2</function>
    </variable>
    <constant name='pi'>3.14159</constant>
    <MonteCarlo name='notNeeded'>
      <samplerInit>
        <limit>3</limit>
```

```
      </samplerInit>
      <variable name='x1'>
        <distribution>norm</distribution>
      </variable>
    </MonteCarlo>
    <Grid name='notNeeded'>
      <variable name='x2'>
        <distribution>unif</distribution>
        <grid construction='custom' type='value'>0.02
           0.6</grid>
      </variable>
    </Grid>
   </EnsembleForward>
</Samplers>
```

In this example note that:

- variables $x1$ and $x2$ are generated by the two samplers (Monte-Carlo and Grid respectively)
- variable $x3$ is generated from the function $funct1$
- variables $x4$ and $x5$ are generated from the function $funct2$
- variables $x3$, $x4$ and $x5$ are defined outside the Monte-Carlo and Grid

12.1.9 Custom Sampling strategy

The **Custom** sampling approach allows the user to specify a predefined set of coordinates (in the input space) that RAVEN should use to inquire the model. For example, the user can provide a CSV file containing a list of samples that RAVEN should use. The specifications of this sampler must be defined within a **<CustomSampler>** XML block. This XML node accepts the following attributes:

- **name**, *required string attribute*, user-defined name of this Sampler. N.B. As for the other objects, this is the name that can be used to refer to this specific entity from other input blocks (xml);

In the **CustomSampler** input block, the user needs to specify the variables need to be sampled. As already mentioned, these variables are inputted within consecutive XML blocks called **<variable>**. In addition, the **<Source>** from which the samples need to be retrieved needs to be specified as well:

- **<variable>**, ***XML node, required parameter*** will specify one attribute:

– `name`, ***required string attribute***, user-defined name of this variable.

- `<Source>`, ***XML node, required parameter*** will specify the following attributes:
  - `class`, ***required string attribute***, class entity of the source where the samples need to be retrieved from. It can be either **Files** or **DataObjects**.
  - `type`, ***required string attribute***, type of the source withing the previously explained "class". If `class` is **Files**, this attribute needs to be kept empty; otherwise it must be **PointSet** (the only allowed DataObjects usable as source in this sampler).
    Note: If the `<Source>` `class` is **Files**, the File needs to be a standard CSV file, specified in the `<Files>` XML block in the RAVEN input.
    In addition, it is important to notice that if in the `<Source>` the **PointProbability** and **ProbabilityWeight** quantities are not found, the samples are assumed to come from a MonteCarlo (from a statistical post-processing prospective).
- `<constant>`, ***XML node, optional parameter*** the user is able to input variables that need to be kept constant. For doing this, as many `<constant>` nodes as needed can be input, where the body of the node contains the constant value that is going to be injected as an additional variable

Example:

Table 1. samples.csv

y	x	z	PointProbability	ProbabilityWeight
0.725675246	0.031099304	0.984988317	0.1	0.2
0.565949127	0.028589754	1.13186372	0.1	0.2
0.72567754	0.031099304	0.967209238	0.1	0.2
0.565951633	0.028589754	1.111431662	0.1	0.2
0.725968307	0.031100307	0.98498835	0.1	0.2

```
<Samplers>
  ...
  <Samplers>
    <CustomSampler name="customSamplerDataObject">
      <Source   class="DataObjects"
         type="PointSet">outCustomSamplerFromFile</Source>
      <variable name="x"/>
      <variable name="y"/>
      <variable name="z"/>
    </CustomSampler>
  </Samplers>
```

```xml
  <Samplers>
    <CustomSampler name="customSamplerFile">
      <Source   class="Files"  type="">samples.csv</Source>
      <variable name="x"/>
      <variable name="y"/>
      <variable name="z"/>
    </CustomSampler>
  </Samplers>
  ...
</Samplers>
```

12.2 Dynamic Event Tree (DET) Samplers

The **Dynamic Event Tree** methodologies are designed to take the timing of events explicitly into account, which can become very important especially when uncertainties in complex phenomena are considered. Hence, the main idea of this methodology is to let a system code determine the pathway of an accident scenario within a probabilistic environment. In this family of methods, a continuous monitoring of the system evolution in the phase space is needed. In order to use the DET-based methods, the generic driven code needs to have, at least, an internal trigger system and, consequently, a "restart" capability. In the RAVEN framework, 4 different DET samplers are available:

- **Dynamic Event Tree (DET)**
- **Hybrid Dynamic Event Tree (HDET)**
- **Adaptive Dynamic Event Tree (ADET)**
- **Adaptive Hybrid Dynamic Event Tree (AHDET)**

The ADET and the AHDET methodologies represent a hybrid between the DET/HDET and adaptive sampling approaches. For this reason, its input requirements are reported in the Adaptive Samplers' section (12.3).

12.2.1 Dynamic Event Tree

The **Dynamic Event Tree** sampling approach is a sampling strategy that is designed to take the timing of events, in transient/accident scenarios, explicitly into account. From an application point of view, an N-Dimensional grid is built on the CDF space. A single simulation is spawned and a set of triggers is added to the system code control logic. Every time a trigger is activated (one of

the CDF thresholds in the grid is exceeded), a new set of simulations (branches) is spawned. Each branch carries its conditional probability. In the RAVEN code, the triggers are defined by specifying a grid using a predefined discretization metric in the mode input space. RAVEN provides two discretization metrics: 1) CDF, and 2) value. Thus, the trigger thresholds can be entered either in probability or value space.

The specifications of this sampler must be defined within a **<DynamicEventTree>** XML block. This XML node accepts the following attributes:

- **name**, ***required string attribute***, user-defined name of this sampler. **Note:** As with other objects, this identifier can be used to reference this specific entity from other input blocks in the XML.
- **printEndXmlSummary**, ***optional string/boolean attribute***, controls the dumping of a "summary" of the DET performed into an external XML.
  Default: False.
- **maxSimulationTime**, ***optional float attribute***, this attribute controls the maximum "mission" time of the simulation underneath.
  Default: None.

In the **<DynamicEventTree>** input block, the user needs to specify the variables to sample. As already mentioned, these variables are specified within consecutive **<variable>** XML blocks:

- **<variable>**, ***XML node, required parameter*** will specify one attribute:
  - **name**, ***required string attribute***, user-defined name of this variable.

  This **<variable>** recognizes the following child nodes:
  - **<distribution>**, ***string, required field***, name of the distribution that is associated to this variable. Its name needs to be contained in the **<Distributions>** block explained in Section 11. In addition, if NDDistribution is used, the attribute **dim** is required. **Note:** Alternatively, this node must be omitted if the **<function>** node is supplied.
  - **<function>**, ***string, required field***, name of the function that defines the calculation of this variable from other distributed variables. Its name needs to be contained in the **<Functions>** block explained in Section 18. This function must implement a method named "evaluate". **Note:** Alternatively, this node must be ommitted if the **<distribution>** node is supplied.
  - **<grid>**, ***space separated floats, required field***, the content of this XML node depends on the definition of the associated attributes:

* type, ***required string attribute***, user-defined discretization metric type: 1) 'CDF', the grid will be specified based on cumulative distribution function probability thresholds, and 2) 'value', the grid will be provided using variable values.
* construction, ***required string attribute***, how the grid needs to be constructed, independent of its type (i.e. 'CDF' or 'value').

Based on the construction type, the content of the <grid> XML node and the requirements for other attributes change:

* construction='equal'. The grid is going to be constructed equally-spaced (type='value') or equally probable (type='CDF'). This construction type requires the definition of additional attributes:
  - steps, ***required integer attribute***, number of equally spaced/probable discretization steps.

  This construction type requires that the content of the <grid> node represents the lower and upper bounds (either in probability or value). Two values need to be specified; the lowest one will be considered as the $lowerBound$, the largest, the $upperBound$. The lower and upper bounds are checked against the associated <distribution> bounds. If one or both of them falls outside the distribution's bounds, the code will raise an error. The $stepSize$ is determined as follows:
  $stepSize = (upperBound - lowerBound)/steps$
* construction='custom'. The grid will be directly specified by the user. No additional attributes are needed. This construction type requires that the <grid> node contains the actual mesh bins. For example, if the grid type is 'CDF', in the body of <grid>, the user will specify the CDF probability thresholds (nodalization in probability). All the bins are checked against the associated <distribution> bounds. If one or more of them falls outside the distribution's bounds, the code will raise an error.

Note: The <grid> node is only required if a <distribution> node is supplied. In the case of a <function> node, no grid information is requested.

- <constant>, ***XML node, optional parameter*** the user is able to input variables that need to be kept constant. For doing this, as many <constant> nodes as needed can be input, where the body of the node contains the constant value that is going to be injected as an additional variable

Example:

```
<Samplers>
  ...
  <DynamicEventTree name='DETname'>
    <variable name='var1'>
      <distribution>aDistributionNameDefinedInDistributionBlock1
         </distribution>
```

```
      <grid type='value' construction='equal' steps='100' >1.0
         201.0</grid>
    </variable>
    <variable name='var2'>
      <distribution>aDistributionNameDefinedInDistributionBlock2
         </distribution>
      <grid type='CDF' construction='equal' steps='5'>0 1</grid>
    </variable>
    <variable name='var3'>
      <distribution>aDistributionNameDefinedInDistributionBlock3
         </distribution>
      <grid type='value' construction='equal' steps='10' >11.0
         21.0</grid>
    </variable>
    <variable name='var4'>
      <distribution>aDistributionNameDefinedInDistributionBlock4
         </distribution>
      <grid type='CDF' construction='equal' steps='5' >0.0
         1.0</grid>
    </variable>
    <variable name='var5'>
      <distribution>aDistributionNameDefinedInDistributionBlock5
         </distribution>
      <grid type='value' construction='custom'>0.2 0.5
         10.0</grid>
    </variable>
    <variable name='var6'>
      <distribution>aDistributionNameDefinedInDistributionBlock6
         </distribution>
      <grid type='CDF' construction='custom'>0.2 0.5 1.0</grid>
    </variable>
  </DynamicEventTree>
  ...
</Samplers>
```

12.2.2 Hybrid Dynamic Event Tree

The **Hybrid Dynamic Event Tree** sampling approach is a sampling strategy that represents an evolution of the Dynamic Event Tree method for the simultaneous exploration of the epistemic and aleatory uncertain space. In similar approaches, the uncertainties are generally treated by employing a Monte-Carlo sampling approach (epistemic) and DET methodology (aleatory). The HDET

methodology, developed within the RAVEN code, can reproduce the capabilities employed by this approach, but provides additional sampling strategies to the user. The epistemic or epistemic-like uncertainties can be sampled through the following strategies:

- Monte-Carlo;
- Grid sampling;
- Stratified (e.g., Latin Hyper Cube).

From a practical point of view, the user defines the parameters that need to be sampled by one or more different approaches. The HDET module samples those parameters creating an N-dimensional grid characterized by all the possible combinations of the input space coordinates coming from the different sampling strategies. Each coordinate in the input space represents a separate and parallel standard DET exploration of the uncertain domain. The HDET methodology allows the user to explore the uncertain domain employing the best approach for each variable kind. The addition of a grid sampling strategy among the usable approaches allows the user to perform a discrete parametric study under aleatory and epistemic uncertainties.

Regarding the input requirements, the HDET sampler is a "sub-type" of the **<DynamicEventTree>** sampler. For this reason, its specifications must be defined within a **<DynamicEventTree>** block. This XML node accepts the following attributes:

- **name**, ***required string attribute***, user-defined name of this sampler. **Note:** As with other objects, this identifier can be used to reference this specific entity from other input blocks in the XML.
- **printEndXmlSummary**, ***optional string/boolean attribute***, controls the dumping of a "summary" of the DET performed into an external XML.
  Default: False.
- **maxSimulationTime**, ***optional float attribute***, this attribute controls the maximum "mission" time of the simulation underneath.
  Default: None.

In the **<DynamicEventTree>** input block, the user needs to specify the variables to sample. As already mentioned, these variables are specified within consecutive **<variable>** XML blocks:

- **<variable>**, ***XML node, required parameter*** will specify one attribute:
  - **name**, ***required string attribute***, user-defined name of this variable.

This **<variable>** recognizes the following child nodes:

- **<distribution>**, ***string, required field***, name of the distribution that is associated to this variable. Its name needs to be contained in the **<Distributions>** block explained in Section 11. In addition, if NDDistribution is used, the attribute **dim** is required. **Note:** Alternatively, this node must be omitted if the **<function>** node is supplied.
- **<function>**, ***string, required field***, name of the function that defines the calculation of this variable from other distributed variables. Its name needs to be contained in the **<Functions>** block explained in Section 18. This function must implement a method named "evaluate". **Note:** Alternatively, this node must be ommitted if the **<distribution>** node is supplied.
- **<grid>**, ***space separated floats, required field***, the content of this XML node depends on the definition of the associated attributes:
  * **type**, ***required string attribute***, user-defined discretization metric type: 1) **'CDF'**, the grid will be specified based on cumulative distribution function probability thresholds, and 2) **'value'**, the grid will be provided using variable values.
  * **construction**, ***required string attribute***, how the grid needs to be constructed, independent of its type (i.e. **'CDF'** or **'value'**).

  Based on the **construction** type, the content of the **<grid>** XML node and the requirements for other attributes change:

  * **construction='equal'**. The grid is going to be constructed equally-spaced (**type='value'**) or equally probable (**type='CDF'**). This construction type requires the definition of additional attributes:
    · **steps**, ***required integer attribute***, number of equally spaced/probable discretization steps.

    This construction type requires that the content of the **<grid>** node represents the lower and upper bounds (either in probability or value). Two values need to be specified; the lowest one will be considered as the $lowerBound$, the largest, the $upperBound$. The lower and upper bounds are checked against the associated **<distribution>** bounds. If one or both of them falls outside the distribution's bounds, the code will raise an error. The $stepSize$ is determined as follows:
    $stepSize = (upperBound - lowerBound)/steps$
  * **construction='custom'**. The grid will be directly specified by the user. No additional attributes are needed. This construction type requires that the **<grid>** node contains the actual mesh bins. For example, if the grid **type** is **'CDF'**, in the body of **<grid>**, the user will specify the CDF probability thresholds (nodalization in probability). All the bins are checked against the associated **<distribution>** bounds. If one or more of them falls outside the distribution's bounds, the code will raise an error.

Note: The `<grid>` node is only required if a `<distribution>` node is supplied. In the case of a `<function>` node, no grid information is requested.

- `<constant>`, ***XML node, optional parameter*** the user is able to input variables that need to be kept constant. For doing this, as many `<constant>` nodes as needed can be input, where the body of the node contains the constant value that is going to be injected as an additional variable

In order to activate the **Hybrid Dynamic Event Tree** sampler, the main `<DynamicEventTree>` block needs to contain, at least, an additional sub-node called `<HybridSampler>`. As already mentioned, the user can combine the Monte-Carlo, Stratified, and Grid approaches in order to create a "pre-sampling" N-dimensional grid, from whose nodes a standard DET method is employed. For this reason, the user can specify a maximum of three `<HybridSampler>` sub-nodes (i.e. one for each of the available Forward samplers). This sub-node needs to contain the following attribute:

- `type`, ***required string attribute***, type of pre-sampling strategy to be used. Available options are `'MonteCarlo'`, `'Grid'`, and `'Stratified'`.

Independent of the type of "pre-sampler" that has been specified, the `<HybridSampler>` must contain the variables that need to be sampled. As already mentioned, these variables are specified within consecutive `<variable>` XML blocks:

- `<variable>`, ***XML node, required parameter*** will specify one attribute:
  - `name`, ***required string attribute***, user-defined name of this variable.

  This `<variable>` recognizes the following child nodes:
  - `<distribution>`, ***string, required field***, name of the distribution that is associated to this variable. Its name needs to be contained in the `<Distributions>` block explained in Section 11. In addition, if NDDistribution is used, the attribute `dim` is required. **Note:** Alternatively, this node must be omitted if the `<function>` node is supplied.
  - `<function>`, ***string, required field***, name of the function that defines the calculation of this variable from other distributed variables. Its name needs to be contained in the `<Functions>` block explained in Section 18. This function must implement a method named "evaluate". **Note:** Alternatively, this node must be ommitted if the `<distribution>` node is supplied.

- `<constant>`, ***XML node, optional parameter*** the user is able to input variables that need to be kept constant. For doing this, as many `<constant>` nodes as needed can be input, where the body of the node contains the constant value that is going to be injected as an additional variable

If a pre-sampling strategy **type** is either **'Grid'** or **'Stratified'**, within the **<variable>** blocks, the user needs to specify the sub-node **<grid>**. As with the standard DET, the content of this XML node depends on the definition of the associated attributes:

- **type**, ***required string attribute***, user-defined discretization metric type:
  - **'CDF'**, the grid is going to be specified based on the cumulative distribution function probability thresholds
  - **'value'**, the grid is going to be provided using variable values.
- **construction**, ***required string attribute***, how the grid needs to be constructed, independent of its type (i.e. **'CDF'** or **'value'**).

Based on the **construction** type, the content of the **<grid>** XML node and the requirements for other attributes change:

- **construction='equal'**. The grid is going to be constructed equally-spaced (**type='value'**) or equally probable (**type='CDF'**). This construction type requires the definition of additional attributes:
  - **steps**, ***required integer attribute***, number of equally spaced/probable discretization steps.

  This construction type requires that the content of the **<grid>** node represents the lower and upper bounds (either in probability or value). Two values need to be specified; the lowest one will be considered as the $lowerBound$, the largest, the $upperBound$. The lower and upper bounds are checked against the associated **<distribution>** bounds. If one or both of them falls outside the distribution's bounds, the code will raise an error. The $stepSize$ is determined as follows:
  $stepSize = (upperBound - lowerBound)/steps$
- **construction='custom'**. The grid will be directly specified by the user. No additional attributes are needed. This construction type requires that the **<grid>** node contains the actual mesh bins. For example, if the grid **type** is **'CDF'**, in the body of **<grid>**, the user will specify the CDF probability thresholds (nodalization in probability). All the bins are checked against the associated **<distribution>** bounds. If one or more of them falls outside the distribution's bounds, the code will raise an error.

Example:

```
<Samplers>
  ...
  <DynamicEventTree name='HybridDETname' print_end_XML="True">
```

```xml
<HybridSampler type='MonteCarlo' limit='2'>
  <variable name='var1' >
    <distribution>aDistributionNameDefinedInDistributionBlock1
       </distribution>
  </variable>
  <variable name='var2' >
    <distribution>aDistributionNameDefinedInDistributionBlock2
       </distribution>
    <grid type='CDF' construction='equal' steps='1'
       lowerBound='0.1'>0.1</grid>
  </variable>
</HybridSampler>
<HybridSampler type='Grid'>
  <!-- Point sampler way (directly sampling the variable) -->
  <variable name='var3' >
    <distribution>aDistributionNameDefinedInDistributionBlock3
       </distribution>
    <grid type='CDF' construction='equal' steps='1'
       lowerBound='0.1'>0.1</grid>
  </variable>
  <variable name='var4' >
    <distribution>aDistributionNameDefinedInDistributionBlock4
       </distribution>
    <grid type='CDF' construction='equal' steps='1'
       lowerBound='0.1'>0.1</grid>
  </variable>
</HybridSampler>
<HybridSampler type='Stratified'>
  <!-- Point sampler way (directly sampling the variable )
     -->
  <variable name='var5' >
    <distribution>aDistributionNameDefinedInDistributionBlock5
       </distribution>
    <grid type='CDF' construction='equal' steps='1'
       lowerBound='0.1'>0.1</grid>
  </variable>
  <variable name='var6' >
    <distribution>aDistributionNameDefinedInDistributionBlock6
       </distribution>
    <grid type='CDF' construction='equal' steps='1'
       lowerBound='0.1'>0.1</grid>
  </variable>
```

```
    </HybridSampler>
    <!-- DYNAMIC EVENT TREE INPUT (it goes outside an inner
       block like HybridSamplerSettings) -->
      <Distribution name='dist7'>
        <distribution>aDistributionNameDefinedInDistributionBlock7
           </distribution>
        <grid type='CDF' construction='custom'>0.1 0.8</grid>
      </Distribution>
  </DynamicEventTree>
  ...
</Samplers>
```

12.3 Adaptive Samplers

The Adaptive Samplers' family provides the possibility to perform smart sampling (also known as adaptive sampling) as an alternative to classical "Forward" techniques. The motivation is that system simulations are often computationally expensive, time-consuming, and high dimensional with respect to the number of input parameters. Thus, exploring the space of all possible simulation outcomes is infeasible using finite computing resources. During simulation-based probabilistic risk analysis, it is important to discover the relationship between a potentially large number of input parameters and the output of a simulation using as few simulation trials as possible.

The description above characterizes a typical context for performing adaptive sampling where a few observations are obtained from the simulation, a reduced order model (ROM) is built to represent the simulation space, and new samples are selected based on the model constructed. The reduced order model (see section 17.3) is then updated based on the simulation results of the sampled points. In this way, an attempt is made to gain the most information possible with a small number of carefully selected sample points, limiting the number of expensive trials needed to understand features of the system space.

Currently, RAVEN provides support for the following adaptive algorithms:

- Limit Surface Search
- Adaptive Dynamic Event Tree
- Adaptive Hybrid Dynamic Event Tree
- Adaptive Sparse Grid
- Adaptive Sobol Decomposition

In the following paragraphs, the input requirements and a small explanation of the different sampling methods are reported.

12.3.1 Limit Surface Search

The **Limit Surface Search** approach is an advanced methodology that employs a smart sampling around transition zones that determine a change in the status of the system (limit surface). To perform such sampling, RAVEN uses ROMs for predicting, in the input space, the location(s) of these transitions, in order to accelerate the exploration of the input space in proximity of the limit surface.

The specifications of this sampler must be defined within an **<LimitSurfaceSearch>** XML block. This XML node accepts one attribute:

- **name**, ***required string attribute***, user-defined name of this sampler. **Note:** As with other objects, this identifier can be used to reference this specific entity from other input blocks in the XML.

In the **<LimitSurfaceSearch>** input block, the user needs to specify the variables to sample. As already mentioned, these variables are specified within consecutive **<variable>** XML blocks:

- This **<variable>** recognizes the following child nodes:
  - **<distribution>**, ***string, required field***, name of the distribution that is associated to this variable. Its name needs to be contained in the **<Distributions>** block explained in Section 11. In addition, if NDDistribution is used, the attribute **dim** is required. **Note:** Alternatively, this node must be omitted if the **<function>** node is supplied.
  - **<function>**, ***string, required field***, name of the function that defines the calculation of this variable from other distributed variables. Its name needs to be contained in the **<Functions>** block explained in Section 18. This function must implement a method named "evaluate". **Note:** Alternatively, this node must be ommitted if the **<distribution>** node is supplied.

In addition to the **<variable>** nodes, the main XML node **<Adaptive>** needs to contain two supplementary sub-nodes:

- **<Convergence>**, ***float, required field***, Convergence tolerance. The meaning of this tolerance depends on the definition of other attributes that might be defined in this XML node:

- `limit`, ***optional integer attribute***, the maximum number of adaptive samples (iterations).
  Default: infinite.
- `forceIteration`, ***optional boolean attribute***, this attribute controls if at least a number of iterations equal to **limit** must be performed.
  Default: False.
- `weight`, ***optional string attribute (case insensitive)***, defines on what the convergence check needs to be performed.
  - **'CDF'**, the convergence is checked in terms of probability (Cumulative Distribution Function). From a practical point of view, this means that full uncertain domain is discretized in a way that the probability volume of each cell is going to be equal to the tolerance specified in the body of the node **<Convergence>**
  - **'value'**, the convergence is checked on the hyper-volume in terms of variable values.From a practical point of view, this means that full uncertain domain is discretized in a way that the "volume" fraction of each cell is going to be equal to the tolerance specified in the body of the node **<Convergence>**. In other words, each cell volume is going to be equal to the total volume times the tolerance.

  Default: CDF.
- `persistence`, ***optional integer attribute***, offers an additional convergence check. It represents the number of times the computed error needs to be below the inputted tolerance before convergence is reported.
  Default: 5.
- `subGridTol`, ***optional float attribute***, this attribute is used to activate the multi-grid approach (adaptive meshing) of the constructed evaluation grid (see attribute `weight`). In case this attribute is specified, the final grid discretization (cell's "volume content" aka convergence confidence) is represented by the value here specified. The sampler converges on the initial coarse grid, defined by the tolerance specified in the body of the node **<Convergence>**. When the Limit Surface has been identified on the coarse grid, the sampler starts refining the grid until the "volume content" of each cell is equal to the value specified in this attribute (Multi-grid approach).
  Default: None.

In summary, this XML node contains the information that is needed in order to control this sampler's convergence criterion.

- **<batchStrategy>**, ***string, optional field***, defines how points should be selected within a batch of size n where n is given by the **<maxBatchSize>** parameter below. Four options are available:
  - **'none'** If this is specified then the **<maxBatchSize>** parameter below will be ignored and the functionality will replicate the LimitSurfaceSearch, in that the limit

surface will be rebuilt and the points will be re-scored after each trial is completed.

- **'naive'** The top n candidates will be queued for adaptive sampling before retraining the limit surface and re-scoring the new candidate set.
- **'maxP'** The topology of the limit surface given the scoring function values will be decomposed and the top n highest topologically persistent features (local maxima) will be queued for adaptive sampling before retraining and re-scoring the new candidate set.
- **'maxV'** The topology of the limit surface given the scoring function values will be decomposed and the top n highest topological features (local maxima) will be queued for adaptive sampling before retraining and re-scoring the new candidate set.

Default: none.

- **<maxBatchSize>**, ***integer, optional field***, specifies the number of points to select for adaptive sampling before retraining the limit surface and re-scoring the candidates. This is the equivalent of the n parameter used in the **<batchStrategy>** description.
Default: 1.

- **<scoring>**, ***string, optional field***, defines the scoring function to use on the candidate limit surface points in order to select the next adaptive point. Two options are available:

  - **'distance'** will scoring the candidate points by their distance to the closest realized point, in this way preference is given to unexplored regions of the limit surface.
  - **'distancePersistence'** augments the distance above by multiplying it with the inverse persistence of a candidate point which measures how many times the label of the candidate point has changed throughout the lifespan of the algorithm.

  Default: distancePersistence.

- **<simplification>**, ***float in the range [0,1], optional field***, specifies the percent of the scoring function range (on the candidate set) as the amount of topological simplification to do before extracting the topological features from the candidate set (local maxima). This only applies when the **<batchStrategy>** is set to **'maxP'** or **'maxV'**. Thus, one may end up with a batch size less than that specified by **<maxBatchSize>**.
Default: 0.

- **<thickness>**, ***positive integer, optional field***, specifies how much the limit surface should be expanded (in terms of grid distance) when constructing a candidate set. A value of 1 implies only the points bounding the limit surface.
Default: 1.

- **<threshold>**, ***float in the range [0,1], optional field***, once the candidates have been ranked and selected, before queueing them for adaptive sampling, this value is used to

threshold any points whose score is less than this percentage of the scoring function range (on the candidate set). Thus, one may end up with a batch size less than that specified by **<maxBatchSize>**.
Default: 0

- **Assembler Objects** These objects are either required or optional depending on the functionality of the LimitSurfaceSearch Sampler. The objects must be listed with a rigorous syntax that, except for the XML node tag, is common among all the objects. Each of these nodes must contain 2 attributes that are used to identify them within the simulation framework:

  - **class**, ***required string attribute***, the main "class" of the listed object. For example, it can be **'Models'**, **'Functions'**, etc.
  - **type**, ***required string attribute***, the object identifier or sub-type. For example, it can be **'ROM'**, **'External'**, etc.

  The **LimitSurfaceSearch** approach requires or optionally accepts the following object types:

  - **<Function>**, ***string, required field***, the body of this XML block needs to contain the name of an external function object defined within the **<Functions>** main block (see Section 18). This object represents the boolean function that defines the transition boundaries. This function must implement a method called __residuumSign(self), that returns either -1 or 1, depending on the system conditions (see Section 18.
  - **<ROM>**, ***, string, optional field***, if used, the body of this XML node must contain the name of a ROM defined in the **<Models>** block (see Section 17.3). The ROM here specified is going to be used as "acceleration model" to speed up the convergence of the sampling strategy. The **<Target>** XML node in the ROM input block (within the **<Models>** section) needs to match the name of the goal **<Function>** (e.g. if the goal function is named "transitionIdentifier", the **<Target>** of the ROM needs to report the same name: **<Target>transitionIdentifier<Target>**).
  - **<TargetEvaluation>**, ***string, required field***, represents the container where the system evaluations are stored. From a practical point of view, this XML node must contain the name of a data object defined in the **<DataObjects>** block (see Section 14). The object here specified must be input as **<Output>** in the Steps that employ this sampling strategy. The Limit Surface Search sampling accepts "DataObjects" of type "PointSet" only.

Example:

```
<Samplers>
  ...
  <LimitSurfaceSearch name='LSSName'>
    <ROM class='Models' type='ROM'>ROMname</ROM>
```

```
    <Function class='Functions' type='External'
       >FunctionName</Function>
    <TargetEvaluation class='DataObjects'
       type='PointSet'>DataName</TargetEvaluation>
    <Convergence limit='3000'  forceIteration='False'
       weight='CDF'  subGridTol='1e-4' persistence='5'>
      1e-2
    </Convergence>
    <variable name='var1'>
      <distribution>aDistributionNameDefinedInDistributionBlock1
         </distribution>
    </variable>
    <variable name='var2'>
      <distribution>aDistributionNameDefinedInDistributionBlock2
         </distribution>
    </variable>
    <variable name='var3'>
      <distribution>aDistributionNameDefinedInDistributionBlock3
         </distribution>
    </variable>
  </LimitSurfaceSearch>
  ...
</Samplers>
```

Batch sampling Example:

```
<Samplers>
  ...
  <LimitSurfaceSearch name='LSBSName'>
    <ROM class='Models' type='ROM'>ROMname</ROM>
    <Function class='Functions' type='External'
       >FunctionName</Function>
    <TargetEvaluation class='DataObjects'
       type='PointSet'>DataName</TargetEvaluation>
    <Convergence limit='3000'  forceIteration='False'
       weight='CDF'  subGridTol='1e-4' persistence='5'>
      1e-2
    </Convergence>
    <scoring>distancePersistence</scoring>
    <batchStrategy>maxP</batchStrategy>
    <thickness>1</thickness>
    <maxBatchSize>4</maxBatchSize>
    <variable name='var1'>
```

```
      <distribution>aDistributionNameDefinedInDistributionBlock1
         </distribution>
    </variable>
    <variable name='var2'>
      <distribution>aDistributionNameDefinedInDistributionBlock2
         </distribution>
    </variable>
    <variable name='var3'>
      <distribution>aDistributionNameDefinedInDistributionBlock3
         </distribution>
    </variable>
  </LimitSurfaceSearch>
  ...
</Samplers>
```

Associated External Python Module:

```
def __residuumSign(self):
  if self.whatEverValue < self.OtherValue :
    return  1
  else:
    return -1
```

12.3.2 Adaptive Dynamic Event Tree

The **Adaptive Dynamic Event Tree** approach is an advanced methodology employing a smart sampling around transition zones that determine a change in the status of the system (limit surface), using the support of a Dynamic Event Tree methodology. The main idea of the application of the previously explained adaptive sampling approach to the DET comes from the observation that the DET, when evaluated from a limit surface perspective, is intrinsically adaptive. For this reason, it appears natural to use the DET approach to perform a goal-function oriented pre-sampling of the input space.

RAVEN uses ROMs for predicting, in the input space, the location(s) of these transitions, in order to accelerate the exploration of the input space in proximity of the limit surface.

The specifications of this sampler must be defined within an `<AdaptiveDynamicEventTree>` XML block. This XML node accepts one attribute:

- `name`, ***required string attribute***, user-defined name of this sampler. **Note:** As with other objects, this identifier can be used to reference this specific entity from other input blocks in the XML.

- `printEndXmlSummary`, ***optional string/boolean attribute***, this attribute controls the dumping of a "summary" of the DET performed in to an external XML.
  Default: False.
- `maxSimulationTime`, ***optional float attribute***, this attribute controls the maximum "mission" time of the simulation underneath.
  Default: None.
- `mode`, ***optional string attribute***, controls when the adaptive search needs to begin. Two options are available:
  - `'post'`, if this option is activated, the sampler first performs a standard Dynamic Event Tree analysis. At end of it, it uses the outcomes to start the adaptive search in conjunction with the DET support.
  - `'online'`, if this option is activated, the adaptive search starts at the beginning, during the initial standard Dynamic Event Tree analysis. Whenever a transition is detected, the **Adaptive Dynamic Event Tree** starts its goal-oriented search using the DET as support;

  Default: post.
- `updateGrid`, ***optional boolean attribute***, if true, each adaptive request is going to update the meshing of the initial DET grid.
  Default: True.

In the `<AdaptiveDynamicEventTree>` input block, the user needs to specify the variables to sample. As already mentioned, these variables are specified within consecutive `<variable>` XML blocks:

- `<variable>`, ***XML node, required parameter*** will specify one attribute:
  - `name`, ***required string attribute***, user-defined name of this variable.

  This `<variable>` recognizes the following child nodes:
  - `<distribution>`, ***string, required field***, name of the distribution that is associated to this variable. Its name needs to be contained in the `<Distributions>` block explained in Section 11. In addition, if NDDistribution is used, the attribute `dim` is required. **Note:** Alternatively, this node must be omitted if the `<function>` node is supplied.
  - `<function>`, ***string, required field***, name of the function that defines the calculation of this variable from other distributed variables. Its name needs to be contained in the `<Functions>` block explained in Section 18. This function must implement a method named "evaluate". **Note:** Alternatively, this node must be ommitted if the `<distribution>` node is supplied.

– **<grid>**, ***space separated floats, required field***, the content of this XML node depends on the definition of the associated attributes:

  * **type**, ***required string attribute***, user-defined discretization metric type: 1) **'CDF'**, the grid will be specified based on cumulative distribution function probability thresholds, and 2) **'value'**, the grid will be provided using variable values.
  * **construction**, ***required string attribute***, how the grid needs to be constructed, independent of its type (i.e. **'CDF'** or **'value'**).

  Based on the **construction** type, the content of the **<grid>** XML node and the requirements for other attributes change:

  * **construction='equal'**. The grid is going to be constructed equally-spaced (**type='value'**) or equally probable (**type='CDF'**). This construction type requires the definition of additional attributes:
    · **steps**, ***required integer attribute***, number of equally spaced/probable discretization steps.

    This construction type requires that the content of the **<grid>** node represents the lower and upper bounds (either in probability or value). Two values need to be specified; the lowest one will be considered as the $lowerBound$, the largest, the $upperBound$. The lower and upper bounds are checked against the associated **<distribution>** bounds. If one or both of them falls outside the distribution's bounds, the code will raise an error. The $stepSize$ is determined as follows:
    $stepSize = (upperBound - lowerBound)/steps$
  * **construction='custom'**. The grid will be directly specified by the user. No additional attributes are needed. This construction type requires that the **<grid>** node contains the actual mesh bins. For example, if the grid **type** is **'CDF'**, in the body of **<grid>**, the user will specify the CDF probability thresholds (nodalization in probability). All the bins are checked against the associated **<distribution>** bounds. If one or more of them falls outside the distribution's bounds, the code will raise an error.

  Note: The **<grid>** node is only required if a **<distribution>** node is supplied. In the case of a **<function>** node, no grid information is requested.

• **<constant>**, ***XML node, optional parameter*** the user is able to input variables that need to be kept constant. For doing this, as many **<constant>** nodes as needed can be input, where the body of the node contains the constant value that is going to be injected as an additional variable

In addition to the **<variable>** nodes, the main **<AdaptiveDynamicEventTree>** node needs to contain two supplementary sub-nodes:

• **<Convergence>**, ***float, required field***, Convergence tolerance. The meaning of this tolerance depends on the definition of other attributes that might be defined in this XML node:

- `limit`, ***optional integer attribute***, the maximum number of adaptive samples (iterations).
  Default: infinite.
- `forceIteration`, ***optional boolean attribute***, this attribute controls if at least a number of iterations equal to **limit** must be performed.
  Default: False.
- `weight`, ***optional string attribute (case insensitive)***, defines on what the convergence check needs to be performed.
  * **'CDF'**, the convergence is checked in terms of probability (Cumulative Distribution Function). From a practical point of view, this means that full uncertain domain is discretized in a way that the probability volume of each cell is going to be equal to the tolerance specified in the body of the node **<Convergence>**
  * **'value'**, the convergence is checked on the hyper-volume in terms of variable values.From a practical point of view, this means that full uncertain domain is discretized in a way that the "volume" fraction of each cell is going to be equal to the tolerance specified in the body of the node **<Convergence>**. In other words, each cell volume is going to be equal to the total volume times the tolerance.

  Default: CDF.
- `persistence`, ***optional integer attribute***, offers an additional convergence check. It represents the number of times the computed error needs to be below the inputted tolerance before convergence is reported.
  Default: 5.
- `subGridTol`, ***optional float attribute***, this attribute is used to activate the multi-grid approach (adaptive meshing) of the constructed evaluation grid (see attribute `weight`). In case this attribute is specified, the final grid discretization (cell's "volume content" aka convergence confidence) is represented by the value here specified. The sampler converges on the initial coarse grid, defined by the tolerance specified in the body of the node **<Convergence>**. When the Limit Surface has been identified on the coarse grid, the sampler starts refining the grid until the "volume content" of each cell is equal to the value specified in this attribute (Multi-grid approach).
  Default: None.

In summary, this XML node contains the information that is needed in order to control this sampler's convergence criterion.

- **Assembler Objects** These objects are either required or optional depending on the functionality of the AdaptiveDynamicEventTree Sampler. The objects must be listed with a rigorous syntax that, except for the XML node tag, is common among all the objects. Each of these nodes must contain 2 attributes that are used to identify them within the simulation framework:

– class, ***required string attribute***, the main "class" of the listed object. For example, it can be **'Models'**, **'Functions'**, etc.

– type, ***required string attribute***, the object identifier or sub-type. For example, it can be **'ROM'**, **'External'**, etc.

The **AdaptiveDynamicEventTree** approach requires or optionally accepts the following object types:

– **<Function>**, ***string, required field***, the body of this XML block needs to contain the name of an external function object defined within the **<Functions>** main block (see Section 18). This object represents the boolean function that defines the transition boundaries. This function must implement a method called __residuumSign(self), that returns either -1 or 1, depending on the system conditions (see Section 18.

– **<ROM>**, ***, string, optional field***, if used, the body of this XML node must contain the name of a ROM defined in the **<Models>** block (see Section 17.3). The ROM here specified is going to be used as "acceleration model" to speed up the convergence of the sampling strategy. The **<Target>** XML node in the ROM input block (within the **<Models>** section) needs to match the name of the goal **<Function>** (e.g. if the goal function is named "transitionIdentifier", the **<Target>** of the ROM needs to report the same name: **<Target>transitionIdentifier<Target>**).

– **<TargetEvaluation>**, ***string, required field***, represents the container where the system evaluations are stored. From a practical point of view, this XML node must contain the name of a data object defined in the **<DataObjects>** block (see Section 14). The adaptive sampling accepts "DataObjects" of type "PointSet" only.

Example:

```
<Samplers>
  ...
  <AdaptiveDynamicEventTree name = 'AdaptiveName'>
    <ROM class = 'Models' type = 'ROM'ROMname</ROM>
    <Function class = 'Functions' type =
       'External'>FunctionName</Function>
    <TargetEvaluation class = 'DataObjects' type =
       'PointSet'>DataName</TargetEvaluation>
    <Convergence limit = '3000' subGridTol= '0.001'
       forceIteration = 'False' weight = 'CDF'
       subGriTol='''1e-5' persistence = '5'>
      1e-2
    </Convergence>
    <variable name = 'var1'>
        <distribution>
```

```xml
        aDistributionNameDefinedInDistributionBlock1
      </distribution>
      <grid type='CDF' construction='custom'>0.1 0.8</grid>
    </variable>
    <variable name = 'var2'>
      <distribution>
        aDistributionNameDefinedInDistributionBlock2
      </distribution>
      <grid type='CDF' construction='custom'>0.1 0.8</grid>
    </variable>
    <variable name = 'var3'>
      <distribution>
        aDistributionNameDefinedInDistributionBlock3
      </distribution>
      <grid type='CDF' construction='custom'>0.1 0.8</grid>
    </variable>
  </AdaptiveDynamicEventTree>
  ...
</Samplers>
```

Associated External Python Module:

```python
def __residuumSign(self):
  if self.whatEverValue < self.OtherValue:
    return  1
  else:
    return -1
```

12.3.3 Adaptive Hybrid Dynamic Event Tree

The **Adaptive Hybrid Dynamic Event Tree** approach is an advanced methodology employing a smart sampling around transition zones that determine a change in the status of the system (limit surface), using the support of the Hybrid Dynamic Event Tree methodology. Practically, this methodology represents a conjunction between the previously described Adaptive DET and the Hybrid DET method for the treatment of the epistemic variables.

Regarding the input requirements, the AHDET sampler is a "sub-type" of the **<AdaptiveDynamicEventTree>** sampler. For this reason, its specifications must be defined within a **<AdaptiveDynamicEventTree>** block.

The specifications of this sampler must be defined within an **<AdaptiveDynamicEventTree>** XML block. This XML node accepts one attribute:

- **name**, ***required string attribute***, user-defined name of this sampler. **Note:** As with other objects, this identifier can be used to reference this specific entity from other input blocks in the XML.
- **printEndXmlSummary**, ***optional string/boolean attribute***, this attribute controls the dumping of a "summary" of the DET performed in to an external XML.
  Default: False.
- **maxSimulationTime**, ***optional float attribute***, this attribute controls the maximum "mission" time of the simulation underneath.
  Default: None.
- **mode**, ***optional string attribute***, controls when the adaptive search needs to begin. Two options are available:
  - **'post'**, if this option is activated, the sampler first performs a standard Dynamic Event Tree analysis. At end of it, it uses the outcomes to start the adaptive search in conjunction with the DET support.
  - **'online'**, if this option is activated, the adaptive search starts at the beginning, during the initial standard Dynamic Event Tree analysis. Whenever a transition is detected, the **Adaptive Dynamic Event Tree** starts its goal-oriented search using the DET as support;

  Default: post.
- **updateGrid**, ***optional boolean attribute***, if true, each adaptive request is going to update the meshing of the initial DET grid.
  Default: True.

In the **<AdaptiveDynamicEventTree>** input block, the user needs to specify the variables to sample. As already mentioned, these variables are specified within consecutive **<variable>** XML blocks:

- **<variable>**, ***XML node, required parameter*** will specify one attribute:
  - **name**, ***required string attribute***, user-defined name of this variable.

  This **<variable>** recognizes the following child nodes:
  - **<distribution>**, ***string, required field***, name of the distribution that is associated to this variable. Its name needs to be contained in the **<Distributions>** block explained in Section 11. In addition, if NDDistribution is used, the attribute **dim** is required. **Note:** Alternatively, this node must be omitted if the **<function>** node is supplied.

- **<function>**, ***string, required field***, name of the function that defines the calculation of this variable from other distributed variables. Its name needs to be contained in the **<Functions>** block explained in Section 18. This function must implement a method named "evaluate". **Note:** Alternatively, this node must be ommitted if the **<distribution>** node is supplied.
- **<grid>**, ***space separated floats, required field***, the content of this XML node depends on the definition of the associated attributes:
  * **type**, ***required string attribute***, user-defined discretization metric type: 1) **'CDF'**, the grid will be specified based on cumulative distribution function probability thresholds, and 2) **'value'**, the grid will be provided using variable values.
  * **construction**, ***required string attribute***, how the grid needs to be constructed, independent of its type (i.e. **'CDF'** or **'value'**).

  Based on the **construction** type, the content of the **<grid>** XML node and the requirements for other attributes change:

  * **construction='equal'**. The grid is going to be constructed equally-spaced (**type='value'**) or equally probable (**type='CDF'**). This construction type requires the definition of additional attributes:
    - **steps**, ***required integer attribute***, number of equally spaced/probable discretization steps.

    This construction type requires that the content of the **<grid>** node represents the lower and upper bounds (either in probability or value). Two values need to be specified; the lowest one will be considered as the $lowerBound$, the largest, the $upperBound$. The lower and upper bounds are checked against the associated **<distribution>** bounds. If one or both of them falls outside the distribution's bounds, the code will raise an error. The $stepSize$ is determined as follows: $stepSize = (upperBound - lowerBound)/steps$
  * **construction='custom'**. The grid will be directly specified by the user. No additional attributes are needed. This construction type requires that the **<grid>** node contains the actual mesh bins. For example, if the grid **type** is **'CDF'**, in the body of **<grid>**, the user will specify the CDF probability thresholds (nodalization in probability). All the bins are checked against the associated **<distribution>** bounds. If one or more of them falls outside the distribution's bounds, the code will raise an error.

  Note: The **<grid>** node is only required if a **<distribution>** node is supplied. In the case of a **<function>** node, no grid information is requested.

- **<constant>**, ***XML node, optional parameter*** the user is able to input variables that need to be kept constant. For doing this, as many **<constant>** nodes as needed can be input, where the body of the node contains the constant value that is going to be injected as an additional variable

In addition to the `<variable>` nodes, the main `<AdaptiveDynamicEventTree>` node needs to contain two supplementary sub-nodes:

- `<Convergence>`, ***float, required field***, Convergence tolerance. The meaning of this tolerance depends on the definition of other attributes that might be defined in this XML node:
  - `limit`, ***optional integer attribute***, the maximum number of adaptive samples (iterations).
    Default: infinite.
  - `forceIteration`, ***optional boolean attribute***, this attribute controls if at least a number of iterations equal to **limit** must be performed.
    Default: False.
  - `weight`, ***optional string attribute (case insensitive)***, defines on what the convergence check needs to be performed.
    * `'CDF'`, the convergence is checked in terms of probability (Cumulative Distribution Function). From a practical point of view, this means that full uncertain domain is discretized in a way that the probability volume of each cell is going to be equal to the tolerance specified in the body of the node `<Convergence>`
    * `'value'`, the convergence is checked on the hyper-volume in terms of variable values.From a practical point of view, this means that full uncertain domain is discretized in a way that the "volume" fraction of each cell is going to be equal to the tolerance specified in the body of the node `<Convergence>`. In other words, each cell volume is going to be equal to the total volume times the tolerance.

    Default: CDF.
  - `persistence`, ***optional integer attribute***, offers an additional convergence check. It represents the number of times the computed error needs to be below the inputted tolerance before convergence is reported.
    Default: 5.
  - `subGridTol`, ***optional float attribute***, this attribute is used to activate the multi-grid approach (adaptive meshing) of the constructed evaluation grid (see attribute `weight`). In case this attribute is specified, the final grid discretization (cell's "volume content" aka convergence confidence) is represented by the value here specified. The sampler converges on the initial coarse grid, defined by the tolerance specified in the body of the node `<Convergence>`. When the Limit Surface has been identified on the coarse grid, the sampler starts refining the grid until the "volume content" of each cell is equal to the value specified in this attribute (Multi-grid approach).
    Default: None.

  In summary, this XML node contains the information that is needed in order to control this sampler's convergence criterion.

- **Assembler Objects** These objects are either required or optional depending on the functionality of the AdaptiveDynamicEventTree Sampler. The objects must be listed with a rigorous syntax that, except for the XML node tag, is common among all the objects. Each of these nodes must contain 2 attributes that are used to identify them within the simulation framework:

  – `class`, ***required string attribute***, the main "class" of the listed object. For example, it can be `'Models'`, `'Functions'`, etc.

  – `type`, ***required string attribute***, the object identifier or sub-type. For example, it can be `'ROM'`, `'External'`, etc.

  The **AdaptiveDynamicEventTree** approach requires or optionally accepts the following object types:

  – `<Function>`, ***string, required field***, the body of this XML block needs to contain the name of an external function object defined within the `<Functions>` main block (see Section 18). This object represents the boolean function that defines the transition boundaries. This function must implement a method called `__residuumSign(self)`, that returns either -1 or 1, depending on the system conditions (see Section 18.

  – `<ROM>`, ***, string, optional field***, if used, the body of this XML node must contain the name of a ROM defined in the `<Models>` block (see Section 17.3). The ROM here specified is going to be used as "acceleration model" to speed up the convergence of the sampling strategy. The `<Target>` XML node in the ROM input block (within the `<Models>` section) needs to match the name of the goal `<Function>` (e.g. if the goal function is named "transitionIdentifier", the `<Target>` of the ROM needs to report the same name: `<Target>`**transitionIdentifier**`<Target>`).

  – `<TargetEvaluation>`, ***string, required field***, represents the container where the system evaluations are stored. From a practical point of view, this XML node must contain the name of a data object defined in the `<DataObjects>` block (see Section 14). The adaptive sampling accepts "DataObjects" of type "PointSet" only.

As it can be noticed, the basic specifications of the Adaptive Hybrid Dynamic Event Tree method are consistent with the ones for the ADET methodology. In order to activate the **Adaptive Hybrid Dynamic Event Tree** sampler, the main `<AdaptiveDynamicEventTree>` block needs to contain an additional sub-node called `<HybridSampler>`. This sub-node needs to contain the following attribute:

- `type`, ***required string attribute***, type of pre-sampling strategy to be used. Up to now only one option is available:

  – `'LimitSurface'`. With this option, the epistemic variables here listed are going to be part of the LS search. This means that the discretization of the domain of these variables is determined by the `<Convergece>` node.

Independent of the type of HybridSampler that has been specified, the **<HybridSampler>** must contain the variables that need to be sampled. As already mentioned, these variables are specified within consecutive **<variable>** XML blocks:

- **<variable>**, ***XML node, required parameter*** will specify one attribute:
  - **name**, ***required string attribute***, user-defined name of this variable.

  This **<variable>** recognizes the following child nodes:
  - **<distribution>**, ***string, required field***, name of the distribution that is associated to this variable. Its name needs to be contained in the **<Distributions>** block explained in Section 11. In addition, if NDDistribution is used, the attribute **dim** is required. **Note:** Alternatively, this node must be omitted if the **<function>** node is supplied.
  - **<function>**, ***string, required field***, name of the function that defines the calculation of this variable from other distributed variables. Its name needs to be contained in the **<Functions>** block explained in Section 18. This function must implement a method named "evaluate". **Note:** Alternatively, this node must be ommitted if the **<distribution>** node is supplied.
- **<constant>**, ***XML node, optional parameter*** the user is able to input variables that need to be kept constant. For doing this, as many **<constant>** nodes as needed can be input, where the body of the node contains the constant value that is going to be injected as an additional variable

Example:

```
<Samplers>
  ...
  <AdaptiveDynamicEventTree name = 'AdaptiveName'>
    <ROM class = 'Models' type = 'ROM'ROMname</ROM>
    <Function class = 'Functions' type =
       'External'>FunctionName</Function>
    <TargetEvaluation class = 'DataObjects' type =
       'PointSet'>DataName</TargetEvaluation>
    <Convergence limit = '3000' subGridTol= '0.001'
       forceIteration = 'False' weight = 'CDF'
       subGriTol='''1e-5' persistence = '5'>
      1e-2
    </Convergence>
    <HybridSampler type='LimitSurface'>
      <variable name = 'epistemicVar1'>
        <distribution>
```

```xml
          aDistributionNameDefinedInDistributionBlock1
        </distribution>
      </variable>
       <variable name = 'epistemicVar2'>
        <distribution>
          aDistributionNameDefinedInDistributionBlock2
        </distribution>
      </variable>
    </HybridSampler>
    <variable name = 'var1'>
        <distribution>
         aDistributionNameDefinedInDistributionBlock3
        </distribution>
        <grid type='CDF' construction='custom'>0.1 0.8</grid>
    </variable>
    <variable name = 'var2'>
        <distribution>
          aDistributionNameDefinedInDistributionBlock4
        </distribution>
        <grid type='CDF' construction='custom'>0.1 0.8</grid>
    </variable>
    <variable name = 'var3'>
        <distribution>
          aDistributionNameDefinedInDistributionBlock5
        </distribution>
        <grid type='CDF' construction='custom'>0.1 0.8</grid>
    </variable>

  </AdaptiveDynamicEventTree>
  ...
</Samplers>
```

Associated External Python Module:

```python
def __residuumSign(self):
  if self.whatEverValue < self.OtherValue:
    return  1
  else:
    return -1
```

12.3.4 Adaptive Sparse Grid

The **Adaptive Sparse Grid** approach is an advanced methodology that employs an intelligent search for the most suitable sparse grid quadrature to characterize a model. To perform such sampling, RAVEN adaptively builds an index set and generates sparse grids in a similar manner to Sparse Grid Collocation samplers. In each iterative step, the adaptive index set determines the next possible quadrature orders to add in each dimension, and determines the index set point that would offer the largest impact to one of the convergence metrics. This process continues until the total impact of all the potential index set points is less than tolerance. For many models, this function converges after fewer runs than a traditional Sparse Grid Collocation sampling. However, it should be noted that this algorithm fails in the event that the partial derivative of the response surface with respect to any single input dimension is zero at the origin of the input domain. For example, the adaptive algorithm fails for the model $f(x) = x \cdot y$.

The specifications of this sampler must be defined within an **<Adaptive Sparse Grid>** XML block. This XML node accepts one attribute:

- **name**, ***required string attribute***, user-defined name of this sampler. **Note:** As with other objects, this identifier can be used to reference this specific entity from other input blocks in the XML.

In the **<Adaptive Sparse Grid>** input block, the user needs to specify the variables to sample. As already mentioned, these variables are specified within consecutive **<variable>** XML blocks:

- **<variable>**, ***XML node, required parameter*** will specify one attribute:
  - **name**, ***required string attribute***, user-defined name of this variable.

  This **<variable>** recognizes the following child nodes:
  - **<distribution>**, ***string, required field***, name of the distribution that is associated to this variable. Its name needs to be contained in the **<Distributions>** block explained in Section 11. In addition, if NDDistribution is used, the attribute **dim** is required. **Note:** Alternatively, this node must be omitted if the **<function>** node is supplied.
  - **<function>**, ***string, required field***, name of the function that defines the calculation of this variable from other distributed variables. Its name needs to be contained in the **<Functions>** block explained in Section 18. This function must implement a method named “evaluate”. **Note:** Alternatively, this node must be ommitted if the **<distribution>** node is supplied.

- **<constant>**, ***XML node, optional parameter*** the user is able to input variables that need to be kept constant. For doing this, as many **<constant>** nodes as needed can be input, where the body of the node contains the constant value that is going to be injected as an additional variable

In addition to the **<variable>** nodes, the main XML node **<AdaptiveSparseGrid>** needs to contain the following supplementary sub-nodes:

- **<Convergence>**,***float, required field***, Convergence tolerance. The meaning of this tolerance depends on the **target** attribute of this node.
  - **target**, ***required string attribute***, the metric for convergence. The following metrics are available: **'variance'**, which converges the sparse quadrature integration of the second moment of the model.
  - **maxPolyOrder**, ***optional integer attribute***, limits the maximum size equivalent polynomial for any one dimension.
    Default: 10.
  - **persistence**, ***optional integer attribute***, defines the number of index set points that are required to be found before calculation can exit. Setting this to a higher value can help if the adaptive process is not finding significant indices on its own.
    Default: 2.

  In summary, this XML node contains the information that is needed in order to control this sampler's convergence criterion.

- **<convergenceStudy>**, ***optional node***, if included, triggers writing state points at particular numbers of model solves for the purpose of a convergence study. The study is performed by writing XML output files as described in the OutStreams for ROMs at the state points requested, using **'all'** as the requested **<what>** values. The state points are identified when a certain number of model runs is passed, as specified by the **<runStatePoints>** node. This node has the following sub-nodes to define its parameters:
  - **<runStatePoints>**, ***list of integers, required node***, lists the number of model runs at which state points should be written. Note that these will be written when the requested number of runs is met or passed, so the actual value is often somewhat more than the requested value, and the exact value will be listed in the XML output.
  - **<baseFilename>**, ***string, optional node***, if specified determines the base file name for the state point outputs. If not specified, defaults to **'out_'**.
  - **<pickle>**, ***no text, optional node***, if this node is included, serialized (pickled) versions of the ROM at each of the run states is also created in the working directory, with the format `<baseFilename><numRuns>.pk`, such as `out_100.pk`.

- **<logFile>**, ***optional node***, if included, the log file onto which the adaptive step progress can be printed. The log includes the values of included polynomial coefficients as well as the expected impacts of polynomial coefficients not yet included. This is different from the convergenceStudy print, which will give statistical moments at certain steps.

- **<maxRuns>**, ***optional node***, if included, the adaptive sampler will track the number of computational solves necessary to construct the associated GaussPolynomialROM. If at any point the number of solves exceeds the value given, it will not initiate any additional solves, and will exit when existing solves finish.

Assembler Objects These objects are either required or optional depending on the functionality of the Adaptive Sparse Grid Sampler. The objects must be listed with a rigorous syntax that, except for the XML node tag, is common among all the objects. Each of these nodes must contain 2 attributes that are used to identify them within the simulation framework:

- **class**, ***required string attribute***, the main "class" of the listed object. For example, it can be **'Models'**, **'Functions'**, etc.

- **type**, ***required string attribute***, the object identifier or sub-type. For example, it can be **'ROM'**, **'External'**, etc.

The **Adaptive Sparse Grid** approach requires or optionally accepts the following object types:

- **<ROM>**, ***string, required field***, the body of this XML node must contain the name of an appropriate ROM defined in the **<Models>** block (see Section 17.3).

- **<TargetEvaluation>**, ***string, required field***, represents the container where the system evaluations are stored. From a practical point of view, this XML node must contain the name of a data object defined in the **<DataObjects>** block (see Section 14). The Adaptive Sparse Grid sampling accepts "DataObjects" of type "PointSet" only.

Example:

```
<Samplers>
  ...
  <AdaptiveSparseGrid name="ASG" verbosity='debug'>
    <Convergence target='coeffs'>1e-2</Convergence>
    <variable name="x1">
      <distribution>UniDist</distribution>
    </variable>
    <variable name="x2">
```

```
      <distribution>UniDist</distribution>
    </variable>
    <ROM class = 'Models' type = 'ROM'>gausspolyrom</ROM>
    <TargetEvaluation class = 'DataObjects' type =
       'PointSet'>solns</TargetEvaluation>
  </AdaptiveSparseGrid>
  ...
</Samplers>
```

Like in the **SparseGridCollocation** sampler, if the covariance matrix is provided and the input parameters are assumed to have the multivariate normal distribution, the **AdaptiveSparseGrid** can be also used. This means one creates the sparse grids of variables listed by **<latentVariables>** in the transformed space. If this is the case, the user needs to provide additional information, i.e. the **<transformation>** under **<MultivariateNormal>** of **<Distributions>** (more information can be found in Section 11.2). In addition, the node **<variablesTransformation>** is also required for **AdaptiveSparseGrid** sampler. This node is used to tranform the variables specified by **<latentVariables>** in the transformed space of input into variables spefified by **<manifestVariables>** in the input space. The variables listed in **<latentVariables>** should be predefined in **<variable>**, and the variables listed in **<manifestVariables>** are used by the **<Models>**.

- **<variablesTransformation>**, *optional field*. this XML node accepts one attribute:
  - distribution, ***required string attribute***, the name for the distribution defined in the XML node **<Distributions>**. This attribute indicates the values of **<manifestVariables>** are drawn from distribution.

  In addition, this XML node also accepts three childen nodes:

  - **<latentVariables>**, ***comma separated string, required field***, user-defined latent variables that are used for the variables transformation. All the variables listed under this node should be also mentioned in **<variable>**.
  - **<manifestVariables>**, ***comma separated string, required field***, user-defined manifest variables that can be used by the model.
  - **<manifestVariablesIndex>**, ***comma separated string, optional field***, user-defined manifest variables indices paired with **<manifestVariables>**. These indices indicate the position of manifest variables associated with multivariate normal distribution defined in the XML node **<Distributions>**. The indices should be postive integer. If not provided, the code will use the positions of manifest variables listed in **<manifestVariables>** as the indices.
  - **<method>**, ***string, required field***, the method that is used for the variables transformation. The currently available method is '**pca**'.

```
...
<Models>
    ...
    <ExternalModel ModuleToLoad="lorentzAttractor_noK"
       name="PythonModule" subType="">
        <variables>sigma,rho,beta,x,y,z,time,x0,y0,z0</variables>
    </ExternalModel>
    <ROM name="gausspolyrom" subType="GaussPolynomialRom">
        <Target>ans</Target>
        <Features>x1,y1,z1</Features>
        <IndexSet>TensorProduct</IndexSet>
        <PolynomialOrder>1</PolynomialOrder>
    </ROM>
    ...
</Models>

<Distributions>
    ...
    <MultivariateNormal name='MVNDist' method='pca'>
        <transformation>
            <rank>3</rank>
        </transformation>
        <mu>0.0 1.0 2.0</mu>
        <covariance type="abs">
            1.0         0.6        -0.4
            0.6         1.0        0.2
            -0.4        0.2        0.8
        </covariance>
    </MultivariateNormal>
    ...
</Distributions>

<Samplers>
  ...
  <AdaptiveSparseGrid name='ASC'>
        <variable name='x0'>
            <distribution dim='1'>MVNDist</distribution>
        </variable>
        <variable name='y0'>
            <distribution dim='2'>MVNDist</distribution>
        </variable>
        <variable name='z0'>
```

```
            <distribution dim='3'>MVNDist</distribution>
        </variable>
        <variablesTransformation model="PythonModule">
            <latentVariables>x1,y1,z1</latentVariables>
            <manifestVariables>x0,y0,z0</manifestVariables>
            <method>pca</method>
        </variablesTransformation>
        <ROM class = 'Models' type = 'ROM'>gausspolyrom</ROM>
        <TargetEvaluation class = 'DataObjects' type =
          'PointSet'>solns</TargetEvaluation>
  </AdaptiveSparseGrid>
  ...
</Samplers>
...
```

12.3.5 Adaptive Sobol Decomposition

The **Adaptive Sobol Decomposition** approach is an advanced methodology that decomposes an uncertainty space into subsets and adaptively includes the most influential ones. For example, for a response function $f(a, b, c)$, the full list of subsets include (a), (b), (c), (a, b), (a, c), (b, c), (a, b, c). A Gauss Polynomial ROM is constructed for each included subset using the Adaptive Sparse Grid sampler. The importance of each subset is estimated based on the importance of preceding subsets; that is, the impact of (a, b) on the representation of f is estimated using the impact of (a) and (b). Because of the excellent performance of Gauss Polynomial ROMs for small-dimension spaces, this sampler used to construct an HDMR ROM can be very efficient. Note that the ROM specified for this sampler *must* be an HDMRRom specified in the Models block.

The specifications of this sampler must be defined within an **<Adaptive Sobol>** XML block. This XML node accepts one attribute:

- **name**, ***required string attribute***, user-defined name of this sampler. **Note:** As with other objects, this identifier can be used to reference this specific entity from other input blocks in the XML.

In the **<Adaptive Sobol>** input block, the user needs to specify the variables to sample. As already mentioned, these variables are specified within consecutive **<variable>** XML blocks:

- **<variable>**, ***XML node, required parameter*** will specify one attribute:
  - **name**, ***required string attribute***, user-defined name of this variable.

This <variable> recognizes the following child nodes:

  - <distribution>, ***string, required field***, name of the distribution that is associated to this variable. Its name needs to be contained in the <Distributions> block explained in Section 11. In addition, if NDDistribution is used, the attribute dim is required. **Note:** Alternatively, this node must be omitted if the <function> node is supplied.
  - <function>, ***string, required field***, name of the function that defines the calculation of this variable from other distributed variables. Its name needs to be contained in the <Functions> block explained in Section 18. This function must implement a method named "evaluate". **Note:** Alternatively, this node must be ommitted if the <distribution> node is supplied.

- <constant>, ***XML node, optional parameter*** the user is able to input variables that need to be kept constant. For doing this, as many <constant> nodes as needed can be input, where the body of the node contains the constant value that is going to be injected as an additional variable

In addition to the <variable> nodes, the main XML node <AdaptiveSobol> needs to contain the following supplementary sub-nodes:

- <Convergence>, ***required node***, Convergence properties. This node contains the following properties that can be set by sub-nodes:
  - <relTolerance>, ***required float***, the relative tolerance to converge. This will compare to the estimate of subset polynomial errors and additional subset polynomials over the variance of the expansion so far to determine convergence.
  - <maxRuns>, ***optional integer field***, a limit for the number of model calls. Once this limit is reached, no additional subsets will be generated or considered; however, existing subsets will continue to be trained. If not specified, no limit on solves is imposed.
  - <maxSobolOrder>, ***optional integer field***, the largest polynomials orders to use in subset GaussPolynomialRom objects. If specified, polynomial indices with a value larger than the value given will be rejected during adaptive construction.
  - <progressParam>, ***optional float field***, a favoritism parameter ranging between 0 and 2. At 0, the algorithm will always prefer adding polynomials to adding new subsets in the HDMR expansion. At 2, the opposite is true. Default is 1.
  - <logFile>, ***optional string field***, a file to which adaptive progress is recorded. If specified, each adaptive step will trigger printing progress to the file given, including the estimated error at the step, the next adaptive step to take, the coefficient of each polynomial within each gPC expansion, and the actual and expected Sobol sensitivities of each HDMR subset. Default is no printing.

- **<subsetVerbosity>**, ***optional string field***, the verbosity for components constructed during the adaptive HDMR process. Options are *silent*, *quiet*, *all*, or *debug*, in order of verbosity. If an invalid entry is provided, will resort to default. Default is *quiet*.

  In summary, this XML node contains the information that is needed in order to control this sampler's convergence criterion.

- **<convergenceStudy>**, ***optional node***, if included, triggers writing state points at particular numbers of model solves for the purpose of a convergence study. The study is performed by writing XML output files as described in the OutStreams for ROMs at the state points requested, using **'all'** as the requested **<what>** values. The state points are identified when a certain number of model runs is passed, as specified by the **<runStatePoints>** node. This node has the following sub-nodes to define its parameters:

  - **<runStatePoints>**, ***list of integers, required node***, lists the number of model runs at which state points should be written. Note that these will be written when the requested number of runs is met or passed, so the actual value is often somewhat more than the requested value, and the exact value will be listed in the XML output.
  - **<baseFilename>**, ***string, optional node***, if specified determines the base file name for the state point outputs. If not specified, defaults to **'out_'**.
  - **<pickle>**, ***no text, optional node***, if this node is included, serialized (pickled) versions of the ROM at each of the run states is also created in the working directory, with the format `<baseFilename><numRuns>.pk`, such as `out_100.pk`.

  Like the **Sobol**, if multivariate normal distribution is provided, the following node need to be specified:

-
  - **<variablesTransformation>**, ***optional field***. this XML node accepts one attribute:
    * **distribution**, ***required string attribute***, the name for the distribution defined in the XML node **<Distributions>**. This attribute indicates the values of **<manifestVariables>** are drawn from **distribution**.

    In addition, this XML node also accepts three childen nodes:

    * **<latentVariables>**, ***comma separated string, required field***, user-defined latent variables that are used for the variables transformation. All the variables listed under this node should be also mentioned in **<variable>**.
    * **<manifestVariables>**, ***comma separated string, required field***, user-defined manifest variables that can be used by the **model**.
    * **<manifestVariablesIndex>**, ***comma separated string, optional field***, user-defined manifest variables indices paired with **<manifestVariables>**. These indices indicate the position of manifest variables associated with multivariate normal distribution defined in the XML node **<Distributions>**. The indices

should be postive integer. If not provided, the code will use the positions of manifest variables listed in **<manifestVariables>** as the indices.

* **<method>**, ***string, required field***, the method that is used for the variables transformation. The currently available method is '**pca**'.

Assembler Objects These objects are either required or optional depending on the functionality of the AdaptiveSobol Sampler. The objects must be listed with a rigorous syntax that, except for the XML node tag, is common among all the objects. Each of these nodes must contain 2 attributes that are used to identify them within the simulation framework:

- **class**, ***required string attribute***, the main "class" of the listed object. For example, it can be **'Models'**, **'Functions'**, etc.
- **type**, ***required string attribute***, the object identifier or sub-type. For example, it can be **'ROM'**, **'External'**, etc.

The **AdaptiveSobol** approach requires or optionally accepts the following object types:

- **<ROM>**, ***string, required field***, the body of this XML node must contain the name of an appropriate ROM defined in the **<Models>** block (see Section 17.3).

- **<TargetEvaluation>**, ***string, required field***, represents the container where the system evaluations are stored. From a practical point of view, this XML node must contain the name of a data object defined in the **<DataObjects>** block (see Section 14). The Adaptive Sobol sampling accepts "DataObjects" of type "PointSet" only.

Example:

```
<Samplers>
  ...
  <AdaptiveSobol name="AS" verbosity='debug'>
    <Convergence>
      <relTolerance>1e-5</relTolerance>
      <maxRuns>150</maxRuns>
      <maxSobolOrder>3</maxSobolOrder>
      <progressParam>1</progressParam>
      <logFile>progress.txt</logFile>
      <subsetVerbosity>silent</subsetVerbosity>
    </Convergence>
    <variable name="x1">
```

```
      <distribution>UniDist</distribution>
    </variable>
    <variable name="x2">
      <distribution>UniDist</distribution>
    </variable>
    <ROM class = 'Models' type = 'ROM'>hdmrrom</ROM>
    <TargetEvaluation class = 'DataObjects' type =
       'PointSet'>solns</TargetEvaluation>
  </AdaptiveSobol>
  ...
</Samplers>
```

13 Optimizers

The optimizer is another important entity in the RAVEN framework. It performs the driving of a specific goal function over the model for value optimization. The difference between an optimizer and a sampler is that the former does not require sampling over a distribution, although certain specific optimizers may utilize stochastic approach to locate the optimality. The optimizers currently available in RAVEN can be categorized into the following class(es):

- **Gradient Based Optimizer** (see Section 13.1)

Before analyzing each optimizer in detail, it is important to mention that each type needs to be contained in the main XML node <Optimizers>, as reported below:

Example:

```
<Simulation>
  ...
  <Optimizers>
    ...
    <WhatEverOptimizer name='whatever'>
      ...
    </WhatEverOptimizer>
    ...
  </Optimizers>
  ...
</Simulation>
```

It should be noted that gradient-based optimizers will not function without including a <SolutionExport> HistorySet in the <MultiRun> step using the optimizer.

13.1 Gradient Based Optimizers

The Gradient Based Optimizer category collects all the strategies that perform the optimization based on gradient information, either directly provided or estimated by optimization strategy. In the RAVEN framework, currently implemented optimizer in this category are:

- **Simultaneous Perturbation Stochastic Approximation (SPSA)**
- **Finite Difference Gradient Optimizer Forward (FiniteDifferenceGradientOptimizer)**

From a practical point of view, these optimization strategies represent different ways to estimate the gradient based on information from previously performed model evaluation. In the following paragraphs, the input requirements and a small explanation of the different sampling methodologies are reported.

Note that in addition to the input variables and response variable as well as other model outputs, several other parameters are available to request for the output of a Gradient-Based Optimizer run. They include the following:

- **'varsUpdate'**, is the iteration number for each new optimal point;
- **'stepSize'**, is the step size used to go from the previous optimal point to the current step;
- **'gradient_var'**, where *var* is replaced by an input variable name, provides the gradient in the *var* direction followed to arrive at the current optimal point (evaluated at the previous optimal point);
- **'convergenceRel'**, the last-calculated relative convergence of the loss function value (see the description of the **<convergence>** node for more details);
- **'convergenceAbs'**, the last-calculated absolute convergence of the loss function value (see the description of the **<convergence>** node for more details);
- **'convergenceGrad'**, the last-calculated norm of the gradient used to arrive at the current point (see the description of the **<convergence>** node for more details).

Note that none of these additional parameters will be provided to the output DataObject by default; they must be specifically requested by listing them in the output space when defining the optimizing step's output data object. Also note that if any of these parameters are not available (for instance, on the first iteration), their output value will be set to -1, as this value is nonsensical for the step size and convergence values.

Example:

```
<Optimizers>
  ...
  <AnyGradientBasedOptimizer name="anyname">
    <initialization>
      <limit>300</limit>
    </initialization>
    <TargetEvaluation class="DataObjects"
       type="PointSet">TEdataObjectName</TargetEvaluation>
    <convergence>
      <iterationLimit>50</iterationLimit>
```

```xml
      <relativeThreshold>1e-3</relativeThreshold>
      <absoluteThreshold>1e-1</absoluteThreshold>
      <gradientThreshold>1e-5</gradientThreshold>
      <persistence>1</persistence>
    </convergence>
    <parameter>
      <numGradAvgIterations>3</numGradAvgIterations>
      <normalize>False</normalize>
    </parameter>
    <variable name="var1">
      <upperBound>100</upperBound>
      <lowerBound>-100</lowerBound>
      <initial>0</initial>
    </variable>
    <objectVar>c</objectVar>
  </SPSA>
  ...
</Optimizers>
..
<DataObjects>
  ...
  <PointSet name='optOut'>
    <Input>x,y</Input>
    <Output>z</Output>
  </PointSet>
  <HistorySet name='opt_export'>
    <Input>trajID</Input>
    <Output>
      x,y,z,varsUpdate,stepSize,
      gradient_x,gradient_y,
      convergenceAbs,convergenceRel,convergenceGrad
    </Output>
  </HistorySet>
  ...
</DataObjects>
```

13.1.1 Simultaneous Perturbation Stochastic Approximation (SPSA)

The **SPSA** optimization approach is one of the optimization strategies that are based on gradient estimation. The main idea is to simultaneously perturb all decision variables in order to estimate the gradient. Consequently a minimal number of two model evaluations are required in order to

approximate the gradient. The theory behind SPSA can be found in [3].

In addition to the algorithm in [3], current implementation of **SPSA** can also handles constrained optimization problem. This paragraph briefly describes how current implementation ensures the input satisfies the constraints. When when updating the variables (not perturbing), if constraint is violated, **SPSA** does the following in sequence:

- Try to find, through bisection method, the longest fraction of gradient vector so that the variable update satisfies the constraints;
- When such fraction cannot be found, then find a random vector orthogonal to gradient vector so that, by using this orthogonal vector as gradient, the variable update satisfies the constraints. Rotate the orthogonal vector towards the gradient, through bisection methods, until constraints can no longer be satisfied;
- If all above cannot return a constraint satisfying variable update, then do not update the variables and the **SPSA** will terminate.

It is important to notice that the gradient and the feature space is always normalized. This means that the gradient is going to be normalized with respect to its norm (versor of the gradient); hence, the optimization advancement is not going to be influenced by the magnitude of the gradient, but just on its "direction" information content. All the following parameters that can be optionally be inputted should be calibrated with this information in mind.

The specifications of this optimizer must be defined within a **<SPSA>** XML block. This XML node accepts the following attributes:

- **name**, ***required string attribute***, user-defined name of this optimizer. **Note:** As for the other objects, this is the name that can be used to refer to this specific entity from other input blocks (xml);

In the **<SPSA>** input block, the user needs to specify the objective variable to be optimized, the decision variables, the DataObject storing previously performed model evaluation, as well as convergence criteria. In addition, the settings for this optimization can be specified in the **<initialization>** and **<parameter>** XML blocks:

- **<initialization>**, ***XML node, optional parameter***. In this xml-node,the following xml sub-nodes can be specified:
  - **<limit>**, ***integer,optional field***, number of samples to be generated, which is same as the number of model evaluation.
    Default: 2000.
  - **<initialSeed>**, ***integer, optional field***, initial seeding of random number generator for stochastic perturbations;

- **<type>**, ***string (case insensitive), optional field***, specifies whether this optimizer performs maximization or minimization. Available options are `'max'` and `'min'`.
  Default: Min;
- **<thresholdTrajRemoval>**, ***float, optional field***, this will be used to determine the convergence of different optimization trajectories on each other when multiple trajectories is handled by **<SPSA>**. When one trajectory comes within tolerance of a point on another trajectory, the first will be removed in interest of the second. Note that this value is calculated as Euclidean distance in a normalized 0 to 1 cubic domain, not the original input space domain.
  Default: 0.05

- **<TargetEvaluation>**, ***XML node, required parameter***, represents the container where the model evaluations are stored. From a practical point of view, this XML node must contain the name of a data object defined in the **<DataObjects>** block (see Section 14). The object here specified must be input as **<Output>** in the Steps that employ this optimization strategy. The **<SPSA>** optimizer accepts "DataObjects" of type "PointSet" only;
- **<objectVar>**, ***XML node, required parameter***. The objective variable to be optimized. This variable must be output of the DataObject specified in **<TargetEvaluation>**.
- **<Sampler>**, ***XML node, optional parameter***, represents a Sampler (Forward) that can be used to initialize the starting points for the trajectories of some of the variables. From a practical point of view, this XML node must contain the name of a Sampler (Forward) defined in the **<Samplers>** block (see Section 12.1). The Sampler will be used to initialize the trajectories' initial points for some of the variables. For example, if the Sampler here specified "samples" only 2 variables over 5, the **<initial>** XML node (see below) is required only for the remaining 3 variables.
- **<Function>**, ***XML node, optional parameter***, indicates the external function where the constraints are stored. From a practical point of view, this XML node must contain the name of a function defined in the **<Functions>** block (see Section 18). This external function must contain a method called "constrain", which returns 1 for inputs satisfying the constraints and 0 otherwise.
- **<Preconditioner>**, ***XML node, optional parameter***, provides a model that can be used as a preconditioner in Multilevel optimization calculations. Only affects optimizers with a **<multilevel>** node. As many preconditioners as desired can be added to the optimizer, each defined with a **<Preconditioner>** node. From a practical point of view, this XML node must contain the name of an **<ExternalModel>** defined in the **<Models>** block (see Section 17.4). In multilevel optimization, the preconditioner is attached to a particular subspace. Whenever subspaces that are "higher" (early in **<sequence>**) are perturbed, before moving to a lower subspace, the preconditioner will be called to provide a new value for each variable in the lower subspace. For example, if an input space is divided into one

subspace **'subx'** with the input variable x and another subspace **'suby'** with input variable y, and if the sequence is specified as **'subx,suby'**, and a preconditioner is attached to subspace **'suby'**, then when a step is taken for subspace **'subx'**, the preconditioner will provide a new value for y before starting a convergence search for y.

- **<multilevel>**, ***XML node, optional node***, engages the optimizer in *multilevel* mode. When in multilevel mode, the input space is divided into multiple subspaces. The subspaces are then aligned in a sequence, and optimizing follows the following procedure:

  1. Hold all variable values in all subspaces constant EXCEPT the last subspace in the sequence.
  2. Converge the optimizer considering only input variables in the last listed subspace.
  3. Hold all variables in the last subspace constant, and take a single optimizing step in the second-to-last subspace.
  4. If the second to last subspace is converged, go up one more subspace and take a step, then repeat the process thus far.
  5. If the second to last subspace is not converged, go back to the last subspace and converge it again.
  6. Et cetera.

  Once the outermost subspace is converged, the entire space is considered converged. Note that multilevel optimization is not in general better than not using it. Multilevel works especially well when some variables in the input space are connected and have relatively difficult-to-converge optimization, while other variables in the input space are easily converged. In this case, the difficult-to-converge variables should make up the last subspace in the sequence, while the easily-converging variables should make up the outer subspace. RAVEN places no limit on the number of subspaces that are defined, but each variable should only exist in a single subspace. The **<multilevel>** node requires the definition of subspaces and the sequence as follows:

  - **<sequence>**, ***comma-separated string, required parameter***, lists the order in which subspaces should be converged. Each subspace is listed as identified by its **name** parameter in the **<subspace>** definition. Note that the first subspace listed will be the slowest to converge and converge only once, and the last subspace listed will be converged frequently and quickly.
  - **<subspace>**, ***comma-separated string, required parameter***, lists the variables included in this subspace. This node additionally has the following attributes:
    * **name**, ***string, require parameter***, provides the identifier that RAVEN will use for this subspace group, both in the **<sequence>** node as well as in log prints.
    * **precond**, ***model name, optional parameter***, provides the option to attach a preconditioner to this subset, chosen from the **<Preconditioner>** nodes defined

within the optimzer, and identified by the text of those nodes. See the documentation for the preconditioner node above for details on how they affect the calculation flow.

* **holdOutputSpace**, ***parameter names, optional comma separated parameter***, provides the option to identify some output parameters that need to be kept on hold at this subspace optimization level. This capability is currently implemented for the *EnsembleModel* only. In other words, all the models that have, in their output spaces, the parameter here specified will not be re-run for the iteration i but the solution at iteration $i-1$ will be used.

- **<variable>**, ***XML node, required parameter*** will specify one attribute:
  - **name**, ***required string attribute***, user-defined name of this variable.

  The variable specified here must be input of the DataObject specified in **<TargetEvaluation>**. This **<variable>** recognizes the following child nodes:
  - **<upperBound>**, ***float, required field***, the upper bound of this variable;
  - **<lowerBound>**, ***float, required field***, the lower bound of this variable;
  - **<initial>**, ***comma separated strings, optional field***, the initial value(s) for this variable. If there are more than one initial values specified for a variable, then all the variables need to have the same number of initial values. In this case, **<SPSA>** optimizer will maintain multiple trajectories to fully utilize potential parallel computing capability. Every input variable must have an initial value specified either through this node, or through a preconditioner in multilevel optimization or through a linked Sampler (see above).

<constant>, ***XML node, optional parameter*** the user is able to input variables that need to be kept constant. For doing this, as many **<constant>** nodes as needed can be input, where the body of the node contains the constant value that is going to be injected as an additional variable

- **<convergence>**, ***XML node, optional parameter*** will specify parameters associated with optimization convergence. This node accepts the following sub-nodes:
  - **<iterationLimit>**, ***integer, optional field***, user-defined maximum number of optimization iterations.
    Default: 650.
  - **<persistence>**, ***integer, optional field***, number of consecutive successful convergences required before completing calculation (per trajectory). Any value less than 1 will be treated as 1. Float values are rounded down to the nearest integer.
    Default: 1.

- <relativeThreshold>, ***float, optional field***, specifies the convergence criteria to determine the optimality in a "relative" sense: when the relative change of the objective variable in two successive model evaluations is smaller than this specified threshold, the <SPSA> optimizer is in convergence and terminates the simulation.
  Default: 1e-3
- <absoluteThreshold>, ***float, optional field***, specifies the convergence criteria to determine the optimality, in an "absolute" sense: when the absolute change of objective variable in two successive model evaluations is smaller than this specified threshold, the <SPSA> optimizer is in convergence and terminates the simulation.
  Default: 0.0
- <gradientThreshold>, ***float, optional field***, specifies the convergence criteria to determine the optimality, as function of the L2 norm of the gradient (useful for unconstrained problems): when the L2 norm od the gradient falls below this threshold, the <SPSA> optimizer is in convergence and terminates the simulation.
  Default: 1e-3
- <minStepSize>, ***float, optional field***, specifies the minimum allowable step size in the normalized input space, ranging from 0 (no movement) to 1 (spans any dimension).

  Default: 1e-9
- <gainGrowthSize>, ***float, optional field***, specifies the rate at which the step size should grow when it does grow, for instance when multiple steps are in the same direction. Increasing this will increase the likelihood that an optimization path travels quickly across the domain along a consistent gradient.
  Default: 2
- <gainShrinkSize>, ***float, optional field***, specifies the rate at which the step size should shrink when it does shrink, for instance when switching directions on successive steps. Increasing this will slow convergence, but decrease the likelihood of achieving false convergence due to small step sizes.
  Default: same value as grainGrowthSize

• <parameter>, ***XML node, optional parameter*** will accepts the following sub-nodes:

- <numGradAvgIterations>, ***integer, optional field*** is the number of iterations for gradient estimation. When this parameter is > 1, multiple gradient evaluations are going to be performed. Since the main goal for this parameter is to get a better gradient estimation (performing a denoising), the current point x_k is evaluated multiple times in order to be able to converge in average.
  Default: 1
- <stochasticDistribution>, ***string, optional field*** determines the process used to find gradient evaluation perturbation points as part of SPSA. Choice include the following:

    * **'Hypersphere'**, which chooses from all possible directions with equal probability,
    * **'Bernoulli'**, which limits directions closely to the diagonal directions (corners of a hypercube).

  Default: Hypersphere
- Optimizer Gradient Evaluation parameters:
    * **<gamma>**, ***float, optional field*** Inverse exponent for gradient evaluation distance. Increasing this parameter will greatly decrease the distance between points sampled in evaluating the gradient [3]. A practical suggestion for γ is 0.101 (paired with an α value of 0.602); however, the asymptotic limit is $\gamma = 1/6$ ($\alpha = 1$).
      Default: 0.101
    * **<c>**, ***float, optional field*** Step size coefficient. This term determines the nominal step size, and increasing it will directly increase the distance between points sampled in evaluating the gradient [3]. It is suggested this parameter be approximately equal to the standard deviation of the measurement noise in the response for stochastic responses. For regular responses, it can be a small arbitrary value.
      Default: 0.005
- Optimizer Step Size parameters:
    * **<a>**, ***float, optional field*** Nominal optimizer step size parameter. Increasing this parameter will directly increase the distance traversed in each optimizer step [3]. In contrast to A, this parameter will be unchanged by increasing iterations, and so will be more impacting as the optimization algorithm iterates.
      Default: 0.16
    * **<alpha>**, ***float, optional field*** Inverse exponent for optimizer step size. Increasing this parameter will greatly decrease the distance traversed in each optimizer step [3]. Values less than 1 for α usually yield better performance by keeping a large step size. See the description of γ above for some suggested values.
      Default: 0.602
    * **<A>**, ***float, optional field*** Nominal step damping stability parameter. Increasing this parameter will directly decrease the distance traversed in each optimizer step [3]. This parameter will have greater affect in reducing step size early in the calculation, and reduced affect as iterations increase.
      Default: ***<limit>*** *divided by 10*
- **<innerBisectionThreshold>**, ***float, optional field*** a parameter specifying the convergence threshold of the bisection method used in constraint handling (See above). This parameter shall be in the open inverval $(0, 1)$.
  Default: 0.01
- **<innerLoopLimit>**, ***integer, optional field*** a parameter specifying the number of orthogonal vectors to try when handling the constraints (See above).
  Default: 1000

Example:

```
<Optimizers>
  ...
  <SPSA name="SPSAname">
    <initialization>
      <limit>300</limit>
      <type>min</type>
      <initialSeed>30</initialSeed>
    </initialization>
    <TargetEvaluation class="DataObjects"
       type="PointSet">dataObjectName</TargetEvaluation>
    <convergence>
      <iterationLimit>50</iterationLimit>
      <relativeThreshold>1e-3</relativeThreshold>
      <absoluteThreshold>1e-1</absoluteThreshold>
      <gradientThreshold>1e-5</gradientThreshold>
      <persistence>1</persistence>
    </convergence>
    <parameter>
      <numGradAvgIterations>3</numGradAvgIterations>
    </parameter>
    <variable name="var1">
      <upperBound>100</upperBound>
      <lowerBound>-100</lowerBound>
      <initial>0</initial>
    </variable>
    <objectVar>c</objectVar>
  </SPSA>
  ...
</Optimizers>
```

13.1.2 Finite Difference Gradient Optimizer (FiniteDifferenceGradientOptimizer)

The **FiniteDifferenceGradientOptimizer** optimization approach is the simplest Gradient based approach since it is based on the first order evaluation of the Gradient. A minimal number of $nvariable$ model evaluations are required in order to get a first order approximation of the gradient.

Current implementation of **FiniteDifferenceGradientOptimizer** can also handles constrained optimization problem. This paragraph briefly describes how current implementation ensures the input satisfies the constraints. When when updating the variables (not perturbing), if constraint is violated, **FiniteDifferenceGradientOptimizer** does the following in sequence:

- Try to find, through bisection method, the longest fraction of gradient vector so that the variable update satisfies the constraints;
- When such fraction cannot be found, then find a random vector orthogonal to gradient vector so that, by using this orthogonal vector as gradient, the variable update satisfies the constraints. Rotate the orthogonal vector towards the gradient, through bisection methods, until constraints can no longer be satisfied;
- If all above cannot return a constraint satisfying variable update, then do not update the variables and the **FiniteDifferenceGradientOptimizer** will terminate.

It is important to notice that the gradient and the feature space is always normalized. This means that the gradient is going to be normalized with respect to its norm (versor of the gradient); hence, the optimization advancement is not going to be influenced by the magnitude of the gradient, but just on its "direction" information content. All the following parameters that can be optionally be inputted should be calibrated with this information in mind.

The specifications of this optimizer must be defined within a **<FiniteDifferenceGradientOptimize** XML block. This XML node accepts the following attributes:

- **name**, ***required string attribute***, user-defined name of this optimizer. **Note:** As for the other objects, this is the name that can be used to refer to this specific entity from other input blocks (xml);

In the **<FiniteDifferenceGradientOptimizer>** input block, the user needs to specify the objective variable to be optimized, the decision variables, the DataObject storing previously performed model evaluation, as well as convergence criteria. In addition, the settings for this optimization can be specified in the **<initialization>** and **<parameter>** XML blocks:

- **<initialization>**, ***XML node, optional parameter***. In this xml-node,the following xml sub-nodes can be specified:
  - **<limit>**, ***integer,optional field***, number of samples to be generated, which is same as the number of model evaluation.
    Default: 2000.
  - **<initialSeed>**, ***integer, optional field***, initial seeding of random number generator for stochastic perturbations;
  - **<type>**, ***string (case insensitive), optional field***, specifies whether this optimizer performs maximization or minimization. Available options are **'max'** and **'min'**.
    Default: Min;
  - **<thresholdTrajRemoval>**, ***float, optional field***, this will be used to determine the convergence of different optimization trajectories on each other when multiple trajectories is handled by **<FiniteDifferenceGradientOptimizer>**. When one

trajectory comes within tolerance of a point on another trajectory, the first will be removed in interest of the second. Note that this value is calculated as Euclidean distance in a normalized 0 to 1 cubic domain, not the original input space domain.
Default: 0.05

- **<TargetEvaluation>**, ***XML node, required parameter***, represents the container where the model evaluations are stored. From a practical point of view, this XML node must contain the name of a data object defined in the **<DataObjects>** block (see Section 14). The object here specified must be input as **<Output>** in the Steps that employ this optimization strategy. The **<FiniteDifferenceGradientOptimizer>** optimizer accepts "DataObjects" of type "PointSet" only;

- **<objectVar>**, ***XML node, required parameter***. The objective variable to be optimized. This variable must be output of the DataObject specified in **<TargetEvaluation>**.

- **<Sampler>**, ***XML node, optional parameter***, represents a Sampler (Forward) that can be used to initialize the starting points for the trajectories of some of the variables. From a practical point of view, this XML node must contain the name of a Sampler (Forward) defined in the **<Samplers>** block (see Section 12.1). The Sampler will be used to initialize the trajectories' initial points for some of the variables. For example, if the Sampler here specified "samples" only 2 variables over 5, the **<initial>** XML node (see below) is required only for the remaining 3 variables.

- **<Function>**, ***XML node, optional parameter***, indicates the external function where the constraints are stored. From a practical point of view, this XML node must contain the name of a function defined in the **<Functions>** block (see Section 18). This external function must contain a method called "constrain", which returns 1 for inputs satisfying the constraints and 0 otherwise.

- **<Preconditioner>**, ***XML node, optional parameter***, provides a model that can be used as a preconditioner in Multilevel optimization calculations. Only affects optimizers with a **<multilevel>** node. As many preconditioners as desired can be added to the optimizer, each defined with a **<Preconditioner>** node. From a practical point of view, this XML node must contain the name of an **<ExternalModel>** defined in the **<Models>** block (see Section 17.4). In multilevel optimization, the preconditioner is attached to a particular subspace. Whenever subspaces that are "higher" (early in **<sequence>**) are perturbed, before moving to a lower subspace, the preconditioner will be called to provide a new value for each variable in the lower subspace. For example, if an input space is divided into one subspace **'subx'** with the input variable x and another subspace **'suby'** with input variable y, and if the sequence is specified as **'subx,suby'**, and a preconditioner is attached to subspace **'suby'**, then when a step is taken for subspace **'subx'**, the preconditioner will provide a new value for y before starting a convergence search for y.

- **<multilevel>**, ***XML node, optional node***, engages the optimizer in *multilevel* mode. When in multilevel mode, the input space is divided into multiple subspaces. The subspaces are then aligned in a sequence, and optimizing follows the following procedure:

  1. Hold all variable values in all subspaces constant EXCEPT the last subspace in the sequence.
  2. Converge the optimizer considering only input variables in the last listed subspace.
  3. Hold all variables in the last subspace constant, and take a single optimizing step in the second-to-last subspace.
  4. If the second to last subspace is converged, go up one more subspace and take a step, then repeat the process thus far.
  5. If the second to last subspace is not converged, go back to the last subspace and converge it again.
  6. Et cetera.

  Once the outermost subspace is converged, the entire space is considered converged. Note that multilevel optimization is not in general better than not using it. Multilevel works especially well when some variables in the input space are connected and have relatively difficult-to-converge optimization, while other variables in the input space are easily converged. In this case, the difficult-to-converge variables should make up the last subspace in the sequence, while the easily-converging variables should make up the outer subspace. RAVEN places no limit on the number of subspaces that are defined, but each variable should only exist in a single subspace. The **<multilevel>** node requires the definition of subspaces and the sequence as follows:

  – **<sequence>**, ***comma-separated string, required parameter***, lists the order in which subspaces should be converged. Each subspace is listed as identified by its **name** parameter in the **<subspace>** definition. Note that the first subspace listed will be the slowest to converge and converge only once, and the last subspace listed will be converged frequently and quickly.

  – **<subspace>**, ***comma-separated string, required parameter***, lists the variables included in this subspace. This node additionally has the following attributes:

    * **name**, ***string, require parameter***, provides the identifier that RAVEN will use for this subspace group, both in the **<sequence>** node as well as in log prints.
    * **precond**, ***model name, optional parameter***, provides the option to attach a preconditioner to this subset, chosen from the **<Preconditioner>** nodes defined within the optimzer, and identified by the text of those nodes. See the documentation for the preconditioner node above for details on how they affect the calculation flow.
    * **holdOutputSpace**, ***parameter names, optional comma separated parameter***, provides the option to identify some output parameters that need to be kept on hold

at this subspace optimization level. This capability is currently implemented for the *EnsembleModel* only. In other words, all the models that have, in their output spaces, the parameter here specified will not be re-run for the iteration i but the solution at iteration $i - 1$ will be used.

- **<variable>**, ***XML node, required parameter*** will specify one attribute:
  - **name**, ***required string attribute***, user-defined name of this variable.

  The variable specified here must be input of the DataObject specified in **<TargetEvaluation>**. This **<variable>** recognizes the following child nodes:
  - **<upperBound>**, ***float, required field***, the upper bound of this variable;
  - **<lowerBound>**, ***float, required field***, the lower bound of this variable;
  - **<initial>**, ***comma separated strings, optional field***, the initial value(s) for this variable. If there are more than one initial values specified for a variable, then all the variables need to have the same number of initial values. In this case, **<FiniteDifferenceGradient** optimizer will maintain multiple trajectories to fully utilize potential parallel computing capability. Every input variable must have an initial value specified either through this node, or through a preconditioner in multilevel optimization or through a linked Sampler (see above).

<constant>, ***XML node, optional parameter*** the user is able to input variables that need to be kept constant. For doing this, as many **<constant>** nodes as needed can be input, where the body of the node contains the constant value that is going to be injected as an additional variable

- **<convergence>**, ***XML node, optional parameter*** will specify parameters associated with optimization convergence. This node accepts the following sub-nodes:
  - **<iterationLimit>**, ***integer, optional field***, user-defined maximum number of optimization iterations.
    Default: 650.
  - **<persistence>**, ***integer, optional field***, number of consecutive successful convergences required before completing calculation (per trajectory). Any value less than 1 will be treated as 1. Float values are rounded down to the nearest integer.
    Default: 1.
  - **<relativeThreshold>**, ***float, optional field***, specifies the convergence criteria to determine the optimality in a "relative" sense: when the relative change of the objective variable in two successive model evaluations is smaller than this specified threshold, the **<FiniteDifferenceGradientOptimizer>** optimizer is in convergence and terminates the simulation.
    Default: 1e-3

- **<absoluteThreshold>**, ***float, optional field***, specifies the convergence criteria to determine the optimality, in an "absolute" sense: when the absolute change of objective variable in two successive model evaluations is smaller than this specified threshold, the **<FiniteDifferenceGradientOptimizer>** optimizer is in convergence and terminates the simulation.
  Default: 0.0
- **<gradientThreshold>**, ***float, optional field***, specifies the convergence criteria to determine the optimality, as function of the L2 norm of the gradient (useful for unconstrained problems): when the L2 norm od the gradient falls below this threshold, the **<FiniteDifferenceGradientOptimizer>** optimizer is in convergence and terminates the simulation.
  Default: 1e-3
- **<minStepSize>**, ***float, optional field***, specifies the minimum allowable step size in the normalized input space, ranging from 0 (no movement) to 1 (spans any dimension).

  Default: 1e-9
- **<gainGrowthSize>**, ***float, optional field***, specifies the rate at which the step size should grow when it does grow, for instance when multiple steps are in the same direction. Increasing this will increase the likelihood that an optimization path travels quickly across the domain along a consistent gradient.
  Default: 2
- **<gainShrinkSize>**, ***float, optional field***, specifies the rate at which the step size should shrink when it does shrink, for instance when switching directions on successive steps. Increasing this will slow convergence, but decrease the likelihood of achieving false convergence due to small step sizes.
  Default: same value as grainGrowthSize

- **<parameter>**, ***XML node, optional parameter*** will accepts the following sub-nodes:
  - **<numGradAvgIterations>**, ***integer, optional field*** is the number of iterations for gradient estimation. When this parameter is > 1, multiple gradient evaluations are going to be performed. Since the main goal for this parameter is to get a better gradient estimation (performing a denoising), the current point x_k is evaluated multiple times in order to be able to converge in average.
    Default: 1
  - Optimizer Gradient Evaluation parameters:
    * **<gamma>**, ***float, optional field*** Inverse exponent for gradient evaluation distance. Increasing this parameter will greatly decrease the distance between points sampled in evaluating the gradient [3]. A practical suggestion for γ is 0.101 (paired with an α value of 0.602); however, the asymptotic limit is $\gamma = 1/6$ ($\alpha = 1$).
      Default: 0.101

* **<c>**, ***float, optional field*** Step size coefficient. This term determines the nominal step size, and increasing it will directly increase the distance between points sampled in evaluating the gradient. It is suggested this parameter be approximately equal to the standard deviation of the measurement noise in the response for stochastic responses. For regular responses, it can be a small arbitrary value.
  Default: 0.005

– Optimizer Step Size parameters:

* **<a>**, ***float, optional field*** Nominal optimizer step size parameter. Increasing this parameter will directly increase the distance traversed in each optimizer step [3]. In contrast to A, this parameter will be unchanged by increasing iterations, and so will be more impacting as the optimization algorithm iterates.
  Default: 0.16
* **<alpha>**, ***float, optional field*** Inverse exponent for optimizer step size. Increasing this parameter will greatly decrease the distance traversed in each optimizer step [3]. Values less than 1 for α usually yield better performance by keeping a large step size. See the description of γ above for some suggested values.
  Default: 0.602
* **<A>**, ***float, optional field*** Nominal step damping stability parameter. Increasing this parameter will directly decrease the distance traversed in each optimizer step [3]. This parameter will have greater affect in reducing step size early in the calculation, and reduced affect as iterations increase.
  *Default: **<limit>** divided by 10*

– **<innerBisectionThreshold>**, ***float, optional field*** a parameter specifying the convergence threshold of the bisection method used in constraint handling (See above). This parameter shall be in the open inverval $(0, 1)$.
Default: 0.01

– **<innerLoopLimit>**, ***integer, optional field*** a parameter specifying the number of orthogonal vectors to try when handling the constraints (See above).
Default: 1000

Example:

```
<Optimizers>
  ...
  <FiniteDifferenceGradientOptimizer name="SPSAname">
    <initialization>
      <limit>300</limit>
      <type>min</type>
      <initialSeed>30</initialSeed>
    </initialization>
    <TargetEvaluation class="DataObjects"
       type="PointSet">dataObjectName</TargetEvaluation>
```

```xml
    <convergence>
      <iterationLimit>50</iterationLimit>
      <relativeThreshold>1e-3</relativeThreshold>
      <absoluteThreshold>1e-1</absoluteThreshold>
      <gradientThreshold>1e-5</gradientThreshold>
      <persistence>1</persistence>
    </convergence>
    <parameter>
      <numGradAvgIterations>3</numGradAvgIterations>
    </parameter>
    <variable name="var1">
      <upperBound>100</upperBound>
      <lowerBound>-100</lowerBound>
      <initial>0</initial>
    </variable>
    <objectVar>c</objectVar>
  </FiniteDifferenceGradientOptimizer>
  ...
</Optimizers>
```

14 DataObjects

As seen in the previous chapters, different entities in the RAVEN code interact with each other in order to create, ideally, an infinite number of different calculation flows. These interactions are made possible through a data handling system that each entity understands. This system, neglecting the grammar imprecision, is called the "DataObjects" system.

The **<DataObjects>** tag is a container of data objects of various types that can be constructed during the execution of a particular calculation flow. These data objects can be used as input or output for a particular **Model** (see Roles' meaning in section 17), etc. Currently, RAVEN supports 4 different data types, each with a particular conceptual meaning. These data types are instantiated as sub-nodes in the **<DataObjects>** block of an input file:

- **<PointSet>** is a collection of individual objects, each describing the state of the system at a certain point (e.g. in time). It can be considered a mapping between multiple sets of parameters in the input space and the resulting sets of outcomes in the output space at a particular point (e.g. in time).
- **<HistorySet>** is a collection of individual objects each describing the temporal evolution of the state of the system within a certain input domain. It can be considered a mapping between multiple sets of parameters in the input space and the resulting sets of temporal evolutions in the output space.

As noted above, each data object represents a mapping between a set of parameters and the resulting outcomes. The data objects are defined within the main XML block called **<DataObjects>**:

```
<Simulation>
   ...
  <DataObjects>
    <PointSet name='***'>...</PointSet>
    <HistorySet name='***'>...</HistorySet>
  </DataObjects>
   ...
</Simulation>
```

Independent of the type of data, the respective XML node has the following available attributes:

- name, ***required string attribute***, is a user-defined identifier for this data object. **Note:** As with other objects, this name can be used to refer to this specific entity from other input blocks in the XML.

- **hierarchical**, ***optional boolean attribute***, if True this data object will be constructed, if possible, in a hierarchical fashion.
  Default: False

In each XML node (e.g. **<PointSet>** or **<HistorySet>**), the user needs to specify the following sub-nodes:

- **<Input>**, ***comma separated string, required field*** lists the input parameters to which this data is connected.
- **<Output>**, ***comma separated string, required field*** lists the output parameters to which this data is connected.

In addition to the XML nodes **<Input>** and **<Output>** explained above, the user can optionally specify a XML node named **<options>**. The **<options>** node can contain the following optional XML sub-nodes:

- **<inputRow>**, ***integer, optional field***, used to specify the row (in the CSV file or HDF5 table) from which the input space needs to be retrieved (e.g. the time-step);
- **<outputRow>**, ***integer, optional field***, used to specify the row (in the CSV file or HDF5 table) from which the output space needs to be retrieved (e.g. the time-step). If this node is inputted, the nodes **<operator>** and **<outputPivotValue>** can not be inputted (mutually exclusive).
  Note: This XML node is available for DataObjects of type **<PointSet>** only;
- **<operator>**, ***string, optional field***, is aimed to perform simple operations on the data to be stored. The 3 options currently available are:
  - **'max'**
  - **'min'**
  - **'average'**

  If this node is inputted, the nodes **<outputRow>** and **<outputPivotValue>** can not be inputted (mutually exclusive).
  Note: This XML node is available for DataObjects of type **<PointSet>** only; It needs to be noticed that if the optional nodes in the block **<options>** are not inputted, the following default are applied:
  - the Input space is retrieved from the first row in the CSVs files or HDF5 tables (if the parameters specified are not among the variables sampled by RAVEN);
  - the output space defaults are as follows:

* if <PointSet>, the output space is retrieved from the last row in the CSVs files or HDF5 tables;
* if <HistorySet>, the output space is represented by all the rows found in the CSVs or HDF5 tables.

```
<DataObjects>
  <PointSet name='outTPS1'>
    <options>
     <inputRow>1</inputRow>
     <outputRow>-1</outputRow>
    </options>
    <Input>pipe_Area,pipe_Dh,Dummy1</Input>
    <Output>pipe_Hw,pipe_Tw,time</Output>
  </PointSet>
  <HistorySet name='stories1'>
      <options>
          <inputRow>1</inputRow>
          <outputRow>-1</outputRow>
      </options>
    <Input>pipe_Area,pipe_Dh</Input>
    <Output>pipe_Hw,pipe_Tw,time</Output>
  </HistorySet>
</DataObjects>
```

15 Databases

The RAVEN framework provides the capability to store and retrieve data to/from an external database. Currently RAVEN has support for only a database type called **HDF5**. This database, depending on the data format it is receiving, will organize itself in a "parallel" or "hierarchical" fashion. The user can create as many database objects as needed. The Database objects are defined within the main XML block called **<Databases>**:

```
<Simulation>
  ...
  <Databases>
    ...
    <HDF5 name="aDatabaseName1" readMode="overwrite"/>
    <HDF5 name="aDatabaseName2" readMode="overwrite"/>
    ...
  </Databases>
  ...
</Simulation>
```

The specifications of each Database of type HDF5 needs to be defined within the XML block **<HDF5>**, that recognizes the following attributes:

- `name`, ***required string attribute***, a user-defined identifier of this object. **Note:** As with other objects, this is name can be used to reference this specific entity from other input blocks in the XML.
- `readMode`, ***required string attribute***, defines whether an existing database should be read when loaded (`'read'`) or overwritten (`'overwrite'`). **Note:** if in `'read'` mode and the database is not found, RAVEN will read in the data as empty and raise a warning, NOT an error.
- `directory`, ***optional string attribute***, this attribute can be used to specify a particular directory path where the database will be created or read from. If an absolute path is given, RAVEN will respect it; otherwise, the path will be assumed to be relative to the **<WorkingDir>** from the **<RunInfo>** block. RAVEN recognizes path expansion tools such as tildes (*user dir*), single dots (*current dir*), and double dots (*parent dir*).
  Default: workingDir/DatabaseStorage. The **<workingDir>** is the one defined within the **<RunInfo>** XML block (see Section 8).
- `filename`, ***optional string attribute***, specifies the filename of the HDF5 that will be created in the `directory`. **Note:** When this attribute is not specified, the newer database filename will be named `name`.h5, where *name* corresponds to the `name` attribute of this object.
  Default: None
- `compression`, ***optional string attribute***, compression algorithm to be used. Available are:

– **'gzip'**, best where portability is required. Good compression, moderate speed.
– **'lzf'**, Low to moderate compression, very fast.

Default: None

In addition, the **<HDF5>** recognizes the following subnodes:

- **<variables>**, ***optional, comma-separated string***, allows only a pre-specified set of variables to be included in the HDF5 when it is written to. If this node is not included, by default the HDF5 will include ALL of the input/output variables as a result of the step it is part of. If included, only the comma-separated variable names will be included if found.

  Note: RAVEN will not error if one of the requested variables is not found; instead, it will silently pass. It is recommended that a small trial run is performed, loading the HDF5 back into a data object, to check that the correct variables are saved to the HDF5 before performing large-scale calculations.

Example:

```
<Databases>
  <HDF5 name="aDatabaseName1" directory=''path_to_a_dir''
     compression=''lzf'' readMode='overwrite'/>
  <HDF5 name="aDatabaseName2" filename=''aDatabaseName2.h5''
     readMode='read'/>
</Databases>
```

16 OutStream system

The PRA and UQ framework provides the capabilities to visualize and dump out the data that are generated, imported (from a system code) and post-processed during the analysis. These capabilities are contained in the "OutStream" system. Actually, two different OutStream types are available:

- **Print**, module that lets the user dump the data contained in the internal objects;
- **Plot**, module, based on MatPlotLib [4], aimed to provide advanced plotting capabilities.

Both the types listed above accept as "input" a *DataObjects* object type. This choice is due to the "*DataObjects*" system (see section 14) having the main advantage of ensuring a standardized approach for exchanging the data/meta-data among the different framework entities. Every module can project its outcomes into a *DataObjects* object. This provides the user with the capability to visualize/dump all the modules' results. Additionally, the **Print** system can accept a ROM and inquire some of its specialized methods. As already mentioned, the RAVEN framework input is based on the e**X**tensible **M**arkup **L**anguage (**XML**) format. Thus, in order to activate the "*OutStream*" system, the input needs to contain a block identified by the **<OutStreams>** tag (as shown below).

```
<OutStreams>
  <!-- "OutStream" objects that need to be created-->
</OutStreams>
```

In the "OutStreams" block an unlimited number of "Plot" and "Print" sub-blocks can be specified. The input specifications and the main capabilities for both types are reported in the following sections.

16.1 Printing system

The Printing system has been created in order to let the user dump the data, contained in the internal data objects (see Section 14), out at anytime during the calculation. Additionally, the user can inquire special methods of a **ROM** after training it, through a printing step. Currently, the only available output is a **C**omma **S**eparated **V**alue (**CSV**) file for **DataObjects**, and **XML** for **ROM** objects. This will facilitate the exchanging of results and provide the possibility to dump the solution of an analysis and "restart" another one constructing a data object from scratch, as well as access advanced features of particular reduced order models.

16.1.1 DataObjects Printing

The XML code, that is reported below, shows different ways to request a *Print* OutStream for **DataObjects**. The user needs to provide a name for each sub-block (XML attribute). These names are then used in the *Step* blocks to activate the Printing keywords at any time. The XML node has the following available attributes:

- `name`, ***required string attribute***, is a user-defined identifier for this data object. **Note:** As with other objects, this name can be used to refer to this specific entity from other input blocks in the XML.

As shown in the examples below, every `<Print>` block must contain, at least, the two required tags:

- `<type>`, the output file type (csv or xml). **Note:** Only **csv** is currently available for `<DataObjects>`
- `<source>`, the *Data* name (one of the *Data* items defined in the `<DataObjects>` block.

An optional tag `<filename>` can be used to specify the filename for the output. If this is not defined, then the default name will be the `name` identifier of the tag.

If only these two tags are provided (as in the "first-example" below), the output file will be filled with the whole content of the "d-name" *Data* object.

```
<OutStreams>
  <Print name='first-example'>
    <type>csv</type>
    <source>d-name</source>
  </Print>
  <Print name='second-example'>
    <type>csv</type>
    <source>d-name</source>
    <what>Output</what>
  </Print>
  <Print name='third-example'>
    <type>csv</type>
    <source>d-name</source>
    <what>Input</what>
  </Print>
  <Print name='fourth-example'>
    <type>csv</type>
    <source>d-name</source>
    <what>Input|var-name-in,Output|var-name-out</what>
```

```xml
  </Print>
  <Print name='fifth-example'>
    <type>csv</type>
    <source>d-name</source>
    <filename>example5</filename>
  </Print>
</OutStreams>
```

If just part of the **<source>** is important for a particular analysis, the additional XML tag **<what>** can be provided. In this block, the variables that need to be dumped must be specified, in comma separated format. The available options, for the **<what>** sub-block, are listed below:

- **Output**, the output space will be dumped out (see "second-example")
- **Input**, the input space will be dumped out (see "third-example")
- **Input—var-name-in/Output—var-name-out**, only the particular variables "var-name-in" and "var-name-out" will be reported in the output file (see "fourth-example")

Note all of the XML tags are case-sensitive but not their content.

16.1.2 ROM Printing

While all **ROM**s in RAVEN are designed to be used as surrogate models, some **ROM**s additionally offer information about the original model that isn't accessible through another means. For instance, **HDMRRom** objects can calculate sensitivity coefficients for subsets of the input domain. The XML code shown below demonstrates the methods to request these features from a **ROM**. The user needs to provide a **<name>** for each sub-block (XML attribute). These names are then used in the *Step* blocks to activate the Printing keywords at any time. As shown in the examples below, every **<Print>** block for ROMs must contain, at least, the three required tags

- **<type>**, the output file type (csv or xml). **Note:** Only **xml** is currently available for ROMs
- **<source>**, the *ROM* name (one of the **<ROM>** items defined in the **<Models>** block.
- **<what>**, the comma-separated list of desired metrics. The list of metrics available in each ROM is listed under that ROM type in Section 17.3. Alternatively, the keyword **'all'** can be provided to request all available metrics, if any.

Addtionally, when printing ROMs one optional node is available,

- **<target>**, the ROM target for which to inquire data

If the ROM is time-dependent, the printed properties will be collected by time step. ROM printing uses the same naming conventions as DataObjects printing. Examples:

```xml
<OutStreams>
  <Print name='first-ROM-example'>
    <type>xml</type>
    <source>mySobolRom</source>
    <what>all</what>
  </Print>
  <Print name='second-ROM-example'>
    <type>xml</type>
    <source>myGaussPolyRom</source>
    <what>mean,variance</what>
  </Print>
</OutStreams>
```

16.2 Plotting system

The Plotting system provides all the capabilities to visualize the analysis outcomes, in real-time or as a post-processing stage. The system is based on the Python library MatPlotLib [4]. MatPlotLib is a 2D/3D plotting library which produces publication quality figures in a variety of hardcopy formats and interactive environments across platforms. This external tool has been wrapped in the RAVEN framework, and is available to the user. Since it was unfeasible to support, in the source code, all the interfaces for all the available plot types, the RAVEN Plotting system directly provide a formatted input structure for 11 different plot types (2D/3D). The user may request a plot not present among the supported ones, since the RAVEN Plotting system has the capability to construct on the fly the interface for a Plot, based on XML instructions.

16.2.1 Plot input structure

In order to create a plot, the user needs to add, within the **<OutStreams>** block, a **<Plot>** sub-block. Similar to the **<Print>** OutStream, the user needs to specify a **name** as an attribute of the plot. This name will then be used to request the plot in the **<Steps>** block. In addition, the Plot block accepts the following attributes:

- **interactive**, ***optional bool attribute***, specifies if the Plot needs to be interactively created (real-time screen visualization).
  Default: False
- **overwrite**, ***optional bool attribute***, used when the plot needs to be dumped into picture file/s. This attribute determines whether the code needs to overwrite the image files every time a new plot (with the same name) is requested.
  Default: False

An optional tag **<filename>** can be used to specify the filename for the output. If this is not defined, then the default base name will be the **name** identifier of the tag prepended and appended with extra information that identifies the plot further.

As shown, in the XML input example below, the body of the Plot XML input contains two main sub-nodes:

- **<actions>**, where general control options for the figure layout are defined (see Section 16.2.1.1).
- **<plotSettings>**, where the actual plot options are provided.

These two main sub-block are discussed in the following paragraphs.

16.2.1.1 "Actions" input block

The input in the **<actions>** sub-node is common to all the Plot types, since, in it, the user specifies all the controls that need to be applied to the figure style. This block must be unique in the definition of the **<Plot>** main block. In the following list, all the predefined "actions" are reported:

- **<how>**, comma separated list of output types:
  - `screen`, show the figure on the screen in interactive mode
  - `pdf`, save the figure as a Portable Document Format file (PDF). **Note:** The pdf format does not support multiple layers that lay on the same pixel. If the user gets an error about this, he/she should move to another format.
  - `png`, save the figure as a Portable Network Graphics file (PNG)
  - `eps`, save the figure as an Encapsulated Postscript file (EPS)
  - `pgf`, save the figure as a LaTeX PGF Figure file (PGF)
  - `ps`, save the figure as a Postscript file (PS)
  - `gif`, save the figure as a Graphics Interchange Format (GIF)
  - `svg`, save the figure as a Scalable Vector Graphics file (SVG)
  - `jpeg`, save the figure as a jpeg file (JPEG)
  - `raw`, save the figure as a Raw RGBA bitmap file (RAW)
  - `bmp`, save the figure as a Windows bitmap file (BMP)
  - `tiff`, save the figure as a Tagged Image Format file (TIFF)
  - `svgz`, save the figure as a Scalable Vector Graphics file (SVGZ)
- **<title>**, as the name suggests, within this block the user can specify the title of the figure. In the body of this node, a few other tags are available:

- <text>, ***string type***, title of the figure
- <kwargs>, within this block the user can specify optional parameters with the following format:

```
<kwargs>
 <param1>value1</param1>
 <param2>value2</param2>
</kwargs>
```

  The kwargs block is able to convert whatever string into a python type (for example <param1>{'1stKey':45}</param1> will be converted into a dictionary, <param2>[56,67]</param2> into a list, etc.). For reference regarding the available kwargs, see "matplotlib.pyplot.title" method in [4].

- <labelFormat>, within this block the default scale formatting can be modified. In the body, a few other tags are available:
  - <axis>, ***string***, the axis where to apply the defined format, 'x,' 'y,' or 'both'.
    Default: 'both' **Note:** If this action will be used in a 3-D plot, the user can input 'z' as well and 'both' will apply this format to all three axis.
  - <style>, ***string***, the style of the number notation, 'sci' or 'scientific' for scientific, 'plain' for plain notation.
    Default: scientific
  - <scilimits>, ***tuple, (m, n), pair of integers***, if style is 'sci', scientific notation will be used for numbers outside the range 10^m to 10^n. Use (0,0) to include all numbers.
    Note: The value for this keyword, needs to be specified between brackets [for example, (5,6)].
    Default: (0,0)
  - <useOffset>, ***bool or double***, if True, the offset will be calculated as needed; if False, no offset will be used; if a numeric offset is specified, it will be used.
    Default: False
- <figureProperties>, within this block the user specifies how to customize the figure style/quality. Thus, through this "action" the user has got full control on the quality of the figure, its dimensions, etc. This control is performed by the following keywords:
  - <figsize>, ***tuple (optional)***, (width, hight), in inches.
  - <dpi>, ***integer***, dots per inch.
  - <facecolor>, ***string***, set the figure background color (please refer to "matplotlib.figure.Figure" in [4] for a list of all the colors available).
  - <edgecolor>, ***string***, the figure edge background color (please refer to "matplotlib.figure.Figure" in [4] for a list of all the colors available).

  - `<linewidth>`, ***float***, the width of lines drawn on the plot.
  - `<frameon>`, ***bool***, if False, suppress drawing the figure frame.

- `<range>`, the range "action" specifies the ranges of all the axis. All the keywords in the body of this block are optional:

  - `<ymin>`, ***double (optional)***, lower boundary for the y axis.
  - `<ymax>`, ***double (optional)***, upper boundary for the y axis.
  - `<xmin>`, ***double (optional)***, lower boundary for the x axis.
  - `<xmax>`, ***double (optional)***, upper boundary for the x axis.
  - `<zmin>`, ***double (optional)***, lower boundary for the z axis. **Note:** This keyword is effective in 3-D plots only.
  - `<zmax>`, ***double (optional)***, upper boundary for the z axis. **Note:** This keyword is effective in 3-D plots only.

- `<camera>`, the camera item is available in 3-D plots only. Through this "action," it is possible to orientate the plot as one wishes. The controls are:

  - `<elevation>`, ***double (optional)***, stores the elevation angle in the z plane.
  - `<azimuth>`, ***double (optional)***, stores the azimuth angle in the x,y plane.

- `<scale>`, the scale block allows the specification of the axis scales:

  - `<xscale>`, ***string (optional)***, scale of the x axis. Three options are available: "linear,""log," or "symlog."
    Default: linear
  - `<yscale>`, ***string (optional)***, scale of the y axis. Three options are available: "linear," "log," or "symlog."
    Default: linear
  - `<zscale>`, ***string (optional)***, scale of the z axis. Three options are available: "linear," "log," or "symlog."
    Default: linear **Note:** This keyword is effective in 3-D plots only.

- `<addText>`, same as title.
- `<autoscale>`, is a convenience method for simple axis view autoscaling. It turns autoscaling on or off, and then, if autoscaling for either axis is on, it performs the autoscaling on the specified axis or axes. The following sub-nodes may be specified:

  - `<enable>`, ***bool (optional)***, True turns autoscaling on, False turns it off. None leaves the autoscaling state unchanged.
    Default: True

- <axis>, ***string (optional)***, determines which axis to apply the defined format, 'x,' 'y,' or 'both.'
  Default: 'both' **Note:** If this action is used in a 3-D plot, the user can input 'z' as well and 'both' will apply this format to all three axis.
- <tight>, ***bool (optional)***, if True, sets the view limits to the data limits; if False, let the locator and margins expand the view limits; if None, use tight scaling if the only output is an image file, otherwise treat tight as False.

• <horizontalLine>, this "action" provides the ability to draw a horizontal line in the current figure. This capability might be useful, for example, if the user wants to highlight a trigger function of a variable. The following sub-nodes may be specified:

  - <y>, ***double (optional)***, sets the y-value for the line.
    Default: 0
  - <xmin>, ***double (optional)***, is the starting coordinate on the x axis.
    Default: 0
  - <xmax>, ***double (optional)***, is the ending coordinate on the x axis.
    Default: 1
  - <kwargs>, within this block the user can specify optional parameters with the following format:

```
<kwargs>
 <param1>value1</param1>
 <param2>value2</param2>
</kwargs>
```

    The kwargs block is able to convert whatever string into a python type (for example <param1>{'1stKey':45}</param1> will be converted into a dictionary, <param2>[56,67]</param2> into a list, etc.). For reference regarding the available kwargs, see "matplotlib.pyplot.axhline" method in [4].

  Note: This capability is not available for 3-D plots.

• <verticalLine>, similar to the "horizontalLine" action, this block provides the ability to draw a vertical line in the current figure. This capability might be useful, for example, if the user wants to highlight a trigger function of a variable. The following sub-nodes may be specified:

  - <x>, ***double (optional)***, sets the x coordinate of the line.
    Default: 0
  - <ymin>, ***double (optional)***, starting coordinate on the y axis.
    Default: 0
  - <ymax>, ***double (optional)***, ending coordinate on the y axis.
    Default: 1

- **<kwargs>**, within this block the user can specify optional parameters with the following format:

```
<kwargs>
 <param1>value1</param1>
 <param2>value2</param2>
</kwargs>
```

The kwargs block is able to convert whatever string into a python type (for example **<param1>**`{'1stKey':45}`**</param1>** will be converted into a dictionary, **<param2>**`[56,67]`**</param2>** into a list, etc.). For reference regarding the available kwargs, see "matplotlib.pyplot.axvline" method in [4].

Note: This capability is not available for 3-D plots.

- **<horizontalRectangle>**, this "action" provides the ability to draw, in the current figure, a horizontally orientated rectangle. This capability might be useful, for example, if the user wants to highlight a zone in the plot. The following sub-nodes may be specified:
  - **<ymin>**, ***double (required)***, starting coordinate on the y axis.
  - **<ymax>**, ***double (required)***, ending coordinate on the y axis.
  - **<xmin>**, ***double (optional)***, starting coordinate on the x axis.
    Default: 0
  - **<xmax>**, ***double (optional)***, ending coordinate on the x axis. Default = 1
  - **<kwargs>**, within this block the user can specify optional parameters with the following format:

```
<kwargs>
 <param1>value1</param1>
 <param2>value2</param2>
</kwargs>
```

The kwargs block is able to convert whatever string into a python type (for example **<param1>**`{'1stKey':45}`**</param1>** will be converted into a dictionary, **<param2>**`[56,67]`**</param2>** into a list, etc.). For reference regarding the available kwargs, see "matplotlib.pyplot.axhspan" method in [4].

Note: This capability is not available for 3D plots.

- **<verticalRectangle>**, this "action" provides the possibility to draw, in the current figure, a vertically orientated rectangle. This capability might be useful, for example, if the user wants to highlight a zone in the plot. The following sub-nodes may be specified:
  - **<xmin>**, ***double (required)***, starting coordinate on the x axis.
  - **<xmax>**, ***double (required)***, ending coordinate on the x axis.

- **<ymin>**, ***double (optional)***, starting coordinate on the y axis.
  Default: 0
- **<ymax>**, ***double (optional)***, ending coordinate on the y axis.
  Default: 1
- **<kwargs>**, within this block the user can specify optional parameters with the following format:

```
<kwargs>
 <param1>value1</param1>
 <param2>value2</param2>
</kwargs>
```

  The kwargs block is able to convert whatever string into a python type (for example **<param1>**{'1stKey':45}**</param1>** will be converted into a dictionary, **<param2>**[56,67]**</param2>** into a list, etc.). For reference regarding the available kwargs, see "matplotlib.pyplot.axvspan" method in [4].

Note: This capability is not available for 3D plots.

- **<axesBox>**, this keyword controls the axes' box. Its value can be 'on' or 'off'.
- **<axisProperties>**, this block is used to set axis properties. There are no fixed keywords. If only a single property needs to be set, it can be specified as the body of this block, otherwise a dictionary-like string needs to be provided. For reference regarding the available keys, refer to "matplotlib.pyplot.axis" method in [4].
- **<grid>**, this block is used to define a grid that needs to be added in the plot. The following keywords can be inputted:
  - **<b>**, ***boolean (required)***, toggles the grid lines on or off.
  - **<which>**, ***double (required)***, ending coordinate on the x axis.
  - **<axis>**, ***double (optional)***, starting coordinate on the y axis.
    Default: 0
  - **<kwargs>**, within this block the user can specify optional parameters with the following format:

```
<kwargs>
 <param1>value1</param1>
 <param2>value2</param2>
</kwargs>
```

    The kwargs block is able to convert whatever string into a python type (for example **<param1>**{'1stKey':45}**</param1>** will be converted into a dictionary, **<param2>**[56,67]**</param2>** into a list, etc.).

16.2.1.2 "plotSettings" input block

The sub-block identified by the keyword **<plotSettings>** is used to define the plot characteristics. Within this sub-section at least a **<plot>** block must be present. the **<plot>** sub-section may not be unique within the **<plotSettings>** definition; the number of **<plot>** sub-blocks is equal to the number of plots that need to be placed in the same figure.

If sub-plots are to be defined then **<gridSpace>** needs to be present. **<gridSpace>** specifies the geometry of the grid that a subplot will be placed. The number of rows and number of columns of the grid need to be set.

For example, in the following XML cut, a "line" and a "scatter" type are combined in the same figure.

```xml
<OutStreams>
  <Plot name='example2PlotsCombined'>
    <actions>
      <!-- Actions -->
    </actions>
    <plotSettings>
      <gridSpace>2 2</gridSpace>
      <plot>
        <type>line</type>
        <x>d-type|Output|x1</x>
        <y>d-type|Output|y1</y>
        <xlabel>label X</xlabel>
        <ylabel>label Y</ylabel>
        <gridLocation>
          <x>0 2</x>
          <y>0</y>
        </gridLocation>
      </plot>
      <plot>
        <type>scatter</type>
        <x>d-type|Output|x2</x>
        <y>d-type|Output|y2</y>
        <xlabel>label X</xlabel>
        <ylabel>label Y</ylabel>
        <gridLocation>
          <x>0 2</x>
          <y>1</y>
        </gridLocation>
      </plot>
```

```
    </plotSettings>
  </Plot>
</OutStreams>
```

The axis labels are conditionally optional nodes that can be defined under the **<plotSetting>**. If the plot does not contain any sub-plots, i.e. **<gridSpace>** is not defined then the axis labels are global parameters for the figure which are defined under **<plotSettings>**, otherwise the axis labels can be defined under **<plot>** for each sub-plot seperately.

- **<xlabel>**, ***string, optional parameter***, the x axis label.
- **<ylabel>**, ***string, optional parameter***, the y axis label.
- **<zlabel>**, ***string, optional parameter (3D plots only)***, the z axis label.

One may also specify a **<legend>** tag that will place a legend on the plot. The legend accepts the following sub-nodes:

- **<loc>**, ***string, optional parameter***, the location where the legend will be placed on the plot. Valid values are:
  - **'best'**
  - **'upper right'**
  - **'upper left'**
  - **'lower left'**
  - **'lower right'**
  - **'right'**
  - **'center left'**
  - **'center right'**
  - **'lower center'**
  - **'upper center'**
  - **'center'**

  Default: ***'best'***
- **<ncol>**, ***integer, optional parameter***, the number of columsn to include in the legend. *Default: 1*
- **<fontsize>**, ***string, optional parameter***, the font size of the legend. Valid values are:

  - 'xx-small'
  - x-small
  - 'small'
  - 'medium'
  - 'large'
  - 'x-large'
  - 'xx-large'

- **<title>**, ***string, optional parameter***, the title of the legend.

Note: The text associated to each **<plot>** tag in the legend is defined in the **<kwargs>** of that plot by specifying a **<label>** within the kwargs. An example usage is given below:

```
<Plot ...>
  ...
  <plotSettings>
    <plot>
      <type>scatter</type>
        <x>...</x>
        <y>...</y>
        <kwargs>
          <label>dots</label>
        </kwargs>
     </plot>
    <plot>
      <type>line</type>
        <x>...</x>
        <y>...</y>
        <kwargs>
          <label>line</label>
        </kwargs>
     </plot>
    <legend>
      <loc>best</loc>
      <ncol>2</ncol>
    </legend>
  </plotSettings>
</Plot>
```

This will create a plot with both scattered points and a line. The plot will also have a legend specifying the labels “dots” and “line” in two columns with the best location selected by matplotlib.

As already mentioned, within the **<plotSettings>** block, at least a **<plot>** sub-block needs to be specified. Independent of the plot type, some keywords are mandatory:

- **<type>**, ***string, required parameter***, the plot type (for example, line, scatter, wireframe, etc.).
- **<x>**, ***string, required parameter***, specifies the DataObject parameter to be plotted as the x coordinate. This parameter must be described in a specific manner, see Section 16.2.1.2.1 below for details.
- **<y>**, ***string, required parameter***, specifies the DataObject parameter to be plotted as the y coordinate. This parameter must be described in a specific manner, see Section 16.2.1.2.1 below for details.
- **<z>**, ***string, required parameter for plots with three dimensions***, specifies the DataObject parameter to be plotted as the z coordinate. This parameter must be described in a specific manner, see Section 16.2.1.2.1 below for details.

In addition, other plot-dependent keywords, reported in Section 16.2.1.3, can be provided.

Under the **<plot>** sub-block other optional keywords can be specified, such as:

- **<xlabel>**, ***string, optional parameter***, the x axis label.
- **<ylabel>**, ***string, optional parameter***, the y axis label.
- **<zlabel>**, ***string, optional parameter (3D plots only)***, the z axis label.
- **<gridLocation>**, ***xmlNode, optional xmlNode (depending on the grid geometry)***
  - **<x>**, ***integer, required parameter***, the position of the subPlot in the grid Space. if this node has a single value then the subplot occupies a single node at the specified location, otherwise the second integer represents the number of nodes that this subplot occupies, i.e. in the example above the first subplot occupies 2 nodes starting from the zero node in x direction.
  - **<y>**, ***integer, required parameter***, the position of the subPlot in the grid Space. if this node has a single value then the subplot occupies a single node at the specified location, otherwise the second integer represents the number of nodes that this subplot occupies, i.e. in the example above the first subplot occupies a single node at the zero node in y direction.
- **<colorMap>**, ***string, optional parameter***, specifies a DataObject parameter whose value will be used to vary the color of plotted points. This parameter must be described in a specific manner, see Section 16.2.1.2.1 below for details.

16.2.1.2.1 Specifying What Values to Plot

As already mentioned, the Plot system accepts as input for the visual parameters (i.e., x, y, z, colorMap), data only from a **DataObjects** object. Considering the structure of "DataObjects", the parameters are specified as three values separated by the vertical bar character ('|') as follows:

```
DataObject Name|Parameter Type|Parameter Name
```

Where:

Value	Description
`DataObject Name`	Name of the DataObject that contains the parameter
`Parameter Type`	Either `Input` or `Output` depending on whether the parameter is defined in the **<Input>** or **<Output>** part of the DataObject
`Parameter Name`	The name of the parameter in the DataObject to plot

Note: If the `Parameter Name` part of the variable specification itself contains the vertical bar character ('|') used to separate the three values, it must be enclosed in parenthesis to be interpreted properly. For example:

```
DataObject Name|Parameter Type|(parameter|name)
```

16.2.1.3 Predefined Plotting System: 2D/3D

As already mentioned above, the Plotting system provides a specialized input structure for several different kind of plots specified in the **<type>** node:

- *2 Dimensional plots*:
  - `scatter` creates a scatter plot of x vs y, where x and y are sequences of numbers of the same length.
  - `line` creates a line plot of x vs y, where x and y are sequences of numbers of the same length.
  - `histogram` computes and draws the histogram of x. **Note:** This plot accepts only the XML node **<x>** even if it is considered as a 2D plot type.
  - `stem` plots vertical lines at each x location from the baseline to y, and places a marker there.
  - `step` creates a 2 dimensional step plot.
  - `pseudocolor` creates a pseudocolor plot of a two dimensional array. The two dimensional array is built creating a mesh from **<x>** and **<y>** data, in conjunction with the data specified in the **<colorMap>** node.

  - `contour` builds a contour plot creating a plot from **<x>** and **<y>** data, in conjunction with the data specified in the **<colorMap>** node.
  - `filledContour` creates a filled contour plot from **<x>** and **<y>** data, in conjunction with the data specified in the **<colorMap>** node.

- *3 Dimensional plots*:
  - `scatter` creates a scatter plot of (x,y) vs z, where x, y, z are sequences of numbers of the same length.
  - `line` creates a line plot of (x,y) vs z, where x, y, z are sequences of numbers of the same length.
  - `stem` creates a 3 Dimensional stem plot of (x,y) vs z.
  - `surface` creates a surface plot of (x,y) vs z. By default it will be colored in shades of a solid color, but it also supports color mapping.
  - `wireframe` creates a 3D wire-frame plot. No color mapping is supported.
  - `tri-surface` creates a 3D tri-surface plot. It is a surface plot with automatic triangulation.
  - `contour3D` builds a 3D contour plot creating the plot from **<x>**, **<y>** and **<z>** data, in conjunction with the data specified in **<colorMap>**.
  - `filledContour3D` builds a filled 3D contour plot creating the plot from **<x>**, **<y>** and **<z>** data, in conjunction with the data specified in **<colorMap>**.
  - `histogram` computes and draws the histogram of x and y. **Note:** This plot accepts only the XML nodes **<x>** and **<y>** even if it is considered as 3D plot type since the frequency is mapped to the third dimension.

As already mentioned, the settings for each plot type are specified within the XML block called **<plot>**. The sub-nodes that are available depends on the plot type as each plot type has its own set of parameters that can be specified.

In the following sub-sections all the options for the plot types listed above are reported.

16.2.2 2D & 3D Scatter plot

In order to create a "scatter" plot, either 2D or 3D, the user needs to write in the **<type>** body the keyword "scatter." In order to customize the plot, the user can define the following XML sub nodes:

- **<s>**, ***integer, optional field***, represents the size in points2̂. The "points" have the same meaning of the font size (e.g. Times New Roman, pts 10). In here the user specifies the area

of the marker size.
Default: 20

- **<c>**, ***string, optional field***, specifies the color or sequence of color to use. **<c>** can be a single color format string, a sequence of color specifications of length N, or a sequence of N numbers to be mapped to colors using the cmap and norm specified via **<kwargs>**. **Note:** **<c>** should not be a single numeric RGB or RGBA sequence because that is indistinguishable from an array of values to be colormapped. **<c>** can be a 2D array in which the rows are RGB or RGBA. **Note:** **<colorMap>** will overwrite **<c>**. If **<colorMap>** is defined then the color set used can be defined by **<cmap>**. If no **<cmap>** is given then the default color set of "matplotlib.pyplot.scatter" method in [4] is used. If **<colorMap>** is not defined then the plot is in solid color (default *blue*) as defined with **<color>** in **<kwargs>**.

- **<marker>**, ***string, optional field***, specifies the type of marker to use.
  Default: o

- **<alpha>**, ***string, optional field***, sets the alpha blending value, between 0 (transparent) and 1 (opaque).
  Default: None

- **<linewidths>**, ***string, optional field***, widths of lines used in the plot. Note that this is a tuple, and if you set the linewidths argument you must set it as a sequence of floats.
  Default: None;

- **<kwargs>**, within this block the user can specify optional parameters with the following format:

```
<kwargs>
 <param1>value1</param1>
 <param2>value2</param2>
</kwargs>
```

  The kwargs block is able to convert whatever string into a python type (for example `<param1>{'1stKey':45}</param1>` will be converted into a dictionary, `<param2>[56,67]</param2>` into a list, etc.). For reference regarding the available kwargs, see "matplotlib.pyplot.scatter" method in [4].

16.2.3 2D & 3D Line plot

In order to create a "line" plot, either 2D or 3D, the user needs to write in the **<type>** body the keyword "line." In order to customize the plot, the user can define the following XML sub nodes:

- **<interpolationType>**, ***string, optional field***, is the type of interpolation algorithm to use for the data. Available options are "nearest," "linear," "cubic," "multiquadric," "inverse,"

"gaussian," "Rbflinear," "Rbfcubic," "quintic," and "thin_plate."
Default: linear

- **<interpPointsX>**, ***integer, optional field***, sets the number of points need to be used for interpolation of the x axis.
- **<interpPointsY>**, ***integer, optional field***, sets the number of points need to be used for interpolation of the y axis. (only 3D line plot). **Note:** If **<colorMap>** is used then a **scatter plot** will be plotted.

16.2.4 2D & 3D Histogram plot

In order to create a "histogram" plot, either 2D or 3D, the user needs to write in the **<type>** body the keyword "histogram." In order to customize the plot, the user can define the following XML sub nodes:

- **<bins>**, ***integer or array_like, optional field***, sets the number of bins if an integer is used or a sequence of edges if a python list is used.
  Default: 10
- **<normed>**, ***boolean, optional field***, if True then the the histogram will be normalized to 1.
  Default: False
- **<weights>**, ***sequence, optional field***, represents an array of weights, of the same shape as x. Each value in x only contributes its associated weight towards the bin count (instead of 1). If normed is True, the weights are normalized, so that the integral of the density over the range remains 1.
  Default: None
- **<cumulative>**, ***boolean, optional field***, if True, then a histogram is computed where each bin gives the counts in that bin plus all bins for smaller values. The last bin gives the total number of data points. If normed is also True then the histogram is normalized such that the last bin equals 1. If cumulative evaluates to less than 0 (e.g., -1), the direction of accumulation is reversed. In this case, if normed is also True, then the histogram is normalized such that the first bin equals 1.
  Default: False
- **<histtype>**, ***string, optional field***, The type of histogram to draw:
  - **bar** is a traditional bar-type histogram. If multiple data sets are given the bars are arranged side by side.
  - **barstacked** is a bar-type histogram where multiple data sets are stacked on top of each other.

- **step** generates a line plot that is by default unfilled.
- **stepfilled** generates a line plot that is by default filled.

Default: bar

- **<align>**, ***string, optional field***, controls how the histogram is plotted.
  - **left** bars are centered on the left bin edge.
  - **mid** bars are centered between the bin edges.
  - **right** bars are centered on the right bin edges.

  Default: mid

- **<orientation>**, ***string, optional field***, specifies the orientation of the histogram:
  - **horizontal**
  - **vertical**

  Default: vertical

- **<rwidth>**, ***float, optional field***, sets the relative width of the bars as a fraction of the bin width.
  Default: None

- **<log>**, ***boolean, optional field***, sets a log scale.
  Default: False

- **<color>**, ***string, optional field***, specifies the color of the histogram.
  Default: blue;

- **<stacked>**, ***boolean, optional field***, if True, multiple data elements are stacked on top of each other. If False, multiple data sets are aranged side by side if histtype is ‘bar’ or on top of each other if histtype is ‘step.’
  Default: False

- **<kwargs>**, within this block the user can specify optional parameters with the following format:

```
<kwargs>
 <param1>value1</param1>
 <param2>value2</param2>
</kwargs>
```

The kwargs block is able to convert whatever string into a python type (for example `<param1>` `{'1stKey':45}` `</param1>` will be converted into a dictionary, `<param2>` `[56,67]` `</param2>` into a list, etc.). For reference regarding the available kwargs, see "matplotlib.pyplot.hist" method in [4].

16.2.5 2D & 3D Stem plot

In order to create a "stem" plot, either 2D or 3D, the user needs to write in the `<type>` body the keyword "stem." In order to customize the plot, the user can define the following XML sub nodes:

- `<linefmt>`, ***string, optional field***, sets the line style used in the plot.
  Default: b-
- `<markerfmt>`, ***string, optional field***, sets the type of marker format to use in the plot.
  Default: bo
- `<basefmt>`, ***string, optional field***, sets the base format.
  Default: r-
- `<kwargs>`, within this block the user can specify optional parameters with the following format:

```
<kwargs>
 <param1>value1</param1>
 <param2>value2</param2>
</kwargs>
```

  The kwargs block is able to convert whatever string into a python type (for example `<param1>` `{'1stKey':45}` `</param1>` will be converted into a dictionary, `<param2>` `[56,67]` `</param2>` into a list, etc.).

  For reference regarding the available kwargs, see "matplotlib.pyplot.stem" method in [4].

16.2.6 2D Step plot

In order to create a 2D "step" plot, the user needs to write in the `<type>` body the keyword "step." In order to customize the plot, the user can define the following XML sub nodes:

- `<where>`, ***string, optional field***, specifies the positioning:
  - **pre**, the interval from x[i] to x[i+1] has level y[i+1]
  - **post**, that interval has level y[i]

- **mid**, the jumps in y occur half-way between the x-values

Default: mid

- **<kwargs>**, within this block the user can specify optional parameters with the following format:

```
<kwargs>
 <param1>value1</param1>
 <param2>value2</param2>
</kwargs>
```

The kwargs block is able to convert whatever string into a python type (for example `<param1>{'1stKey':45}</param1>` will be converted into a dictionary, `<param2> [56,67]</param2>` into a list, etc.). For reference regarding the available kwargs, see "matplotlib.pyplot.step" method in [4].

16.2.7 2D Pseudocolor plot

In order to create a 2D "pseudocolor" plot, the user needs to write in the **<type>** body the keyword "pseudocolor." In order to customize the plot, the user can define the following XML sub nodes:

- **<interpolationType>**, ***string, optional field***, is the type of interpolation algorithm to use for the data. Available options are "nearest," "linear," "cubic," "multiquadric," "inverse," "gaussian," "Rbflinear," "Rbfcubic," "quintic," and "thin_plate."
  Default: [linear]

- **<interpPointsX>**, ***integer, optional field***, sets the number of points need to be used for interpolation of the x axis.

- **<kwargs>**, within this block the user can specify optional parameters with the following format:

```
<kwargs>
 <param1>value1</param1>
 <param2>value2</param2>
</kwargs>
```

The kwargs block is able to convert whatever string into a python type (for example `<param1>{'1stKey':45}</param1>` will be converted into a dictionary, `<param2> [56,67]</param2>` into a list, etc.). For reference regarding the available kwargs, see "matplotlib.pyplot.pcolor" method in [4].

16.2.8 2D Contour or filledContour plots

In order to create a 2D "contour" or "filledContour" plot, the user needs to write in the <type> body the keyword "contour" or "filledContour," respectively. In order to customize the plot, the user can define the following XML sub-nodes:

- <numberBins>, ***integer, optional field***, sets the number of bins.
  Default: 5
- <interpolationType>, ***string, optional field***, is the type of interpolation algorithm to use for the data. Available options are "nearest," "linear," "cubic," "multiquadric," "inverse," "gaussian," "Rbflinear," "Rbfcubic," "quintic," and "thin_plate."
  Default: linear
- <interpPointsX>, ***integer, optional field***, sets the number of points need to be used for interpolation of the x axis.
- <colorMap> vector is the array to visualize. If <colorMap> is defined then the color set used can be defined by <cmap>. If no <cmap> is given then the plot is in solid color (default *blue*) as defined with <color> in <kwargs>.
- <cmap>, ***string, optional field***, defines the color map to use for this plot.
  Default: None
- <kwargs>, within this block the user can specify optional parameters with the following format:

```
<kwargs>
 <param1>value1</param1>
 <param2>value2</param2>
</kwargs>
```

  The kwargs block is able to convert whatever string into a python type (for example <param1> {'1stKey':45}</param1> will be converted into a dictionary, <param2> [56,67] </param2> into a list, etc.). For reference regarding the available kwargs, see "matplotlib.pyplot.contour" method in [4].

16.2.9 3D Surface Plot

In order to create a 3D "surface" plot, the user needs to write in the <type> body the keyword "surface." In order to customize the plot, the user can define the following XML sub nodes:

- <rstride>, ***integer, optional field***, specifies the array row stride (step size).
  Default: 1

- **<cstride>**, ***integer, optional field***, specifies the array column stride (step size).
  Default: 1

- **<cmap>**, ***string, optional field***, defines the color map to use for this plot.
  Default: None **Note:** If **<colorMap>** is defined then the plot will always use a color set even if no **<cmap>** is given. In such a case, if no **<cmap>** is given, then the default color set of "matplotlib.pyplot.surface" method in [4] is used. If **<colorMap>** and **<cmap>** are both not defined then the plot is in solid color (*default blue*) as defined with **<color>** in **<kwargs>**.

- **<antialiased>**, ***boolean, optional field***, determines whether or not the rendering should be antialiased.
  Default: False

- **<linewidth>**, ***integer, optional field***, defines the widths of lines rendered on the plot.
  Default: 0

- **<interpolationType>**, ***string, optional field***, is the type of interpolation algorithm to use for the data. Available options are "nearest," "linear," "cubic," "multiquadric," "inverse," "gaussian," "Rbflinear," "Rbfcubic," "quintic," and "thin_plate."
  Default: linear

- **<interpPointsX>**, ***integer, optional field***, sets the number of points need to be used for interpolation of the x axis.

- **<interpPointsY>**, ***integer, optional field***, sets the number of points need to be used for interpolation of the y axis.

- **<kwargs>**, within this block the user can specify optional parameters with the following format:

```
<kwargs>
 <param1>value1</param1>
 <param2>value2</param2>
</kwargs>
```

  The kwargs block is able to convert whatever string into a python type (for example **<param1>** `{'1stKey':45}`**</param1>** will be converted into a dictionary, **<param2>** `[56,67]` **</param2>** into a list, etc.). For reference regarding the available kwargs, see "matplotlib.pyplot.surface" method in [4].

16.2.10 3D Wireframe Plot

In order to create a 3D "wireframe" plot, the user needs to write in the **<type>** body the keyword "wireframe." In order to customize the plot, the user can define the following XML sub nodes:

- **<rstride>**, ***integer, optional field***, sets the array row stride (step size).
  Default: 1

- **<cstride>**, ***integer, optional field***, sets the array column stride (step size).
  Default: 1

- **<cmap>**, ***string, optional field***, defines the color map to use for this plot.
  Default: None **Note:** **<cmap>** is not applicable in the current version of MatPlotLib for wireframe plots. However, if the colorMap option is set then a surface plot is plotted with a transparency of 0.4 on top of wireframe to give a visual colormap. **Note:** If **<colorMap>** is defined then the plot will always use a color set even if no **<cmap>** is given. In such a case, if no **<cmap>** is given, then the default color set of "matplotlib.pyplot.surface" method in [4] is used. If **<colorMap>** and **<cmap>** are both not defined then the plot is in solid color (*default blue*) as defined with **<color>** in **<kwargs>**.

- **<interpolationType>**, ***string, optional field***, is the type of interpolation algorithm to use for the data. Available options are "nearest," "linear," "cubic," "multiquadric," "inverse," "gaussian," "Rbflinear," "Rbfcubic," "quintic," and "thin_plate."
  Default: linear

- **<interpPointsX>**, ***integer, optional field***, sets the number of points need to be used for interpolation of the x axis.

- **<interpPointsY>**, ***integer, optional field***, sets the number of points need to be used for interpolation of the y axis.

- **<kwargs>**, within this block the user can specify optional parameters with the following format:

```
<kwargs>
 <param1>value1</param1>
 <param2>value2</param2>
</kwargs>
```

  The kwargs block is able to convert whatever string into a python type (for example **<param1>** `{'1stKey':45}`**</param1>** will be converted into a dictionary, **<param2>** `[56,67]` **</param2>** into a list, etc.). For reference regarding the available kwargs, see "matplotlib.pyplot.wireframe" method in [4].

16.2.11 3D Tri-surface Plot

In order to create a 3D "tri-surface" plot, the user needs to write in the **<type>** body the keyword "tri-surface." In order to customize the plot, the user can define the following XML sub nodes:

- **<color>**, ***string, optional field***, sets the color of the surface patches.
  Default: b

- **<shade>**, ***boolean, optional field***, determines whether to apply shading or not.
  Default: False

- **<cmap>**, ***string, optional field***, defines the color map to use for this plot.
  Default: None **Note:** If **<colorMap>** is defined then the plot will always use a color set even if no **<cmap>** is given. In such a case, if no **<cmap>** is given, then the default color set of "matplotlib.pyplot.trisurface" method in [4] is used. If **<colorMap>** and **<cmap>** are both not defined then the plot is in solid color (*default blue*) as defined with **<color>** in **<kwargs>**.

- **<kwargs>**, within this block the user can specify optional parameters with the following format:

```
<kwargs>
 <param1>value1</param1>
 <param2>value2</param2>
</kwargs>
```

  The kwargs block is able to convert whatever string into a python type (for example **<param1>** `{'1stKey':45}`**</param1>** will be converted into a dictionary, **<param2>** `[56,67]` **</param2>** into a list, etc.). For reference regarding the available kwargs, see "matplotlib.pyplot.trisurface" method in [4].

16.2.12 3D Contour or filledContour plots

In order to create a 3D "Contour" or "filledContour" plot, the user needs to write in the **<type>** body the keyword "contour3D" or "filledContour3D," respectively. In order to customize these plots, the user can define the following XML sub nodes:

- **<numberBins>**, ***integer, optional field***, sets the number of bins to use.
  Default: 5

- **<interpolationType>**, ***string, optional field***, is the type of interpolation algorithm to use for the data. Available options are "nearest," "linear," "cubic," "multiquadric," "inverse," "gaussian," "Rbflinear," "Rbfcubic," "quintic," and "thin_plate."
  Default: linear

- **<interpPointsX>**, ***integer, optional field***, sets the number of points need to be used for interpolation of the x axis.

- **<interpPointsY>**, ***integer, optional field***, sets the number of points need to be used for interpolation of the y axis.

- **<colorMap>** vector is the array to visualize. If **<colorMap>** is defined then the color set used can be defined by **<cmap>**. If no **<cmap>** is given then the plot is in solid color (default *blue*) as defined with **<color>** in **<kwargs>**.

- **<kwargs>**, within this block the user can specify optional parameters with the following format:

```
<kwargs>
 <param1>value1</param1>
 <param2>value2</param2>
</kwargs>
```

  The kwargs block is able to convert whatever string into a python type (for example **<param1>** {'1stKey':45}**</param1>** will be converted into a dictionary, **<param2>** [56,67] **</param2>** into a list, etc.). For reference regarding the available kwargs, see "matplotlib.pyplot.contour3d" method in [4].

16.2.13 DataMining plots

In order to create a "DataMining" plot, the user needs to write in the **<type>** body the keyword "dataMining". "DataMining" plots are based on 2D or 3D Scattering plots, depending on the method/algorithm used in the "DataMining" postprocessor [see 17.5.9]. These plots are created to ease the color labeling the clusters, etc parameters in the data. The following are the optional or required input parameters that can be used in these plots additional to the coordinate inputs **<x>**, **<y>**, or **<z>** depending on the dimension:

- **<type>**, ***string, required field***, this block should read "dataMining" in order to create a data mining plot.

- **<SKLtype>**, ***string, required field***, name of the algorithm used in the "dataMining" postprocessor. It is one of:
  - cluster: for clustering algorithms, such as KMeans clustering.
  - bicluster (**Note:** not implemented yet!)
  - mixture: for Gaussian mixture algorithms, such as GMM classifier
  - manifold: for Manifold Learning algorithms, such as Spectral Embedding
  - decomposition: for decomposing signals in components algorithms, such as Principal Component Analysis (PCA)

- **<clusterLabels>**, ***string, optional field***, defines the place where the labels of the clusters are located. As in the visual parameters (i.e., x,y,z and colorMap) this is also from a **DataObjects** object. Considering the structure of "DataObjects", the labels inputted as follows: `DataObjectName|Output|DataMiningPPNameLabels`.
Default: None

- **<noClusters>**, ***integer, optional field***, defines the number of clusters used in the "dataMining" postprocessor
Default: 1

- **<kwargs>**, within this block the user can specify optional parameters with the following format:

```
<kwargs>
 <param1>value1</param1>
 <param2>value2</param2>
</kwargs>
```

The kwargs block is able to convert whatever string into a python type (for example `<param1>{'1stKey':45}</param1>` will be converted into a dictionary, `<param2>[56,67]</param2>` into a list, etc.).

For reference regarding the other available kwargs, see "matplotlib.pyplot.scatter" method in [4].

16.2.14 Example XML input

```
<OutStreams>
  <Plot name='2DHistoryPlot' interactive='False'
     overwrite='False'>
    <actions>
      <how>pdf,png,eps</how>
      <title>
        <text>***</text>
      </title>
    </actions>
    <plotSettings>
       <plot>
        <type>line</type>
        <x>stories|Output|time</x>
        <y>stories|Output|pipe1_Hw</y>
        <kwargs>
         <color>green</color>
         <label>pipe1-Hw</label>
```

```xml
        </kwargs>
      </plot>
       <plot>
        <type>line</type>
        <x>stories|Output|time</x>
        <y>stories|Output|pipe1_aw</y>
        <kwargs>
         <color>blue</color>
         <label>pipe1-aw</label>
        </kwargs>
      </plot>
      <xlabel>time [s]</xlabel>
      <ylabel>evolution</ylabel>
    </plotSettings>
  </Plot>
</OutStreams>
```

17 Models

In RAVEN, **Models** are important entities. A model is an object that employs a mathematical representation of a phenomenon, either of a physical or other nature (e.g. statistical operators, etc.). From a practical point of view, it can be seen, as a "black box" that, given an input, returns an output.

RAVEN has a strict classification of the different types of models. Each "class" of models is represented by the definition reported above, but it can be further classified based on its particular functionalities:

- **<Code>** represents an external system code that employs a high fidelity physical model.
- **<Dummy>** acts as "transfer" tool. The only action it performs is transferring the the information in the input space (inputs) into the output space (outputs). For example, it can be used to check the effect of a sampling strategy, since its outputs are the sampled parameters' values (input space) and a counter that keeps track of the number of times an evaluation has been requested.
- **<ROM>**, or reduced order model, is a mathematical model trained to predict a response of interest of a physical system. Typically, ROMs trade speed for accuracy representing a faster, rough estimate of the underlying phenomenon. The "training" process is performed by sampling the response of a physical model with respect to variation of its parameters subject to probabilistic behavior. The results (outcomes of the physical model) of the sampling are fed into the algorithm representing the ROM that tunes itself to replicate those results.
- **<ExternalModel>**, as its name suggests, is an entity existing outside the RAVEN framework that is embedded in the RAVEN code at run time. This object allows the user to create a Python module that will be treated as a predefined internal model object.
- **<EnsembleModel>** is model that is able to combine **Code**, **ExternalModel** and **ROM** models. It is aimed to create a chain of Models (whose execution order is determined by the Input/Output relationships among them). If the relationships among the models evolve in a non-linear system, a Picard's Iteration scheme is employed.
- **<PostProcessor>** is a container of all the actions that can manipulate and process a data object in order to extract key information, such as statistical quantities, clustering, etc.

Before analyzing each model in detail, it is important to mention that each type needs to be contained in the main XML node **<Models>**, as reported below:

Example:

```xml
<Simulation>
  ...
  <Models>
    ...
    <WhatEverModel name='whatever'>
      ...
    </WhatEverModel>
    ...
  </Models>
  ...
</Simulation>
```

In the following sub-sections each **Model** type is fully analyzed and described.

17.1 Code

As already mentioned, the **Code** model represents an external system software employing a high fidelity physical model. The link between RAVEN and the driven code is performed at run-time, through coded interfaces that are the responsible for transferring information from the code to RAVEN and vice versa. In Section 21, all of the available interfaces are reported and, for advanced users, Section 22 explains how to couple a new code.

The specifications of this model must be defined within a `<Code>` XML block. This XML node accepts the following attributes:

- `name`, ***required string attribute***, user-defined identifier of this model. **Note:** As with other objects, this identifier can be used to reference this specific entity from other input blocks in the XML.
- `subType`, ***required string attribute***, specifies the code that needs to be associated to this Model. **Note:** See Section 21 for a list of currently supported codes.

This model can be initialized with the following children:

- `<executable>` ***string, required field*** specifies the path of the executable to be used. **Note:** Either an absolute or relative path can be used.
- `<alias>` ***string, optional field*** specifies alias for any variable of interest in the input or output space for the Code. These aliases can be used anywhere in the RAVEN input to refer to the Code variables. In the body of this node the user specifies the name of the variable that the model is going to use (during its execution). The actual alias, usable throughout the RAVEN input, is instead defined in the `variable` attribute of this tag.

The user can specify aliases for both the input and the output space. As sanity check, RAVEN requires an additional required attribute **type**. This attribute can be either "input" or "output". **Note:** The user can specify as many aliases as needed.
Default: None

- **<clargs>** ***string, optional field*** allows addition of command-line arguments to the execution command. If the code interface specified in **<Code>** **subType** does not specify how to determine the input file(s), this node must be used to specify them. There are several types of **<clargs>**, based on the **type**:
  - **type** ***string, required field*** specifies the type of command-line argument to add. Options include **'input'**, **'output'**, **'prepend'**, **'postpend'**, and **'text'**.
  - **arg** ***string, optional field*** specifies the flag to be used before the entry. For example, **arg='-i'** would place a `-i` before the entry in the execution command. Required for the **'output'** **type**.
  - **extension** ***string, optional field*** specifies the type of file extension to use (for example, `-i` or `-o`). This links the **<Input>** file in the **<Step>** to this location in the execution command. Required for **'input'** **type**.

  The execution command is combined in the order **'prepend'**, **<executable>**, **'input'**, **'output'**, **'text'**, **'postpend'**.

- **<fileargs>** ***string, optional field*** like **<clargs>**, but allows editing of input files to specify the output filename and/or auxiliary file names. The location in the input files to edit using these arguments are identified in the input file using the prefix-postfix notation, which defaults to `$RAVEN-var$` for variable keyword *var*. The variable keyword is then listed in the **<fileargs>** node in the attribute **arg** to couple it in Raven. If the code interface specified in **<Code>** **subType** does not specify how to name the output file, that must be specified either through **<clargs>** or **<filargs>**, with **type** **'output'**. The attributes required for **<fileargs>** are as follows:
  - **type** ***string, required field*** specifies the type of entry to replace in the file. Possible values for **<fileargs>** **type** are **'input'** and **'output'**.
  - **arg** ***string, required field*** specifies the Raven variable with which to replace the file of interest. This should match the entry in the template input file; that is, if `$RAVEN-auxinp$` is in the input file, the arg for the corresponding input file should be **'auxinp'**.
  - **extension** ***string, optional field*** specifies the extension of the input file that should replace the Raven variable in the input file. This attribute is required for the **'input'** **type** and ignored for the **'output'** **type**. **Note:** Currently, there can only be a one-to-one pairing between input files and extensions; that is, multiple Raven-editable input files cannot have the same extension.

Example:

```
<Simulation>
  ...
  <Models>
    ...
    <Code name='aUserDefinedName' subType='RAVEN_Driven_code'>
      <executable>path_to_executable</executable>
      <alias variable='internal_RAVEN_input_variable_name1'
         type="input">
         External_Code_input_Variable_Name_1
      </alias>
      <alias variable='internal_RAVEN_input_variable_name2'
         type='input'>
         External_Code_input_Variable_Name_2
      </alias>
      <alias variable='internal_RAVEN__output_variable_name'
         type='output'>
         External_Code_output_Variable_Name_2
      </alias>
      <clargs type='prepend' arg='python'/>
      <clargs type='input' arg='-i' extension='.i'/>
      <fileargs type='input' arg='aux' extension='.two'
      <fileargs type='output' arg='out' />
    </Code>
    ...
  </Models>
  ...
</Simulation>
```

17.2 Dummy

The **Dummy** model is an object that acts as a pass-through tool. The only action it performs is transferring the information in the input space (inputs) to the output space (outputs). For example, it can be used to check the effect of a particular sampling strategy, since its outputs are the sampled parameters' values (input space) and a counter that keeps track of the number of times an evaluation has been requested.

The specifications of this model must be defined within a **<Dummy>** XML block. . This XML node accepts the following attributes:

- name, ***required string attribute***, user-defined identifier of this model. **Note:** As with other

objects, this identifier can be used to reference this specific entity from other input blocks in the XML.

- **subType**, ***required string attribute***, this attribute must be kept empty.

This model can be initialized with the following children:

- **<alias>** ***string, optional field*** specifies alias for any variable of interest in the input or output space for the Dummy. These aliases can be used anywhere in the RAVEN input to refer to the Dummy variables. In the body of this node the user specifies the name of the variable that the model is going to use (during its execution). The actual alias, usable throughout the RAVEN input, is instead defined in the **variable** attribute of this tag.
  The user can specify aliases for both the input and the output space. As sanity check, RAVEN requires an additional required attribute **type**. This attribute can be either "input" or "output". **Note:** The user can specify as many aliases as needed.
  Default: None
  Since the **Dummy** model represents a transfer function only, the usage of the alias is relatively meaningless.

Given a particular *Step* using this model, if this model is linked to a *Data* with the role of **Output**, it expects one of the output parameters will be identified by the keyword "OutputPlaceHolder" (see Section 20).

Example:

```xml
<Simulation>
  ...
  <Models>
    ...
    <Dummy name='aUserDefinedName1' subType=''/>

    <Dummy name='aUserDefinedName2' subType=''>
      <alias variable="a_RAVEN_input_variable" type="input">
      another_name_for_this_variable_in_the_model
      </alias>
    </Dummy>
    ...
  </Models>
  ...
</Simulation>
```

17.3 ROM

A Reduced Order Model (ROM) is a mathematical model consisting of a fast solution trained to predict a response of interest of a physical system. The "training" process is performed by sampling the response of a physical model with respect to variations of its parameters subject, for example, to probabilistic behavior. The results (outcomes of the physical model) of the sampling are fed into the algorithm representing the ROM that tunes itself to replicate those results. RAVEN supports several different types of ROMs, both internally developed and imported through an external library called "scikit-learn" [5].

Currently in RAVEN, the ROMs are classified into several sub-types that, once chosen, provide access to several different algorithms. These sub-types are specified in the **subType** attribute and should be one of the following:

- **'NDspline'**
- **'GaussPolynomialRom'**
- **'HDMRRom'**
- **'NDinvDistWeight'**
- **'SciKitLearn'**
- **'MSR'**
- **'ARMA'**

The specifications of this model must be defined within a **<ROM>** XML block. This XML node accepts the following attributes:

- **name**, ***required string attribute***, user-defined identifier of this model. **Note:** As with other objects, this identifier can be used to reference this specific entity from other input blocks in the XML.
- **subType**, ***required string attribute***, defines which of the sub-types should be used, choosing among the previously reported types. This choice conditions the subsequent the required and/or optional **<ROM>** sub-nodes.

In the **<ROM>** input block, the following XML sub-nodes are required, independent of the **subType** specified:

- **<Features>**, ***comma separated string, required field***, specifies the names of the features of this ROM. **Note:** These parameters are going to be requested for the training of this object (see Section 20.4);

- **<Target>**, ***comma separated string, required field***, contains a comma separated list of the targets of this ROM. These parameters are the Figures of Merit (FOMs) this ROM is supposed to predict. **Note:** These parameters are going to be requested for the training of this object (see Section 20.4).

In addition, if the user wants to use the alias system, the following XML block can be inputted:

- **<alias>** ***string, optional field*** specifies alias for any variable of interest in the input or output space for the ROM. These aliases can be used anywhere in the RAVEN input to refer to the ROM variables. In the body of this node the user specifies the name of the variable that the model is going to use (during its execution). The actual alias, usable throughout the RAVEN input, is instead defined in the **variable** attribute of this tag.
  The user can specify aliases for both the input and the output space. As sanity check, RAVEN requires an additional required attribute **type**. This attribute can be either "input" or "output". **Note:** The user can specify as many aliases as needed.
  Default: None

The types and meaning of the remaining sub-nodes depend on the sub-type specified in the attribute **subType**.

Note that if an HistorySet is provided in the training step then a temporal ROM is created, i.e. a ROM that generates not a single value prediction of each element indicated in the **<Target>** block but its full temporal profile.

It is important to NOTE that RAVEN uses a Z-score normalization of the training data before constructing most of the Reduced Order Models (e.g. most of the SciKitLearn-based ROMs):

$$\mathbf{X}' = \frac{(\mathbf{X} - \mu)}{\sigma} \tag{3}$$

In the following sections the specifications of each ROM type are reported, highlighting when a **Z-score normalization** is performed by RAVEN before constructing the ROM or when it is not performed.

17.3.1 NDspline

The NDspline sub-type contains a single ROM type, based on an N-dimensional spline interpolation/extrapolation scheme. In spline interpolation, the interpolant is a special type of piecewise polynomial called a spline. The interpolation error can be made small even when using low degree polynomials for the spline. Spline interpolation avoids the problem of Runge's phenomenon, in which oscillation can occur between points when interpolating using higher degree polynomials.

In order to use this ROM, the <ROM> attribute **subType** needs to be **'NDspline'** (see the example below). No further XML sub-nodes are required. **Note:** This ROM type must be trained from a regular Cartesian grid. Thus, it can only be trained from the outcomes of a grid sampling strategy.

It is important to NOTE that RAVEN uses a Z-score normalization of the training data before constructing the *NDspline* ROM:

$$\mathbf{X}' = \frac{(\mathbf{X} - \mu)}{\sigma} \tag{4}$$

Example:

```
<Simulation>
  ...
  <Models>
    ...
    <ROM name='aUserDefinedName' subType='NDspline'>
       <Features>var1,var2,var3</Features>
       <Target>result1,result2</Target>
     </ROM>
    ...
  </Models>
  ...
</Simulation>
```

17.3.2 pickledROM

It is not uncommon for a reduced-order model (ROM) to be created and trained in one RAVEN run, then serialized to file (*pickled*), then loaded into another RAVEN run to be used as a model. When this is the case, a <ROM> with subtype **'pickledROM'** is used to hold the place of the ROM that will be loaded from file. The notation for this ROM is much less than a typical ROM; it only requires a name and its subtype.

Note that when loading ROMs from file, RAVEN will not perform any checks on the expected inputs or outputs of a ROM; it is expected that a user know at least the I/O of a ROM before trying to use it as a model. However, RAVEN does require that pickled ROMs be trained before pickling in the first place.

Initially, a pickledROM is not usable. It cannot be trained or sampled; attempting to do so will raise an error. An <IOStep> is used to load the ROM from file, at which point the ROM will have all the same characteristics as when it was pickled in a previous RAVEN run.

Example: For this example the ROM has already been created and trained in another RAVEN run, then pickled to a file called rom_pickle.pk. In the example, the file is identified in **<Files>**, the model is defined in **<Models>**, and the model loaded in **<Steps>**.

```
<Simulation>
  ...
  <Files>
    <Input name="rompk" type="">rom_pickle.pk</Input>
  </Files>
  ...
  <Models>
    ...
    <ROM name="myRom" subType="pickledROM"/>
    ...
  </Models>
  ...
  <Steps>
    ...
    <IOStep name="loadROM">
      <Input class="Files" type="">rompk</Input>
      <Output class="Models" type="ROM">myRom</Output>
    </IOStep>
    ...
  </Steps>
  ...
</Simulation>
```

17.3.3 GaussPolynomialRom

The GaussPolynomialRom sub-type contains a single ROM type, based on a characteristic Gaussian polynomial fitting scheme: generalized polynomial chaos expansion (gPC). In gPC, sets of polynomials orthogonal with respect to the distribution of uncertainty are used to represent the original model. The method converges moments of the original model faster than Monte Carlo for small-dimension uncertainty spaces ($N < 15$). In order to use this ROM, the **<ROM>** attribute **subType** needs to be **'GaussPolynomialRom'** (see the example below). The GaussPolynomialRom is dependent on specific sampling; thus, this ROM cannot be trained unless a SparseGridCollocation or similar Sampler specifies this ROM in its input and is sampled in a MultiRun step. In addition to the common **<Target>** and **<Features>**, this ROM requires two more nodes and can accept multiple entries of a third optional node.

- **<IndexSet>**, ***string, required field***, specifies the rules by which to construct multidimensional polynomials. The options are **'TensorProduct'**, **'TotalDegree'**, **'HyperbolicCross'**, and **'Custom'**. Total degree is efficient for uncertain inputs with a large degree of regularity, while hyperbolic cross is more efficient for low-regularity input spaces. If **'Custom'** is chosen, the **<IndexPoints>** is required.

- **<PolynomialOrder>**, ***integer, required field***, indicates the maximum polynomial order in any one dimension to use in the polynomial chaos expansion. **Note:** If non-equal importance weights are supplied in the optional **<Interpolation>** node, the actual polynomial order in dimensions with high importance might exceed this value; however, this value is still used to limit the relative overall order.

- **<SparseGrid>**,***string, optional field***, allows specification of the multidimensional quadrature construction strategy. Options are **'smolyak'** and **'tensor'**. Default is **'smolyak'**.

- **<IndexPoints>**, ***list of tuples, required field***, used to specify the index set points in a **'Custom'** index set. The tuples are entered as comma-seprated values between parenthesis, with each tuple separated by a comma. Any amount of whitespace is acceptable. For example, **<IndexPoints>**`(0,1),(0,2),(1,1),(4,0)`**</IndexPoints>** **Note:** Using custom index sets does not guarantee accurate convergence.

- **<Interpolation>**, ***string, optional field***, offers the option to specify quadrature, polynomials, and importance weights for the given variable name. The ROM accepts any number of **<Interpolation>** nodes up to the dimensionality of the input space. This node accepts several attributes, all of which are optional and default to the code-defined optimal choices based on the input dimension uncertainty distribution:

  – **quad**, ***string, optional field***, specifies the quadrature type to use for collocation in this dimension. The default options depend on the uncertainty distribution of the input dimension, as shown in Table 2. Additionally, Clenshaw Curtis quadrature can be used for any distribution that doesn't include an infinite bound.
  Default: see Table 2. **Note:** For an uncertain distribution aside from the four listed on Table 2, this ROM makes use of the uniform-like range of the distribution's CDF to apply quadrature that is suited uniform uncertainty (Legendre). It converges more slowly than the four listed, but are viable choices. Choosing polynomial type Legendre for any non-uniform distribution will enable this formulation automatically.

  – **poly**, ***string,optional field***, specifies the interpolating polynomial family to use for the polynomial expansion in this dimension. The default options depend on the quadrature type chosen, as shown in Table 2. Currently, no polynomials are available outside the default.
  Default: see Table 2.

  – **weight**, ***float, optional field***, delineates the importance weighting of this dimension. A larger importance weight will result in increased resolution for this dimension at the cost of resolution in lower-weighted dimensions. The algorithm normalizes weights at run-time.
  Default: 1.

Note: This ROM type must be trained from a collocation quadrature set. Thus, it can only be trained from the outcomes of a SparseGridCollocation sampler. Also, this ROM must be referenced

Unc. Distribution	Default Quadrature	Default Polynomials
Uniform	Legendre	Legendre
Normal	Hermite	Hermite
Gamma	Laguerre	Laguerre
Beta	Jacobi	Jacobi
Other	Legendre*	Legendre*

Table 2. GaussPolynomialRom defaults

in the SparseGridCollocation sampler in order to accurately produce the necessary sparse grid points to train this ROM.

It is important to NOTE that RAVEN does not pre-normalize the training data before constructing the *GaussPolynomialRom* ROM.

Example:

```
<Simulation>
  ...
  <Samplers>
    ...
    <SparseGridCollocation name="mySG" parallel="0">
      <variable name="x1">
        <distribution>myDist1</distribution>
      </variable>
      <variable name="x2">
        <distribution>myDist2</distribution>
      </variable>
      <ROM class = 'Models' type = 'ROM' >myROM</ROM>
    </SparseGridCollocation>
    ...
  </Samplers>
  ...
  <Models>
    ...
    <ROM name='myRom' subType='GaussPolynomialRom'>
      <Target>ans</Target>
      <Features>x1,x2</Features>
      <IndexSet>TotalDegree</IndexSet>
      <PolynomialOrder>4</PolynomialOrder>
      <Interpolation quad='Legendre' poly='Legendre'
         weight='1'>x1</Interpolation>
      <Interpolation quad='ClenshawCurtis' poly='Jacobi'
         weight='2'>x2</Interpolation>
    </ROM>
    ...
```

```
  </Models>
  ...
</Simulation>
```

When Printing this ROM via a Print OutStream (see 16.1), the available metrics are:

- **'mean'**, the mean value of the ROM output within the input space it was trained,
- **'variance'**, the variance of the ROM output within the input space it was trained,
- **'samples'**, the number of distinct model runs required to construct the ROM,
- **'indices'**, the Sobol sensitivity indices (in percent), Sobol total indices, and partial variances,
- **'polyCoeffs'**, the polynomial expansion coefficients (PCE moments) of the ROM. These are listed by each polynomial combination, with the polynomial order tags listed in the order of the variables shown in the XML print.

17.3.4 HDMRRom

The HDMRRom sub-type contains a single ROM type, based on a Sobol decomposition scheme. In Sobol decomposition, also known as high-density model reduction (HDMR, specifically Cut-HDMR), a model is approximated as as the sum of increasing-complexity interactions. At its lowest level (order 1), it treats the function as a sum of the reference case plus a functional of each input dimesion separately. At order 2, it adds functionals to consider the pairing of each dimension with each other dimension. The benefit to this approach is considering several functions of small input cardinality instead of a single function with large input cardinality. This allows reduced order models like generalized polynomial chaos (see 17.3.3) to approximate the functionals accurately with few computations runs. In order to use this ROM, the **<ROM>** attribute **subType** needs to be **'HDMRRom'** (see the example below). The HDMRRom is dependent on specific sampling; thus, this ROM cannot be trained unless a Sobol or similar Sampler specifies this ROM in its input and is sampled in a MultiRun step. In addition to the common **<Target>** and **<Features>**, this ROM requires the same nodes as the GaussPolynomialRom (see 17.3.3. Additionally, this ROM requires the **<SobolOrder>** node.

- **<SobolOrder>**, ***integer, required field***, indicates the maximum cardinality of the input space used in the subset functionals. For example, order 1 includes only functionals of each independent dimension separately, while order 2 considers pair-wise interactions.

Note: This ROM type must be trained from a Sobol decomposition training set. Thus, it can only be trained from the outcomes of a Sobol sampler. Also, this ROM must be referenced in the

Sobol sampler in order to accurately produce the necessary sparse grid points to train this ROM. Experience has shown order 2 Sobol decompositions to include the great majority of uncertainty in most models.

It is important to NOTE that RAVEN does not pre-normalize the training data before constructing the *HDMRRom* ROM.

Example:

```
<Samplers>
  ...
  <Sobol name="mySobol" parallel="0">
    <variable name="x1">
      <distribution>myDist1</distribution>
    </variable>
    <variable name="x2">
      <distribution>myDist2</distribution>
    </variable>
    <ROM class = 'Models' type = 'ROM' >myHDMR</ROM>
  </Sobol>
  ...
</Samplers>
...
<Models>
  ...
  <ROM name='myHDMR' subType='HDMRRom'>
    <Target>ans</Target>
    <Features>x1,x2</Features>
    <SobolOrder>2</SobolOrder>
    <IndexSet>TotalDegree</IndexSet>
    <PolynomialOrder>4</PolynomialOrder>
    <Interpolation quad='Legendre' poly='Legendre'
       weight='1'>x1</Interpolation>
    <Interpolation quad='ClenshawCurtis' poly='Jacobi'
       weight='2'>x2</Interpolation>
  </ROM>
  ...
</Models>
```

When Printing this ROM via an OutStream (see 16.1), the available metrics are:

- **'mean'**, the mean value of the ROM output within the input space it was trained,
- **'variance'**, the ANOVA-calculated variance of the ROM output within the input space it was trained.
- **'samples'**, the number of distinct model runs required to construct the ROM,

- **'indices'**, the Sobol sensitivity indices (in percent), Sobol total indices, and partial variances.

17.3.5 MSR

The MSR sub-type contains a class of ROMs that perform a topological decomposition of the data into approximately monotonic regions and fits weighted linear patches to the identified monotonic regions of the input space. Query points have estimated probabilities that they belong to each cluster. These probabilities can eitehr be used to give a smooth, weighted prediction based on the associated linear models, or a hard classification to a particular local linear model which is then used for prediction. Currently, the probability prediction can be done using kernel density estimation (KDE) or through a one-versus-one support vector machine (SVM).

In order to use this ROM, the **<ROM>** attribute **subType** needs to be **'MSR'** (see the associated example). This model can be initialized with the following children:

- **<persistence>**, ***string, optional field***, specifies how to define the hierarchical simplification by assigning a value to each local minimum and maximum according to the one of the strategy options below:
  - `difference` - The function value difference between the extremum and its closest-valued neighboring saddle.
  - `probability` - The probability integral computed as the sum of the probability of each point in a cluster divided by the count of the cluster.
  - `count` - The count of points that flow to or from the extremum.

  Default: `difference`
- **<gradient>**, ***string, optional field***, specifies the method used for estimating the gradient, available options are:
  - `steepest`

  Default: `steepest`
- **<simplification>**, ***float, optional field***, specifies the amount of noise reduction to apply before returning labels.
  Default: 0
- **<graph>** ***, string, optional field***, specifies the type of neighborhood graph used in the algorithm, available options are:

- `beta skeleton`
- `relaxed beta skeleton`
- `approximate knn`

Default: `beta skeleton`

- **<beta>**, ***float in the range: (0,2], optional field***, is only used when the **<graph>** is set to `beta skeleton` or `relaxed beta skeleton`.
  Default: 1.0
- **<knn>**, ***integer, optional field***, is the number of neighbors when using the `approximate knn` for the **<graph>** sub-node and used to speed up the computation of other graphs by using the approximate knn graph as a starting point for pruning. -1 means use a fully connected graph.
  Default: -1
- **<partitionPredictor>**, ***string, optional***, a flag that specifies how the predictions for query point classification should be performed. Available options are:
  - `kde`
  - `svm`

  Default: kde
- **<smooth>**, if this node is present, the ROM will blend the estimates of all of the local linear models weighted by the probability the query point is classified as belonging to that partition of the input space.
- **<kernel>**, ***string, optional field***, this option is only used when the **<partitionPredictor>** is set to `kde` and specifies the type of kernel to use in the kernel density estimation. Available options are:
  - `uniform`
  - `triangular`
  - `gaussian`
  - `epanechnikov`
  - `biweight` or `quartic`
  - `triweight`
  - `tricube`
  - `cosine`

- logistic
- silverman
- exponential

Default: gaussian

- <bandwidth>, ***float or string, optional field***, this option is only used when the <partitionPredictor> is set to kde and specifies the scale of the fall-off. A higher bandwidth implies a smooother blending. If set to variable, then the bandwidth will be set to the distance of the k-nearest neighbor of the query point where k is set by the <knn> parameter.
  Default: 1.

It is important to NOTE that RAVEN does not pre-normalize the training data before constructing the *MSR* ROM.

Example:

```
<Simulation>
  ...
  <Models>
    ...
    </ROM>
    <ROM name='aUserDefinedName' subType='MSR'>
       <Features>var1,var2,var3</Features>
       <Target>result1,result2</Target>
       <!-- <weighted>true</weighted> -->
       <simplification>0.0</simplification>
       <persistence>difference</persistence>
       <gradient>steepest</gradient>
       <graph>beta skeleton</graph>
       <beta>1</beta>
       <knn>8</knn>
       <partitionPredictor>kde</partitionPredictor>
       <kernel>gaussian</kernel>
       <smooth/>
       <bandwidth>0.2</bandwidth>
     </ROM>
    ...
  </Models>
  ...
</Simulation>
```

17.3.6 NDinvDistWeight

The NDinvDistWeight sub-type contains a single ROM type, based on an N-dimensional inverse distance weighting formulation. Inverse distance weighting (IDW) is a type of deterministic method for multivariate interpolation with a known scattered set of points. The assigned values to unknown points are calculated via a weighted average of the values available at the known points.

In order to use this Reduced Order Model, the **<ROM>** attribute **subType** needs to be xml-StringNDinvDistWeight (see the example below). This model can be initialized with the following child:

- **<p>**, ***integer, required field***, must be greater than zero and represents the "power parameter". For the choice of value for **<p>**,it is necessary to consider the degree of smoothing desired in the interpolation/extrapolation, the density and distribution of samples being interpolated, and the maximum distance over which an individual sample is allowed to influence the surrounding ones (lower p means greater importance for points far away).

It is important to NOTE that RAVEN uses a Z-score normalization of the training data before constructing the *NDinvDistWeight* ROM:

$$\mathbf{X}' = \frac{(\mathbf{X} - \mu)}{\sigma} \tag{5}$$

Example:

```
<Simulation>
  ...
  <Models>
    ...
    <ROM name='aUserDefinedName' subType='NDinvDistWeight'>
      <Features>var1,var2,var3</Features>
      <Target>result1,result2</Target>
      <p>3</p>
     </ROM>
    ...
  </Models>
  ...
</Simulation>
```

17.3.7 SciKitLearn

The SciKitLearn sub-type represents the container of several ROMs available in RAVEN through the external library scikit-learn [5].

In order to use this Reduced Order Model, the **<ROM>** attribute **subType** needs to be **'SciKitLearn'** (i.e. **subType='SciKitLearn'**). The specifications of a **'SciKitLearn'** ROM depend on the value assumed by the following sub-node within the main **<ROM>** XML node:

- **<SKLtype>**, ***vertical bar (|) separated string, required field***, contains a string that represents the ROM type to be used. As mentioned, its format is:
  <SKLtype>`mainSKLclass|algorithm`**</SKLtype>** where the first word (before the "|" symbol) represents the main class of algorithms, and the second word (after the "|" symbol) represents the specific algorithm.

Based on the **<SKLtype>** several different algorithms are available. In the following paragraphs a brief explanation and the input requirements are reported for each of them.

17.3.7.1 Linear Models

The LinearModels' algorithms implement generalized linear models. They include Ridge regression, Bayesian regression, lasso, and elastic net estimators computed with least angle regression and coordinate descent. This class also implements stochastic gradient descent related algorithms. In the following, all of the linear models available in RAVEN are reported.

17.3.7.1.1 Linear Model: Automatic Relevance Determination Regression

The *Automatic Relevance Determination* (ARD) regressor is a hierarchical Bayesian approach where hyperparameters explicitly represent the relevance of different input features. These relevance hyperparameters determine the range of variation for the parameters relating to a particular input, usually by modelling the width of a zero-mean Gaussian prior on those parameters. If the width of the Gaussian is zero, then those parameters are constrained to be zero, and the corresponding input cannot have any effect on the predictions, therefore making it irrelevant. ARD optimizes these hyperparameters to discover which inputs are relevant. In order to use the *Automatic Relevance Determination regressor*, the user needs to set the sub-node:

<SKLtype>`linear_model|ARDRegression`**</SKLtype>**.

In addition to this XML node, several others are available: .

- **<n_iter>**, ***integer, optional field***, is the maximum number of iterations.
  Default: 300

- **<tol>**, ***float, optional field***, stop the algorithm if the convergence error felt below the tolerance specified here.
  Default: 1.e-3
- **<alpha_1>**, ***float, optional field***, is a shape hyperparameter for the Gamma distribution prior over the α parameter.
  Default: 1.e-6
- **<alpha_2>**, ***float, optional field***, inverse scale hyperparameter (rate parameter) for the Gamma distribution prior over the α parameter.
  Default: 1.e-6
- **<lambda_1>**, ***float, optional field***, shape hyperparameter for the Gamma distribution prior over the λ parameter.
  Default: 1.e-6
- **<lambda_2>**, ***float, optional field***, inverse scale hyperparameter (rate parameter) for the Gamma distribution prior over the λ parameter.
  Default: 1.e-6
- **<compute_score>**, ***boolean, optional field***, if True, compute the objective function at each step of the model.
  Default: False
- **<threshold_lambda>**, ***float, optional field***, specifies the threshold for removing (pruning) weights with high precision from the computation.
  Default: 1.e+4
- **<fit_intercept>**, ***boolean, optional field***, determines whether to calculate the intercept for this model. If set to False, no intercept will be used in the calculations (e.g. data is expected to be already centered).
  Default: True
- **<normalize>**, ***boolean, optional field***, if True, the regressors X will be normalized before regression.
  Default: False
- **<verbose>**, ***boolean, optional field***, use verbose mode when fitting the model.
  Default: False

It is important to NOTE that RAVEN does not pre-normalize the training data before constructing the *ARDRegression* ROM.

17.3.7.1.2 Linear Model: Bayesian ridge regression

The *Bayesian ridge regression* estimates a probabilistic model of the regression problem as described above. The prior for the parameter w is given by a spherical Gaussian:

$$p(w|\lambda) = \mathcal{N}(w|0, \lambda^{-1}\mathbf{I_p}) \tag{6}$$

The priors over α and λ are chosen to be gamma distributions, the conjugate prior for the precision of the Gaussian. The resulting model is called Bayesian ridge regression, and is similar to the classical ridge regression. The parameters w, α, and λ are estimated jointly during the fit of the model. The remaining hyperparameters are the parameters of the gamma priors over α and λ. These are usually chosen to be non-informative. The parameters are estimated by maximizing the marginal log likelihood. In order to use the *Bayesian ridge regressor*, the user needs to set the sub-node:

<SKLtype>`linear_model|BayesianRidge`**</SKLtype>**.

In addition to this XML node, several others are available:

- **<n_iter>**, ***integer, optional field***, is the maximum number of iterations.
  Default: 300
- **<tol>**, ***float, optional field***, stop the algorithm if the convergence error felt below the tolerance specified here.
  Default: 1.e-3
- **<alpha_1>**, ***float, optional field***, is a shape hyperparameter for the Gamma distribution prior over the α parameter.
  Default: 1.e-6
- **<alpha_2>**, ***float, optional field***, inverse scale hyperparameter (rate parameter) for the Gamma distribution prior over the α parameter.
  Default: 1.e-6
- **<lambda_1>**, ***float, optional field***, shape hyperparameter for the Gamma distribution prior over the λ parameter.
  Default: 1.e-6
- **<lambda_2>**, ***float, optional field***, inverse scale hyperparameter (rate parameter) for the Gamma distribution prior over the λ parameter.
  Default: 1.e-6
- **<compute_score>**, ***boolean, optional field***, if True, compute the objective function at each step of the model.
  Default: False

- **<fit_intercept>**, ***boolean, optional field***, determines whether to calculate the intercept for this model. If set to False, no intercept will be used in the calculations (e.g. data is expected to be already centered).
  Default: True

- **<normalize>**, ***boolean, optional field***, if True, the regressors X will be normalized before regression.
  Default: False

- **<verbose>**, ***boolean, optional field***, use verbose mode when fitting the model.
  Default: False

It is important to NOTE that RAVEN does not pre-normalize the training data before constructing the *BayesianRidge* ROM.

17.3.7.1.3 Linear Model: Elastic Net

The *Elastic Net* is a linear regression technique with combined L1 and L2 priors as regularizers. It minimizes the objective function:

$$1/(2*n_{samples})*||y-Xw||_2^2+alpha*l1_ratio*||w||_1+0.5*alpha*(1-l1_ratio)*||w||_2^2 \quad (7)$$

In order to use the *Elastic Net regressor*, the user needs to set the sub-node:

<SKLtype>linear_model|ElasticNet**</SKLtype>**.

In addition to this XML node, several others are available:

- **<alpha>**, ***float, optional field***, specifies a constant that multiplies the penalty terms. $alpha = 0$ is equivalent to an ordinary least square, solved by the **LinearRegression** object.
  Default: 1.0

- **<l1_ratio>**, ***float, optional field***, specifies the ElasticNet mixing parameter, with $0 <= l1_ratio <= 1$. For $l1_ratio = 0$ the penalty is an L2 penalty. For $l1_ratio = 1$ it is an L1 penalty. For $0 < l1_ratio < 1$, the penalty is a combination of L1 and L2.
  Default: 0.5

- **<fit_intercept>**, ***boolean, optional field***, determines whether to calculate the intercept for this model. If set to False, no intercept will be used in the calculations (e.g. data is expected to be already centered).
  Default: True

- **<normalize>**, ***boolean, optional field***, if True, the regressors X will be normalized before regression.
Default: False

- **<max_iter>**, ***integer, optional field***, specifies the maximum number of iterations.
Default: 1000

- **<tol>**, ***float, optional field***, specifies the tolerance for the optimization: if the updates are smaller than tol, the optimization code checks the dual gap for optimality and continues until it is smaller than tol.
Default: 1.e-4

- **<warm_start>**, ***boolean, optional field***, when set to True, the model will reuse the solution of the previous call to fit as initialization, otherwise, it will just erase the previous solution.
Default: False

- **<positive>**, ***boolean, optional field***, when set to True, this forces the coefficients to be positive.
Default: False

It is important to NOTE that RAVEN does not pre-normalize the training data before constructing the *ElasticNet* ROM.

17.3.7.1.4 Linear Model: Elastic Net CV

The *Elastic Net CV* is a linear regression similar to the Elastic Net model but with an iterative fitting along a regularization path. The best model is selected by cross-validation.

In order to use the *Elastic Net CV regressor*, the user needs to set the sub-node:

<SKLtype>`linear_model|ElasticNetCV`**</SKLtype>**.

In addition to this XML node, several others are available:

- **<l1_ratio>**, ***float, optional field***, Float flag between 0 and 1 passed to ElasticNet (scaling between l1 and l2 penalties). For $l1_ratio = 0$ the penalty is an L2 penalty. For $l1_ratio = 1$ it is an L1 penalty. For $0 < l1_ratio < 1$, the penalty is a combination of L1 and L2 This parameter can be a list, in which case the different values are tested by cross-validation and the one giving the best prediction score is used. Note that a good choice of list of values for $l1_ratio$ is often to put more values close to 1 (i.e. Lasso) and less close to 0 (i.e. Ridge), as in [.1, .5, .7, .9, .95, .99, 1].
Default: 0.5

- **<eps>**, ***float, optional field***, specifies the length of the path. eps=1e-3 means that $alpha_min/alpha_max = 1e-3$.
  Default: 0.001

- **<n_alphas>**, ***integer, optional field***, is the number of alphas along the regularization path used for each $l1_ratio$.
  Default: 100

- **<precompute>**, ***boolean or string, optional field***, determines whether to use a precomputed Gram matrix to speed up calculations. If set to 'auto,' RAVEN will decide. The Gram matrix can also be passed as an argument. Available options are [True — False — 'auto' — array-like].
  Default: 'auto'

- **<max_iter>**, ***integer, optional field***, specifies the maximum number of iterations.
  Default: 1000

- **<tol>**, ***float, optional field***, specifies the tolerance for the optimization: if the updates are smaller than tol, the optimization code checks the dual gap for optimality and continues until it is smaller than tol.
  Default: 1.e-4

- **<positive>**, ***boolean, optional field***, when set to True, this forces the coefficients to be positive.
  Default: False

It is important to NOTE that RAVEN does not pre-normalize the training data before constructing the *ElasticNetCV* ROM.

17.3.7.1.5 Linear Model: Least Angle Regression model

The *Least Angle Regression model* (LARS) is a regression algorithm for high-dimensional data. The LARS algorithm provides a means of producing an estimate of which variables to include, as well as their coefficients, when a response variable is determined by a linear combination of a subset of potential covariates.

In order to use the *Least Angle Regression model*, the user needs to set the sub-node:

<SKLtype>`linear_model|Lars`**</SKLtype>**.

In addition to this XML node, several others are available:

- **<n_nonzero_coefs>**, ***integer, optional field***, represents the target number of non-zero coefficients.
  Default: 500

- **<fit_intercept>**, ***boolean, optional field***, determines whether to calculate the intercept for this model. If set to False, no intercept will be used in the calculations (e.g. data is expected to be already centered).
  Default: True

- **<verbose>**, ***boolean, optional field***, use verbose mode when fitting the model.
  Default: False

- **<precompute>**, ***boolean or string, optional field***, determines whether to use a precomputed Gram matrix to speed up calculations. If set to 'auto,' RAVEN will decide. The Gram matrix can also be passed as an argument. Available options are [True — False — 'auto' — array-like].
  Default: 'auto'

- **<normalize>**, ***boolean, optional field***, if True, the regressors X will be normalized before regression.
  Default: True

- **<eps>**,***float, optional field***, represents the machine precision regularization in the computation of the Cholesky diagonal factors. Increase this for very ill-conditioned systems. Unlike the **<tol>** parameter in some iterative optimization-based algorithms, this parameter does not control the tolerance of the optimization.
  Default: 2.220446049250313e-16

- **<fit_path>**, ***boolean, optional field***, if True the full path is stored in the coef_path_attribute. If you compute the solution for a large problem or many targets, setting fit_path to False will lead to a speedup, especially with a small alpha.
  Default: True

It is important to NOTE that RAVEN does not pre-normalize the training data before constructing the *Lars* ROM.

17.3.7.1.6 Linear Model: Cross-validated Least Angle Regression model

The *Cross-validated Least Angle Regression model* is a regression algorithm for high-dimensional data. It is similar to the LARS method, but the best model is selected by cross-validation. In order to use the *Cross-validated Least Angle Regression model*, the user needs to set the sub-node:

<SKLtype>`linear_model|LarsCV`**</SKLtype>**.

In addition to this XML node, several others are available:

- **<fit_intercept>**, ***boolean, optional field***, determines whether to calculate the intercept for this model. If set to False, no intercept will be used in the calculations (e.g. data is

expected to be already centered).
Default: True

- **<verbose>**, ***boolean, optional field***, use verbose mode when fitting the model.
  Default: False

- **<normalize>**, ***boolean, optional field***, if True, the regressors X will be normalized before regression.
  Default: True

- **<precompute>**, ***boolean or string, optional field***, determines whether to use a precomputed Gram matrix to speed up calculations. If set to 'auto,' RAVEN will decide. The Gram matrix can also be passed as an argument. Available options are [True — False — 'auto' — array-like].
  Default: 'auto'

- **<max_iter>**, ***integer, optional field***, specifies the maximum number of iterations.
  Default: 500

- **<max_n_alphas>**, ***integer, optional field***, specifies the maximum number of points on the path used to compute the residuals in the cross-validation.
  Default: 1000

- **<eps>**, ***float, optional field***, represents the machine-precision regularization in the computation of the Cholesky diagonal factors. Increase this for very ill-conditioned systems. Unlike the *tol* parameter in some iterative optimization-based algorithms, this parameter does not control the tolerance of the optimization.
  Default: 2.2204460492503131e-16

17.3.7.1.7 Linear Model trained with L1 prior as regularizer (aka the Lasso)

The *Linear Model trained with L1 prior as regularizer (Lasso)* is a shrinkage and selection method for linear regression. It minimizes the usual sum of squared errors, with a bound on the sum of the absolute values of the coefficients. In order to use the *Linear Model trained with L1 prior as regularizer (Lasso)*, the user needs to set the sub-node:

<SKLtype>`linear_model|Lasso`**</SKLtype>**.

In addition to this XML node, several others are available:

- **<alpha>**, ***float, optional field***, sets a constant multiplier for the L1 term. alpha = 0 is equivalent to an ordinary least square, solved by the LinearRegression object. For numerical reasons, using alpha = 0 with the Lasso object is not advised and you should instead use the LinearRegression object.
  Default: 1.0

- **<fit_intercept>**, ***boolean, optional field***, determines whether to calculate the intercept for this model. If set to False, no intercept will be used in the calculations (e.g. data is expected to be already centered).
  Default: True

- **<normalize>**, ***boolean, optional field***, if True, the regressors X will be normalized before regression.
  Default: False

- **<precompute>**, ***boolean or string, optional field***, determines whether to use a precomputed Gram matrix to speed up calculations. If set to 'auto,' RAVEN will decide. The Gram matrix can also be passed as an argument. Available options are [True — False — 'auto' — array-like].
  Default: False **Note:** For sparse input this option is always True to preserve sparsity.

- **<max_iter>**, ***integer, optional field***, specifies the maximum number of iterations.
  Default: 1000

- **<tol>**, ***float, optional field***, specifies the tolerance for the optimization: if the updates are smaller than tol, the optimization code checks the dual gap for optimality and continues until it is smaller than tol.
  Default: 1.e-4

- **<warm_start>**, ***boolean, optional field***, when set to True, the model will reuse the solution of the previous call to fit as initialization, otherwise, it will just erase the previous solution.
  Default: False

- **<positive>**, ***boolean, optional field***, when set to True, this forces the coefficients to be positive.
  Default: False

It is important to NOTE that RAVEN does not pre-normalize the training data before constructing the *LarsCV* ROM.

17.3.7.1.8 Lasso linear model with iterative fitting along a regularization path

The *Lasso linear model with iterative fitting along a regularization path* is an algorithm of the Lasso family, that computes the linear regressor weights, identifying the regularization path in an iterative fitting (see http://www.jstatsoft.org/v33/i01/paper)

In order to use the *Lasso linear model with iterative fitting along a regularization path regressor*, the user needs to set the sub-node:

<SKLtype>linear_model|LassoCV**</SKLtype>**.

In addition to this XML node, several others are available:

- **<eps>**, ***float, optional field***, represents the length of the path. eps=1e-3 means that alpha_min / alpha_max = 1e-3.
  Default: 1.0e-3

- **<n_alphas>**, ***int, optional field***, sets the number of alphas along the regularization path.
  Default: 100

- **<alphas>**, ***numpy array, optional field***, lists the locations of the alphas used to compute the models.
  Default: None If None, alphas are set automatically.

- **<precompute>**, ***boolean or string, optional field***, determines whether to use a precomputed Gram matrix to speed up calculations. If set to 'auto,' RAVEN will decide. The Gram matrix can also be passed as an argument. Available options are [True — False — 'auto' — array-like].
  Default: 'auto'

- **<max_iter>**, ***integer, optional field***, specifies the maximum number of iterations.
  Default: 1000

- **<tol>**, ***float, optional field***, specifies the tolerance for the optimization: if the updates are smaller than tol, the optimization code checks the dual gap for optimality and continues until it is smaller than tol.
  Default: 1.e-4

- **<verbose>**, ***boolean or integer, optional field***, use verbose mode when fitting the model.
  Default: False

- **<positive>**, ***boolean, optional field***, when set to True, this forces the coefficients to be positive.
  Default: False

It is important to NOTE that RAVEN does not pre-normalize the training data before constructing the *LassoCV* ROM.

17.3.7.1.9 Lasso model fit with Least Angle Regression

Lasso model fit with Least Angle Regression (aka Lars) It is a Linear Model trained with an L1 prior as regularizer. In order to use the *Least Angle Regression model regressor*, the user needs to

set the sub-node In order to use the *Least Angle Regression model regressor*, the user needs to set the sub-node:

<SKLtype>linear_model|LassoLars**</SKLtype>**.

In addition to this XML node, several others are available:

- **<alpha>**, ***float, optional field***, specifies a constant that multiplies the penalty terms. $alpha = 0$ is equivalent to an ordinary least square, solved by the **LinearRegression** object.
  Default: 1.0
- **<fit_intercept>**, ***boolean, optional field***, determines whether to calculate the intercept for this model. If set to False, no intercept will be used in the calculations (e.g. data is expected to be already centered).
  Default: True
- **<verbose>**, ***boolean or integer, optional field***, use verbose mode when fitting the model.
  Default: False
- **<normalize>**, ***boolean, optional field***, if True, the regressors X will be normalized before regression.
  Default: True
- **<precompute>**, ***boolean or string, optional field***, determines whether to use a precomputed Gram matrix to speed up calculations. If set to 'auto,' RAVEN will decide. The Gram matrix can also be passed as an argument. Available options are [True — False — 'auto' — array-like].
  Default: 'auto'
- **<max_iter>**, ***integer, optional field***, specifies the maximum number of iterations.
  Default: 500
- **<eps>**, ***float, optional field***, sets the machine precision regularization in the computation of the Cholesky diagonal factors. Increase this for very ill-conditioned systems.
  Default: 2.220446049250313e-16

It is important to NOTE that RAVEN does not pre-normalize the training data before constructing the *LassoLars* ROM.

17.3.7.1.10 Cross-validated Lasso, using the LARS algorithm

The *Cross-validated Lasso, using the LARS algorithm* is a cross-validated Lasso, using the LARS algorithm.

In order to use the *Cross-validated Lasso, using the LARS algorithm regressor*, the user needs to set the sub-node:

<SKLtype>linear_model|LassoLarsCV**</SKLtype>**.

In addition to this XML node, several others are available:

- **<fit_intercept>**, ***boolean, optional field***, determines whether to calculate the intercept for this model. If set to False, no intercept will be used in the calculations (e.g. data is expected to be already centered).
Default: True

- **<verbose>**, ***boolean or integer, optional field***, use verbose mode when fitting the model.
Default: False

- **<normalize>**, ***boolean, optional field***, if True, the regressors X will be normalized before regression.
Default: True

- **<precompute>**, ***boolean or string, optional field***, determines whether to use a precomputed Gram matrix to speed up calculations. If set to 'auto,' RAVEN will decide. The Gram matrix can also be passed as an argument. Available options are [True — False — 'auto' — array-like].
Default: 'auto'

- **<max_iter>**, ***integer, optional field***, specifies the maximum number of iterations.
Default: 500

- **<max_n_alphas>**, ***integer, optional field***, specifies the maximum number of points on the path used to compute the residuals in the cross-validation.
Default: 1000

- **<eps>**, ***float, optional field***, specifies the machine precision regularization in the computation of the Cholesky diagonal factors. Increase this for very ill-conditioned systems.
Default: 2.220446049250313e-16

It is important to NOTE that RAVEN does not pre-normalize the training data before constructing the *LassoLarsCV* ROM.

17.3.7.1.11 Lasso model fit with Lars using BIC or AIC for model selection

The *Lasso model fit with Lars using BIC or AIC for model selection* is a Lasso model fit with Lars using BIC or AIC for model selection. The optimization objective for Lasso is: $(1/(2 *$

$n_samples)) * ||y - Xw||_2^2 + alpha * ||w||_1$ AIC is the Akaike information criterion and BIC is the Bayes information criterion. Such criteria are useful in selecting the value of the regularization parameter by making a trade-off between the goodness of fit and the complexity of the model. A good model explains the data well while maintaining simplicity. In order to use the *Lasso model fit with Lars using BIC or AIC for model selection regressor*, the user needs to set the sub-node:

<SKLtype>`linear_model|LassoLarsIC`**</SKLtype>**.

In addition to this XML node, several others are available:

- **<criterion>**, ***'bic' — 'aic'*** , specifies the type of criterion to use.
  Default: 'aic'
- **<fit_intercept>**, ***boolean, optional field***, determines whether to calculate the intercept for this model. If set to False, no intercept will be used in the calculations (e.g. data is expected to be already centered).
  Default: True
- **<verbose>**, ***boolean or integer, optional field***, use verbose mode when fitting the model.
  Default: False
- **<normalize>**, ***boolean, optional field***, if True, the regressors X will be normalized before regression.
  Default: True
- **<precompute>**, ***boolean or string, optional field***, determines whether to use a precomputed Gram matrix to speed up calculations. If set to 'auto,' RAVEN will decide. The Gram matrix can also be passed as an argument. Available options are [True — False — 'auto' — array-like].
  Default: 'auto'
- **<max_iter>**, ***integer, optional field***, specifies the maximum number of iterations.
  Default: 500
- **<eps>**,***float, optional field***, represents the machine precision regularization in the computation of the Cholesky diagonal factors. Increase this for very ill-conditioned systems. Unlike the tol parameter in some iterative optimization-based algorithms, this parameter does not control the tolerance of the optimization.
  Default: 2.220446049250313e-16

It is important to NOTE that RAVEN does not pre-normalize the training data before constructing the *LassoLarsIC* ROM.

17.3.7.1.12 Ordinary least squares Linear Regression

The *Ordinary least squares Linear Regression* is a method for estimating the unknown parameters in a linear regression model, with the goal of minimizing the differences between the observed responses in some arbitrary dataset and the responses predicted by the linear approximation of the data. In order to use the *Ordinary least squares Linear Regressor*, the user needs to set the sub-node:

<SKLtype>`linear_model|LinearRegression`**</SKLtype>**.

In addition to this XML node, several others are available:

- **<fit_intercept>**, ***boolean, optional field***, determines whether to calculate the intercept for this model. If set to False, no intercept will be used in the calculations (e.g. data is expected to be already centered).
  Default: True
- **<normalize>**, ***boolean, optional field***, if True, the regressors X will be normalized before regression.
  Default: False

It is important to NOTE that RAVEN does not pre-normalize the training data before constructing the *LinearRegression* ROM.

17.3.7.1.13 Logistic Regression

The *Logistic Regression* implements L1 and L2 regularized logistic regression using the liblinear library. It can handle both dense and sparse input. This regressor uses C-ordered arrays or CSR matrices containing 64-bit floats for optimal performance; any other input format will be converted (and copied). In order to use the *Logistic Regressor*, the user needs to set the sub-node:

<SKLtype>`linear_model|LogisticRegression`**</SKLtype>**.

In addition to this XML node, several others are available:

- **<penalty>**, ***string, 'l1' or 'l2'***, specifies the norm used in the penalization.
  Default: 'l2'
- **<dual>**, ***boolean***, specifies the dual or primal formulation. Dual formulation is only implemented for the l2 penalty. Prefer dual=False when n_samples > n_features.
  Default: False

- **<C>**, ***float, optional field***, is the inverse of the regularization strength; must be a positive float. Like in support vector machines, smaller values specify stronger regularization.
  Default: 1.0

- **<fit_intercept>**, ***boolean***, specifies if a constant (a.k.a. bias or intercept) should be added to the decision function.
  Default: True

- **<intercept_scaling>**, ***float, optional field***, when self.fit_intercept is True, instance vector x becomes [x, self.intercept_scaling], i.e. a "synthetic" feature with constant value equal to intercept_scaling is appended to the instance vector. The intercept becomes intercept_scaling * synthetic feature weight. **Note:** The synthetic feature weight is subject to l1/l2 regularization as are all other features. To lessen the effect of regularization on synthetic feature weight (and therefore on the intercept) intercept_scaling has to be increased.
  Default: 1.0

- **<class_weight>**, ***dict, or 'balanced', optional*** Weights associated with classes in the form {class_label: weight}. If not given, all classes are supposed to have weight one. The "balanced" mode uses the values of y to automatically adjust weights inversely proportional to class frequencies in the input data as n_samples / (n_classes * np.bincount(y)) Note that these weights will be multiplied with sample_weight (passed through the fit method) if sample_weight is specified. New in version 0.17: class_weight='balanced' instead of deprecated class_weight='auto'.
  Default: None

- **<random_state>**, ***int seed, RandomState instance, or None***, sets the seed of the pseudo random number generator to use when shuffling the data.
  Default: None

- **<tol>**, ***float, optional field***, specifies the tolerance for stopping criteria.
  Default: 0.0001

It is important to NOTE that RAVEN uses a Z-score normalization of the training data before constructing the *LogisticRegression* ROM:

$$\mathbf{X}' = \frac{(\mathbf{X} - \mu)}{\sigma} \tag{8}$$

17.3.7.1.14 Multi-task Lasso model trained with L1/L2 mixed-norm as regularizer

The *Multi-task Lasso model trained with L1/L2 mixed-norm as regularizer* is a regressor where the optimization objective for Lasso is: $(1/(2*n_samples))*||Y-XW||^2_{Fro}+alpha*||W||_{21}$ Where:

$||W||_{21} = \sum_i \sqrt{\sum_j w_{ij}^2}$ i.e. the sum of norm of each row. In order to use the *Multi-task Lasso model trained with L1/L2 mixed-norm as regularizer regressor*, the user needs to set the sub-node:

<SKLtype>linear_model|MultiTaskLasso**</SKLtype>**.

In addition to this XML node, several others are available:

- **<alpha>**, ***float, optional field***, sets the constant multiplier for the L1/L2 term.
  Default: 1.0
- **<fit_intercept>**, ***boolean, optional field***, determines whether to calculate the intercept for this model. If set to False, no intercept will be used in the calculations (e.g. data is expected to be already centered).
  Default: True
- **<normalize>**, ***boolean, optional field***, if True, the regressors X will be normalized before regression.
  Default: False
- **<max_iter>**, ***integer, optional field***, specifies the maximum number of iterations.
  Default: 1000
- **<tol>**, ***float, optional field***, specifies the tolerance for the optimization: if the updates are smaller than tol, the optimization code checks the dual gap for optimality and continues until it is smaller than tol.
  Default: 1.e-4
- **<warm_start>**, ***boolean, optional field***, when set to True, the model will reuse the solution of the previous call to fit as initialization, otherwise, it will just erase the previous solution.
  Default: False

It is important to NOTE that RAVEN does not pre-normalize the training data before constructing the *MultiTaskLasso* ROM.

17.3.7.1.15 Multi-task Elastic Net model trained with L1/L2 mixed-norm as regularizer

The *Multi-task Elastic Net model trained with L1/L2 mixed-norm as regularizer* is a regressor where the optimization objective for MultiTaskElasticNet is: $(1/(2 * n_samples)) * ||Y - XW||_2^{Fro} + alpha * l1_ratio * ||W||_{21} + 0.5 * alpha * (1 - l1_ratio) * ||W||_{Fro}^2$ Where: $||W||_{21} = \sum_i \sqrt{\sum_j w_{ij}^2}$ i.e. the sum of norm of each row. In order to use the *Multi-task ElasticNet model trained with L1/L2 mixed-norm as regularizer regressor*, the user needs to set the sub-node:

<SKLtype>linear_model|MultiTaskElasticNet</SKLtype>.

In addition to this XML node, several others are available:

- **<alpha>**, ***float, optional field***, represents a constant multiplier for the L1/L2 term.
  Default: 1.0
- **<l1_ratio>**, ***float***, represents the Elastic Net mixing parameter, with $0 < l1_ratio \leq 1$. For $l1_ratio = 0$ the penalty is an L1/L2 penalty. For $l1_ratio = 1$ it is an L1 penalty. For $0 < l1_ratio < 1$, the penalty is a combination of L1/L2 and L2.
  Default: 0.5
- **<fit_intercept>**, ***boolean, optional field***, determines whether to calculate the intercept for this model. If set to False, no intercept will be used in the calculations (e.g. data is expected to be already centered).
  Default: True
- **<normalize>**, ***boolean, optional field***, if True, the regressors X will be normalized before regression.
  Default: False
- **<max_iter>**, ***integer, optional field***, specifies the maximum number of iterations.
- **<tol>**, ***float, optional field***, specifies the tolerance for the optimization: if the updates are smaller than tol, the optimization code checks the dual gap for optimality and continues until it is smaller than tol.
  Default: 1.e-4
- **<warm_start>**, ***boolean, optional field***, when set to True, the model will reuse the solution of the previous call to fit as initialization, otherwise, it will just erase the previous solution.
  Default: False

It is important to NOTE that RAVEN does not pre-normalize the training data before constructing the *MultiTaskElasticNet* ROM.

17.3.7.1.16 Orthogonal Mathching Pursuit model (OMP)

The *Orthogonal Mathching Pursuit model (OMP)* is a type of sparse approximation which involves finding the "best matching" projections of multidimensional data onto an over-complete dictionary, D. In order to use the *Orthogonal Mathching Pursuit model (OMP) regressor*, the user needs to set the sub-node:

<SKLtype>linear_model|OrthogonalMatchingPursuit</SKLtype>.

In addition to this XML node, several others are available:

- **<n_nonzero_coefs>**, ***int, optional field***, represents the desired number of non-zero entries in the solution. If None, this value is set to 10% of n_features.
  Default: None

- **<tol>**, ***float, optional field***, specifies the maximum norm of the residual. If not None, overrides n_nonzero_coefs.
  Default: None

- **<fit_intercept>**, ***boolean, optional field***, determines whether to calculate the intercept for this model. If set to False, no intercept will be used in the calculations (e.g. data is expected to be already centered).
  Default: True

- **<normalize>**, ***boolean, optional field***, if True, the regressors X will be normalized before regression.
  Default: True

- **<precompute>**, {***True, False, 'auto'***}, specifies whether to use a precomputed Gram and Xy matrix to speed up calculations. Improves performance when n_targets or n_samples is very large. **Note:** If you already have such matrices, you can pass them directly to the fit method.
  Default: 'auto'

It is important to NOTE that RAVEN does not pre-normalize the training data before constructing the *OrthogonalMatchingPursuit* ROM.

17.3.7.1.17 Cross-validated Orthogonal Mathching Pursuit model (OMP)

The *Cross-validated Orthogonal Mathching Pursuit model (OMP)* is a regressor similar to OMP which has good performance in sparse recovery. In order to use the *Cross-validated Orthogonal Mathching Pursuit model (OMP) regressor*, the user needs to set the sub-node:

<SKLtype>`linear_model|OrthogonalMatchingPursuitCV`**</SKLtype>**.

In addition to this XML node, several others are available:

- **<fit_intercept>**, ***boolean, optional field***, determines whether to calculate the intercept for this model. If set to False, no intercept will be used in the calculations (e.g. data is expected to be already centered).
  Default: True

- **<normalize>**, ***boolean, optional field***, if True, the regressors X will be normalized before regression.
  Default: True

- **<max_iter>**, ***integer, optional field***, specifies the maximum number of iterations.
  Default: None Maximum number of iterations to perform, therefore maximum features to include 10% of n_features but at least 5 if available.

- **<cv>**, ***cross-validation generator, optional***, see sklearn.cross_validation.
  Default: None

- **<verbose>**, ***boolean or integer, optional field***, use verbose mode when fitting the model.
  Default: False

It is important to NOTE that RAVEN does not pre-normalize the training data before constructing the *OrthogonalMatchingPursuitCV* ROM.

17.3.7.1.18 Passive Aggressive Classifier

The *Passive Aggressive Classifier* is a principled approach to linear classification that advocates minimal weight updates i.e., the least required to correctly classify the current training instance. In order to use the *Passive Aggressive Classifier*, the user needs to set the sub-node:

<SKLtype>`linear_model|PassiveAggressiveClassifier`**</SKLtype>**.

In addition to this XML node, several others are available:

- **<C>**, ***float***, specifies the maximum step size (regularization).
  Default: 1.0

- **<fit_intercept>**, ***boolean, optional field***, determines whether to calculate the intercept for this model. If set to False, no intercept will be used in the calculations (e.g. data is expected to be already centered).
  Default: True

- **<n_iter>**, ***int, optional field***, specifies the number of passes over the training data (aka epochs).
  Default: 5

- **<shuffle>**, ***boolean, optional field***, specifies whether or not the training data should be shuffled after each epoch.
  Default: True

- **<random_state>**, ***int seed, RandomState instance, or None***, sets the seed of the pseudo random number generator to use when shuffling the data.
  Default: None

- **<verbose>**, ***boolean or integer, optional field***, use verbose mode when fitting the model.
  Default: 0

- **<loss>**, ***string, optional field***, the loss function to be used:
  - hinge: equivalent to PA-I (http://jmlr.csail.mit.edu/papers/volume7/crammer06a/crammer06a.pdf)
  - squared_hinge: equivalent to PA-II (http://jmlr.csail.mit.edu/papers/volume7/crammer06a/crammer06a

  Default: 'hinge'

- **<warm_start>**, ***boolean, optional field***, when set to True, the model will reuse the solution of the previous call to fit as initialization, otherwise, it will just erase the previous solution.
  Default: False

It is important to NOTE that RAVEN uses a Z-score normalization of the training data before constructing the *PassiveAggressiveClassifier* ROM:

$$\mathbf{X'} = \frac{(\mathbf{X} - \mu)}{\sigma} \tag{9}$$

17.3.7.1.19 Passive Aggressive Regressor

The *Passive Aggressive Regressor* is similar to the Perceptron in that it does not require a learning rate. However, contrary to the Perceptron, this regressor includes a regularization parameter, C.

In order to use the *Passive Aggressive Regressor*, the user needs to set the sub-node:

<SKLtype>`linear_model|PassiveAggressiveRegressor`**</SKLtype>**.

In addition to this XML node, several others are available:

- **<C>**, ***float***, sets the maximum step size (regularization).
  Default: 1.0

- **<epsilon>**, ***float***, if the difference between the current prediction and the correct label is below this threshold, the model is not updated.
  Default: 0.1

- **<fit_intercept>**, ***boolean, optional field***, determines whether to calculate the intercept for this model. If set to False, no intercept will be used in the calculations (e.g. data is expected to be already centered).
  Default: True

- **<n_iter>**, ***int, optional field***, specifies the number of passes over the training data (aka epochs).
  Default: 5

- **<shuffle>**, ***boolean, optional field***, specifies whether or not the training data should be shuffled after each epoch.
  Default: True

- **<random_state>**, ***int seed, RandomState instance, or None***, sets the seed of the pseudo random number generator to use when shuffling the data.
  Default: None

- **<verbose>**, ***boolean or integer, optional field***, use verbose mode when fitting the model.
  Default: 0

- **<loss>**, ***string, optional field***, specifies the loss function to be used:
  - epsilon_insensitive: equivalent to PA-I in the reference paper (http://jmlr.csail.mit.edu/papers/volume7 /crammer06a/crammer06a.pdf).
  - squared_epsilon_insensitive: equivalent to PA-II in the reference paper (http://jmlr.csail.mit.edu/papers /volume7/crammer06a/crammer06a.pdf).

  Default: 'epsilon_insensitive'

- **<warm_start>**, ***boolean, optional field***, when set to True, the model will reuse the solution of the previous call to fit as initialization, otherwise, it will just erase the previous solution.
  Default: False

It is important to NOTE that RAVEN uses a Z-score normalization of the training data before constructing the *PassiveAggressiveRegressor* ROM:

$$\mathbf{X}' = \frac{(\mathbf{X} - \mu)}{\sigma} \tag{10}$$

17.3.7.1.20 Perceptron

The *Perceptron* method is an algorithm for supervised classification of an input into one of several possible non-binary outputs. It is a type of linear classifier, i.e. a classification algorithm that makes its predictions based on a linear predictor function combining a set of weights with the feature vector. The algorithm allows for online learning, in that it processes elements in the training set one at a time. In order to use the *Perceptron classifier*, the user needs to set the sub-node:

<SKLtype>linear_model|Perceptron**</SKLtype>**.

In addition to this XML node, several others are available:

- **<penalty>**, ***None, 'l2' or 'l1' or 'elasticnet'***, defines the penalty (aka regularization term) to be used.
  Default: None

- **<alpha>**, ***float***, sets the constant multiplier for the regularization term if regularization is used.
  Default: 0.0001

- **<fit_intercept>**, ***boolean, optional field***, determines whether to calculate the intercept for this model. If set to False, no intercept will be used in the calculations (e.g. data is expected to be already centered).
  Default: True

- **<n_iter>**, ***int, optional field***, specifies the number of passes over the training data (aka epochs).
  Default: 5

- **<shuffle>**, ***boolean, optional field***, specifies whether or not the training data should be shuffled after each epoch.
  Default: True

- **<random_state>**, ***int seed, RandomState instance, or None***, sets the seed of the pseudo random number generator to use when shuffling the data.
  Default: 0

- **<verbose>**, ***boolean or integer, optional field***, use verbose mode when fitting the model.
  Default: 0

- **<eta0>**, ***double, optional field***, defines the constant multiplier for the updates.
  Default: 1.0

- **<class_weight>**, ***dict, {class_label: weight} or "balanced" or None, optional*** Preset for the class_weight fit parameter. Weights associated with classes. If not given, all classes are supposed to have weight one. The "balanced" mode uses the values of y to automatically

adjust weights inversely proportional to class frequencies in the input data as n_samples / (n_classes * np.bincount(y))

- **<warm_start>**, ***boolean, optional field***, when set to True, the model will reuse the solution of the previous call to fit as initialization, otherwise, it will just erase the previous solution.
  Default: False

It is important to NOTE that RAVEN uses a Z-score normalization of the training data before constructing the *PassiveAggressiveRegressor* ROM:

$$\mathbf{X}' = \frac{(\mathbf{X} - \mu)}{\sigma} \tag{11}$$

17.3.7.1.21 Linear least squares with l2 regularization

The *Linear least squares with l2 regularization* solves a regression model where the loss function is the linear least squares function and the regularization is given by the l2-norm. Also known as Ridge Regression or Tikhonov regularization. This estimator has built-in support for multivariate regression (i.e., when y is a 2d-array of shape [n_samples, n_targets]). In order to use the *Linear least squares with l2 regularization*, the user needs to set the sub-node:

<SKLtype>linear_model|Ridge**</SKLtype>**.

In addition to this XML node, several others are available:

- **<alpha>**, ***float, array-like***, shape = [n_targets] Small positive values of alpha improve the conditioning of the problem and reduce the variance of the estimates. Alpha corresponds to $(2 * C)^{-}1$ in other linear models such as LogisticRegression or LinearSVC. If an array is passed, penalties are assumed to be specific to the targets. Hence they must correspond in number.
  Default: 1.0

- **<fit_intercept>**, ***boolean, optional field***, determines whether to calculate the intercept for this model. If set to False, no intercept will be used in the calculations (e.g. data is expected to be already centered).
  Default: True

- **<max_iter>**, ***integer, optional field***, specifies the maximum number of iterations.
  Default: determined by scipy.sparse.linalg.

- **<normalize>**, ***boolean, optional field***, if True, the regressors X will be normalized before regression.
  Default: False

- **<solver>**, {***'auto', 'svd', 'cholesky', 'lsqr', 'sparse_cg'***}, specifies the solver to use in the computational routines:
  - 'auto' chooses the solver automatically based on the type of data.
  - 'svd' uses a singular value decomposition of X to compute the ridge coefficients. More stable for singular matrices than 'cholesky.'
  - 'cholesky' uses the standard scipy.linalg.solve function to obtain a closed-form solution.
  - 'sparse_cg' uses the conjugate gradient solver as found in scipy.sparse.linalg.cg. As an iterative algorithm, this solver is more appropriate than 'cholesky' for large-scale data (possibility to set tol and max_iter).
  - 'lsqr' uses the dedicated regularized least-squares routine scipy.sparse.linalg.lsqr. It is the fatest but may not be available in old scipy versions. It also uses an iterative procedure.

  All three solvers support both dense and sparse data.
  Default: 'auto'

It is important to NOTE that RAVEN does not pre-normalize the training data before constructing the *Ridge* ROM.

17.3.7.1.22 Classifier using Ridge regression

The *Classifier using Ridge regression* is a classifier based on linear least squares with l2 regularization. In order to use the *Classifier using Ridge regression*, the user needs to set the sub-node:

<SKLtype>linear_model|RidgeClassifier**</SKLtype>**.

In addition to this XML node, several others are available:

- **<alpha>**, ***float***, small positive values of alpha improve the conditioning of the problem and reduce the variance of the estimates. Alpha corresponds to $(2 * C)^-1$ in other linear models such as LogisticRegression or LinearSVC.
  Default: 1.0
- **<class_weight>**, ***dict, optional field***, specifies weights associated with classes in the form class_label: weight. If not given, all classes are assumed to have weight one.
  Default: None
- **<fit_intercept>**, ***boolean, optional field***, determines whether to calculate the intercept for this model. If set to False, no intercept will be used in the calculations (e.g. data is

expected to be already centered).
Default: True

- **<max_iter>**, ***integer, optional field***, specifies the maximum number of iterations.
  Default: determined by scipy.sparse.linalg.

- **<normalize>**, ***boolean, optional field***, if True, the regressors X will be normalized before regression.
  Default: False

- **<solver>**, {***'auto', 'svd', 'cholesky', 'lsqr', 'sparse_cg'***}, specifies the solver to use in the computational routines:
  - 'auto' chooses the solver automatically based on the type of data.
  - 'svd' uses a singular value decomposition of X to compute the ridge coefficients. More stable for singular matrices than 'cholesky.'
  - 'cholesky' uses the standard scipy.linalg.solve function to obtain a closed-form solution.
  - 'sparse_cg' uses the conjugate gradient solver as found in scipy.sparse.linalg.cg. As an iterative algorithm, this solver is more appropriate than 'cholesky' for large-scale data (possibility to set tol and max_iter).
  - 'lsqr' uses the dedicated regularized least-squares routine scipy.sparse.linalg.lsqr. It is the fatest but may not be available in old scipy versions. It also uses an iterative procedure.

  All three solvers support both dense and sparse data.
  Default: 'auto'

- **<tol>**, ***float***, defines the required precision of the solution.
  Default: 0.001

It is important to NOTE that RAVEN does not pre-normalize the training data before constructing the *RidgeClassifier* ROM.

17.3.7.1.23 Ridge classifier with built-in cross-validation

The *Ridge classifier with built-in cross-validation* performs Generalized Cross-Validation, which is a form of efficient leave-one-out cross-validation. Currently, only the n_features > n_samples case is handled efficiently. In order to use the *Ridge classifier with built-in cross-validation classifier*, the user needs to set the sub-node:

<SKLtype>`linear_model|RidgeClassifierCV`**</SKLtype>**.

In addition to this XML node, several others are available:

- **<alphas>**, ***numpy array of shape [n_alphas]***, is an array of alpha values to try. Small positive values of alpha improve the conditioning of the problem and reduce the variance of the estimates. Alpha corresponds to $(2*C)^{-1}$ in other linear models such as LogisticRegression or LinearSVC.
  Default: (0.1, 1.0, 10.0)

- **<fit_intercept>**, ***boolean, optional field***, determines whether to calculate the intercept for this model. If set to False, no intercept will be used in the calculations (e.g. data is expected to be already centered).
  Default: True

- **<normalize>**, ***boolean, optional field***, if True, the regressors X will be normalized before regression.
  Default: False

- **<scoring>**, ***string, callable or None, optional***, is a string (see model evaluation documentation) or a scorer callable object / function with signature scorer(estimator, X, y).
  Default: None

- **<cv>**, ***cross-validation generator, optional***, If None, Generalized Cross-Validation (efficient leave-one-out) will be used.
  Default: None

- **<class_weight>**, ***dic, optional field***, specifies weights associated with classes in the form class_label:weight. If not given, all classes are supposed to have weight one.
  Default: None

It is important to NOTE that RAVEN does not pre-normalize the training data before constructing the *RidgeClassifierCV* ROM.

17.3.7.1.24 Ridge regression with built-in cross-validation

The *Ridge regression with built-in cross-validation* performs Generalized Cross-Validation, which is a form of efficient leave-one-out cross-validation. In order to use the *Ridge regression with built-in cross-validation regressor*, the user needs to set the sub-node:

<SKLtype>`linear_model|RidgeCV`**</SKLtype>**.

In addition to this XML node, several others are available:

- **<alphas>**, ***numpy array of shape [n_alphas]***, specifies an array of alpha values to try. Small positive values of alpha improve the conditioning of the problem and reduce the vari-

ance of the estimates. Alpha corresponds to $(2 * C)^{-1}$ in other linear models such as LogisticRegression or LinearSVC.
Default: (0.1, 1.0, 10.0)

- **<fit_intercept>**, ***boolean, optional field***, determines whether to calculate the intercept for this model. If set to False, no intercept will be used in the calculations (e.g. data is expected to be already centered).
Default: True

- **<normalize>**, ***boolean, optional field***, if True, the regressors X will be normalized before regression.
Default: False

- **<scoring>**, ***string, callable or None, optional***, is a string (see model evaluation documentation) or a scorer callable object / function with signature scorer(estimator, X, y).
Default: None

- **<cv>**, ***cross-validation generator, optional field***, if None, Generalized Cross-Validation (efficient leave-one-out) will be used.
Default: None

- **<gcv_mode>**, ***{None, 'auto,' 'svd,' 'eigen'}, optional field***, is a flag indicating which strategy to use when performing Generalized Cross-Validation. Options are:

  - 'auto:' use svd if n_samples ¿ n_features or when X is a sparse matrix, otherwise use eigen
  - 'svd:' force computation via singular value decomposition of X (does not work for sparse matrices)
  - 'eigen:' force computation via eigendecomposition of $X^T X$

  The 'auto' mode is the default and is intended to pick the cheaper option of the two depending upon the shape and format of the training data.
  Default: None

- **<store_cv_values>**, ***boolean***, is a flag indicating if the cross-validation values corresponding to each alpha should be stored in the cv_values_attribute (see below). This flag is only compatible with cv=None (i.e. using Generalized Cross-Validation).
Default: False

It is important to NOTE that RAVEN does not pre-normalize the training data before constructing the *RidgeCV* ROM.

17.3.7.1.25 Linear classifiers (SVM, logistic regression, a.o.) with SGD training

The *Linear classifiers (SVM, logistic regression, a.o.) with SGD training* implements regularized linear models with stochastic gradient descent (SGD) learning: the gradient of the loss is estimated for each sample at a time and the model is updated along the way with a decreasing strength schedule (aka learning rate). SGD allows minibatch (online/out-of-core) learning, see the partial_fit method. This implementation works with data represented as dense or sparse arrays of floating point values for the features. The model it fits can be controlled with the loss parameter; by default, it fits a linear support vector machine (SVM). The regularizer is a penalty added to the loss function that shrinks model parameters towards the zero vector using either the squared Euclidean norm L2 or the absolute norm L1 or a combination of both (Elastic Net). If the parameter update crosses the 0.0 value because of the regularizer, the update is truncated to 0.0 to allow for learning sparse models and achieves online feature selection. In order to use the *Linear classifiers (SVM, logistic regression, a.o.) with SGD training*, the user needs to set the sub-node:

`<SKLtype>linear_model|SGDClassifier</SKLtype>`.

In addition to this XML node, several others are available:

- **<loss>**, ***str, 'hinge,' 'log,' 'modified_huber,' 'squared_hinge,' 'perceptron,' or a regression loss: 'squared_loss,' 'huber,' 'epsilon_insensitive,' or 'squared_epsilon_insensitive'***, dictates the loss function to be used. The available options are:
  - 'hinge' gives a linear SVM.
  - 'log' loss gives logistic regression, a probabilistic classifier.
  - 'modified_huber' is another smooth loss that brings tolerance to outliers as well as probability estimates.
  - 'squared_hinge' is like hinge but is quadratically penalized.
  - 'perceptron' is the linear loss used by the perceptron algorithm.

  The other losses are designed for regression but can be useful in classification as well; see SGDRegressor for a description.
  Default: 'hinge'

- **<penalty>**, ***str, 'l2' or 'l1' or 'elasticnet'***, defines the penalty (aka regularization term) to be used. 'l2' is the standard regularizer for linear SVM models. 'l1' and 'elasticnet' might bring sparsity to the model (feature selection) not achievable with 'l2.'
  Default: 'l2'

- **<alpha>**, ***float***, is the constant multiplier for the regularization term.
  Default: 0.0001

- **<l1_ratio>**, ***float***, is the Elastic Net mixing parameter, with 0 ¡= l1_ratio ¡= 1. l1_ratio=0 corresponds to L2 penalty, l1_ratio=1 to L1.
  Default: 0.15

- **<fit_intercept>**, ***boolean, optional field***, determines whether to calculate the intercept for this model. If set to False, no intercept will be used in the calculations (e.g. data is expected to be already centered).
  Default: True

- **<n_iter>**, ***int, optional field***, specifies the number of passes over the training data (aka epochs).
  Default: 5

- **<shuffle>**, ***boolean, optional field***, specifies whether or not the training data should be shuffled after each epoch.
  Default: True

- **<random_state>**, ***int seed, RandomState instance, or None***, represents the seed of the pseudo random number generator to use when shuffling the data for probability estimation.
  Default: None

- **<verbose>**, ***boolean or integer, optional field***, use verbose mode when fitting the model.
  Default: 0

- **<epsilon>**, ***float, optional field***, varies meaning depending on the value of **<loss>**. If loss is 'huber', 'epsilon_insensitive' or 'squared_epsilon_insensitive' then this is the epsilon in the epsilon-insensitive loss functions. For 'huber', determines the threshold at which it becomes less important to get the prediction exactly right. For 'epsilon_insensitive, any differences between the current prediction and the correct label are ignored if they are less than this threshold.
  Default: 0.1

- **<learning_rate>**, ***string, optional field***, specifies the learning rate:
  - 'constant:' eta = eta0
  - 'optimal:' eta = 1.0 / (t + t0)
  - 'invscaling:' eta = eta0 / pow(t, power_t)

  Default: 'optimal'

- **<eta0>**, ***double***, specifies the initial learning rate for the 'constant' or 'invscaling' schedules. The default value is 0.0 as eta0 is not used by the default schedule 'optimal.'
  Default: 0.0

- **<power_t>**, ***double***, represents the exponent for the inverse scaling learning rate.
  Default: 0.5

- **<class_weight>**, ***dict, class_label***, is the preset for the class_weight fit parameter. Weights associated with classes. If not given, all classes are assumed to have weight one. The "auto" mode uses the values of y to automatically adjust weights inversely proportional to class frequencies.
  Default: None

- **<warm_start>**, ***boolean, optional field***, when set to True, the model will reuse the solution of the previous call to fit as initialization, otherwise, it will just erase the previous solution.
  Default: False

It is important to NOTE that RAVEN uses a Z-score normalization of the training data before constructing the *SGDClassifier* ROM:

$$\mathbf{X}' = \frac{(\mathbf{X} - \mu)}{\sigma} \tag{12}$$

17.3.7.1.26 Linear model fitted by minimizing a regularized empirical loss with SGD

The *Linear model fitted by minimizing a regularized empirical loss with SGD* is a model where SGD stands for Stochastic Gradient Descent: the gradient of the loss is estimated each sample at a time and the model is updated along the way with a decreasing strength schedule (aka learning rate). The regularizer is a penalty added to the loss function that shrinks model parameters towards the zero vector using either the squared euclidean norm L2 or the absolute norm L1 or a combination of both (Elastic Net). If the parameter update crosses the 0.0 value because of the regularizer, the update is truncated to 0.0 to allow for learning sparse models and achieving online feature selection. This implementation works with data represented as dense numpy arrays of floating point values for the features. In order to use the *Linear model fitted by minimizing a regularized empirical loss with SGD*, the user needs to set the sub-node:

<SKLtype>`linear_model|SGDRegressor`**</SKLtype>**.

In addition to this XML node, several others are available:

- **<loss>**, ***str, 'squared_loss,' 'huber,' 'epsilon_insensitive,' or 'squared_epsilon_insensitive'***, specifies the loss function to be used. Defaults to 'squared_loss' which refers to the ordinary least squares fit. 'huber' modifies 'squared_loss' to focus less on getting outliers correct by switching from squared to linear loss past a distance of epsilon. 'epsilon_insensitive' ignores errors less than epsilon and is linear past that; this is the loss function used in SVR. 'squared_epsilon_insensitive' is the same but becomes squared loss past a tolerance of epsilon.
  Default: 'squared_loss'

- **<penalty>**, ***str, 'l2' or 'l1' or 'elasticnet'***, sets the penalty (aka regularization term) to be used. Defaults to 'l2' which is the standard regularizer for linear SVM models. 'l1' and 'elasticnet' might bring sparsity to the model (feature selection) not achievable with 'l2'.
  Default: 'l2'

- **<alpha>**,***float***, Constant that multiplies the regularization term. Defaults to 0.0001

- **<l1_ratio>**,***float***, is the Elastic Net mixing parameter, with $0 \leq l1_ratio \leq 1$. l1_ratio=0 corresponds to L2 penalty, l1_ratio=1 to L1.
  Default: 0.15

- **<fit_intercept>**, ***boolean, optional field***, determines whether to calculate the intercept for this model. If set to False, no intercept will be used in the calculations (e.g. data is expected to be already centered).
  Default: True

- **<n_iter>**, ***int, optional field***, specifies the number of passes over the training data (aka epochs).
  Default: 5

- **<shuffle>**, ***boolean, optional field***, specifies whether or not the training data should be shuffled after each epoch.
  Default: True

- **<random_state>**, ***int seed, RandomState instance, or None***, sets the seed of the pseudo random number generator to use when shuffling the data.
  Default: None

- **<verbose>**, ***boolean or integer, optional field***, use verbose mode when fitting the model.
  Default: 0

- **<epsilon>**, ***float***, sets the epsilon in the epsilon-insensitive loss functions; only if loss is 'huber,' 'epsilon_insensitive,' or 'squared_epsilon_insensitive.' For 'huber', determines the threshold at which it becomes less important to get the prediction exactly right. For epsilon-insensitive, any differences between the current prediction and the correct label are ignored if they are less than this threshold.
  Default: 0.1

- **<learning_rate>**, ***string, optional field***, Learning rate:
  - constant: eta = eta0
  - optimal: eta = 1.0/(t+t0)
  - invscaling: eta= eta0 / pow(t, power_t)

  Default: invscaling

- **<eta0>**, ***double***, specifies the initial learning rate.
  Default: 0.01

- **<power_t>**, ***double, optional field***, specifies the exponent for inverse scaling learning rate.

  Default: 0.25

- **<warm_start>**, ***boolean, optional field***, when set to True, the model will reuse the solution of the previous call to fit as initialization, otherwise, it will just erase the previous solution.
  Default: False

It is important to NOTE that RAVEN uses a Z-score normalization of the training data before constructing the *SGDRegressor* ROM:

$$\mathbf{X}' = \frac{(\mathbf{X} - \mu)}{\sigma} \quad (13)$$

17.3.7.2 Support Vector Machines

In machine learning, **Support Vector Machines** (SVMs, also support vector networks) are supervised learning models with associated learning algorithms that analyze data and recognize patterns, used for classification and regression analysis. Given a set of training examples, each marked as belonging to one of two categories, an SVM training algorithm builds a model that assigns new examples into one category or the other, making it a non-probabilistic binary linear classifier. An SVM model is a representation of the examples as points in space, mapped so that the examples of the separate categories are divided by a clear gap that is as wide as possible. New examples are then mapped into that same space and predicted to belong to a category based on which side of the gap they fall on. In addition to performing linear classification, SVMs can efficiently perform a non-linear classification using what is called the kernel trick, implicitly mapping their inputs into high-dimensional feature spaces.
It is important to NOTE that RAVEN uses a Z-score normalization of the training data before constructing the *SVM-based* ROM:

$$\mathbf{X}' = \frac{(\mathbf{X} - \mu)}{\sigma} \quad (14)$$

In the following, all the SVM models available in RAVEN are reported.

17.3.7.2.1 Linear Support Vector Classifier

The *Linear Support Vector Classifier* is similar to SVC with parameter kernel='linear', but implemented in terms of liblinear rather than libsvm, so it has more flexibility in the choice of penalties and loss functions and should scale better (to large numbers of samples). This class supports both dense and sparse input and the multiclass support is handled according to a one-vs-the-rest scheme. In order to use the *Linear Support Vector Classifier*, the user needs to set the sub-node:

<SKLtype>`svm|LinearSVC`**</SKLtype>**.

In addition to this XML node, several others are available:

- **<C>**, ***float, optional field***, sets the penalty parameter C of the error term.
  Default: 1.0

- **<loss>**, ***string, 'hinge' or 'squared_hinge'***, specifies the loss function. 'hinge' is the hinge loss (standard SVM) while 'squared_hinge' is the squared hinge loss.
  Default: 'squared_hinge'

- **<penalty>**, ***string, 'l1' or 'l2'***, specifies the norm used in the penalization. The 'l2' penalty is the standard used in SVC. The 'l1' leads to coef_vectors that are sparse.
  Default: 'l2'

- **<dual>**, ***boolean***, selects the algorithm to either solve the dual or primal optimization problem. Prefer dual=False when n_samples > n_features.
  Default: True

- **<tol>**, ***float, optional field***, specifies the tolerance for stopping criteria.
  Default: 1e-4

- **<multi_class>**, ***string, 'ovr' or 'crammer_singer'***, Determines the multi-class strategy if y contains more than two classes. ovr trains n_classes one-vs-rest classifiers, while crammer_singer optimizes a joint objective over all classes. While crammer_singer is interesting from a theoretical perspective as it is consistent, it is seldom used in practice and rarely leads to better accuracy and is more expensive to compute. If crammer_singer is chosen, the options loss, penalty and dual will be ignored.
  Default: 'ovr'

- **<fit_intercept>**, ***boolean, optional field***, determines whether to calculate the intercept for this model. If set to False, no intercept will be used in the calculations (e.g. data is expected to be already centered).
  Default: True

- **<intercept_scaling>**, ***float, optional field***, when True, the instance vector x becomes [x,self.intercept_scaling], i.e. a "synthetic" feature with constant value equals to intercept_scaling

is appended to the instance vector. The intercept becomes intercept_scaling * synthetic feature weight. **Note:** The synthetic feature weight is subject to l1/l2 regularization as are all other features. To lessen the effect of regularization on the synthetic feature weight (and therefore on the intercept) intercept_scaling has to be increased.
Default: 1

- **<class_weight>**, ***dict, 'auto', optional***, sets the parameter C of class i to class_weight[i]*C for SVC. If not given, all classes are assumed to have weight one. The 'auto' mode uses the values of y to automatically adjust weights inversely proportional to class frequencies.
Default: None

- **<verbose>**, ***boolean or integer, optional field***, use verbose mode when fitting the model.
Default: 0 **Note:** This setting takes advantage of a per-process runtime setting in liblinear that, if enabled, may not work properly in a multithreaded context.

- **<random_state>**, ***int seed, RandomState instance, or None***, represents the seed of the pseudo random number generator to use when shuffling the data for probability estimation.
Default: None

17.3.7.2.2 C-Support Vector Classification

The *C-Support Vector Classification* is a based on libsvm. The fit time complexity is more than quadratic with the number of samples which makes it hard to scale to datasets with more than a couple of 10000 samples. The multiclass support is handled according to a one-vs-one scheme. In order to use the *C-Support Vector Classifier*, the user needs to set the sub-node:

<SKLtype>`svm|SVC`**</SKLtype>**.

In addition to this XML node, several others are available:

- **<C>**, ***float, optional field***, sets the penalty parameter C of the error term.
Default: 1.0

- **<kernel>**, ***string, optional***, specifies the kernel type to be used in the algorithm. It must be one of:
  - 'linear'
  - 'poly'
  - 'rbf'
  - 'sigmoid'
  - 'precomputed'
  - a callable object

If a callable is given it is used to pre-compute the kernel matrix.
Default: 'rbf'

- **<degree>**, ***int, optional field***, determines the degree of the polynomial kernel function ('poly'). Ignored by all other kernels.
Default: 3.0

- **<gamma>**, ***float, optional field***, sets the kernel coefficient for the kernels 'rbf,' 'poly,' and 'sigmoid.' If gamma is 'auto' then 1/n_features will be used instead.
Default: 'auto'

- **<coef0>**, ***float, optional field***, is an independent term in kernel function. It is only significant in 'poly' and 'sigmoid.'
Default: 0.0

- **<probability>**, ***boolean, optional field***, determines whether or not to enable probability estimates. This must be enabled prior to calling fit, and will slow down that method.
Default: False

- **<shrinking>**, ***boolean, optional field***, determines whether or not to use the shrinking heuristic.
Default: True

- **<tol>**, ***float, optional field***, specifies the tolerance for stopping criteria.
Default: 1e-3

- **<cache_size>**, ***float, optional field***, specifies the size of the kernel cache (in MB).

- **<class_weight>**, ***dict, 'auto', optional***, sets the parameter C of class i to class_weight[i]*C for SVC. If not given, all classes are assumed to have weight one. The 'auto' mode uses the values of y to automatically adjust weights inversely proportional to class frequencies.
Default: None

- **<verbose>**, ***boolean, optional field***, use verbose mode when fitting the model.
Default: False **Note:** This setting takes advantage of a per-process runtime setting in libsvm that, if enabled, may not work properly in a multithreaded context.

- **<max_iter>**, ***integer, optional field***, specifies the maximum number of iterations.
Default: -1

- **<random_state>**, ***int seed, RandomState instance, or None***, represents the seed of the pseudo random number generator to use when shuffling the data for probability estimation.
Default: None

17.3.7.2.3 Nu-Support Vector Classification

The *Nu-Support Vector Classification* is similar to SVC but uses a parameter to control the number of support vectors. The implementation is based on libsvm. In order to use the *Nu-Support Vector Classifier*, the user needs to set the sub-node:

<SKLtype>svm|NuSVC**</SKLtype>**.

In addition to this XML node, several others are available:

- **<nu>**, ***float, optional field***, is an upper bound on the fraction of training errors and a lower bound of the fraction of support vectors. Should be in the interval (0, 1].
  Default: 0.5
- **<kernel>**, ***string, optional***, specifies the kernel type to be used in the algorithm. It must be one of:
  - 'linear'
  - 'poly'
  - 'rbf'
  - 'sigmoid'
  - 'precomputed'
  - a callable object

  If a callable is given it is used to pre-compute the kernel matrix.
  Default: 'rbf'
- **<degree>**, ***int, optional field***, determines the degree of the polynomial kernel function ('poly'). Ignored by all other kernels.
  Default: 3
- **<gamma>**, ***float, optional field***, sets the kernel coefficient for the kernels 'rbf,' 'poly,' and 'sigmoid.' If gamma is 'auto' then 1/n_features will be used instead.
  Default: 'auto'
- **<coef0>**, ***float, optional field***, is an independent term in kernel function. It is only significant in 'poly' and 'sigmoid.'
  Default: 0.0
- **<probability>**, ***boolean, optional field***, determines whether or not to enable probability estimates. This must be enabled prior to calling fit, and will slow down that method.
  Default: False

- **<shrinking>**, ***boolean, optional field***, determines whether or not to use the shrinking heuristic.
  Default: True

- **<tol>**, ***float, optional field***, specifies the tolerance for stopping criteria.
  Default: 1e-3

- **<cache_size>**, ***float, optional field***, specifies the size of the kernel cache (in MB).

- **<verbose>**, ***boolean, optional field***, use verbose mode when fitting the model.
  Default: False **Note:** This setting takes advantage of a per-process runtime setting in libsvm that, if enabled, may not work properly in a multithreaded context.

- **<max_iter>**, ***integer, optional field***, specifies the maximum number of iterations.
  Default: -1

- **<random_state>**, ***int seed, RandomState instance, or None***, represents the seed of the pseudo random number generator to use when shuffling the data for probability estimation.
  Default: None

17.3.7.2.4 Support Vector Regression

The *Support Vector Regression* is an epsilon-Support Vector Regression. The free parameters in this model are C and epsilon. The implementations is a based on libsvm. In order to use the *Support Vector Regressor*, the user needs to set the sub-node:

<SKLtype>svm|SVR**</SKLtype>**.

In addition to this XML node, several others are available:

- **<C>**, ***float, optional field***, sets the penalty parameter C of the error term.
  Default: 1.0

- **<epsilon>**, ***float, optional field***, specifies the epsilon-tube within which no penalty is associated in the training loss function with points predicted within a distance epsilon from the actual value.
  Default: 0.1

- **<kernel>**, ***string, optional***, specifies the kernel type to be used in the algorithm. It must be one of:
  - 'linear'
  - 'poly'
  - 'rbf'
  - 'sigmoid'

  - ‘precomputed’
  - a callable object

  If a callable is given it is used to pre-compute the kernel matrix.
  Default: ‘rbf’

- **<degree>**, ***int, optional field***, determines the degree of the polynomial kernel function (‘poly’). Ignored by all other kernels.
  Default: 3.0

- **<gamma>**, ***float, optional field***, sets the kernel coefficient for the kernels ‘rbf,’ ‘poly,’ and ‘sigmoid.’ If gamma is ‘auto’ then 1/n_features will be used instead.
  Default: ‘auto’

- **<coef0>**, ***float, optional field***, is an independent term in kernel function. It is only significant in ‘poly’ and ‘sigmoid.’
  Default: 0.0

- **<shrinking>**, ***boolean, optional field***, determines whether or not to use the shrinking heuristic.
  Default: True

- **<tol>**, ***float, optional field***, specifies the tolerance for stopping criteria.
  Default: 1e-3

- **<cache_size>**, ***float, optional field***, specifies the size of the kernel cache (in MB).

- **<verbose>**, ***boolean, optional field***, use verbose mode when fitting the model.
  Default: False **Note:** This setting takes advantage of a per-process runtime setting in libsvm that, if enabled, may not work properly in a multithreaded context.

- **<max_iter>**, ***integer, optional field***, specifies the maximum number of iterations.
  Default: -1

17.3.7.3 Multi Class

Multiclass classification means a classification task with more than two classes; e.g., classify a set of images of fruits which may be oranges, apples, or pears. Multiclass classification makes the assumption that each sample is assigned to one and only one label: a fruit can be either an apple or a pear but not both at the same time.

It is important to NOTE that RAVEN does not pre-normalize the training data before constructing the *multi-class-based* ROM.

In the following, all the multi-class models available in RAVEN are reported.

17.3.7.3.1 One-vs-the-rest (OvR) multiclass/multilabel strategy

The *One-vs-the-rest (OvR) multiclass/multilabel strategy*, also known as one-vs-all, consists in fitting one classifier per class. For each classifier, the class is fitted against all the other classes. In addition to its computational efficiency (only n_classes classifiers are needed), one advantage of this approach is its interpretability. Since each class is represented by one and one classifier only, it is possible to gain knowledge about the class by inspecting its corresponding classifier. This is the most commonly used strategy and is a fair default choice.

In order to use the *One-vs-the-rest (OvR) multiclass/multilabel classifier*, the user needs to set the sub-node:

<SKLtype>`multiClass|OneVsRestClassifier`**</SKLtype>**.

In addition to this XML node, several others are available:

- **<estimator>**, ***boolean, required field***, An estimator object implementing fit and one of decision_function or predict_proba. This XML node needs to contain the following attribute:
  - **estimatorType**, ***required string attribute***, this attribute is another reduced order mode type that needs to be used for the construction of the multi-class algorithms. Each sub-sequential node depends on the chosen ROM.

17.3.7.3.2 One-vs-one multiclass strategy

The *One-vs-one multiclass strategy* consists in fitting one classifier per class pair. At prediction time, the class which received the most votes is selected. Since it requires to fit n_classes * (n_classes - 1) / 2 classifiers, this method is usually slower than one-vs-the-rest, due to its O(n_classes^2) complexity. However, this method may be advantageous for algorithms such as kernel algorithms which do not scale well with n_samples. This is because each individual learning problem only involves a small subset of the data whereas, with one-vs-the-rest, the complete dataset is used n_classes times.

In order to use the *One-vs-one multiclass classifier*, the user needs to set the sub-node:

<SKLtype>`multiClass|OneVsOneClassifier`**</SKLtype>**.

In addition to this XML node, several others are available:

- **<estimator>**, ***boolean, required field***, An estimator object implementing fit and one of decision_function or predict_proba. This XML node needs to contain the following attribute:
  - **estimatorType**, ***required string attribute***, this attribute is another reduced order mode type that needs to be used for the construction of the multi-class algorithms.

Each sub-sequential node depends on the chosen ROM.

17.3.7.3.3 Error-Correcting Output-Code multiclass strategy

The *Error-Correcting Output-Code multiclass strategy* consists in representing each class with a binary code (an array of 0s and 1s). At fitting time, one binary classifier per bit in the code book is fitted. At prediction time, the classifiers are used to project new points in the class space and the class closest to the points is chosen. The main advantage of these strategies is that the number of classifiers used can be controlled by the user, either for compressing the model ($0 < code_size < 1$) or for making the model more robust to errors ($code_size > 1$).

In order to use the *Error-Correcting Output-Code multiclass classifier*, the user needs to set the sub-node:

<SKLtype>`multiClass|OutputCodeClassifier`**</SKLtype>**.

In addition to this XML node, several others are available:

- **<estimator>**, ***boolean, required field***, An estimator object implementing fit and one of decision_function or predict_proba. This XML node needs to contain the following attribute:
  - **estimatorType**, ***required string attribute***, this attribute is another reduced order mode type that needs to be used for the construction of the multi-class algorithms. Each sub-sequential node depends on the chosen ROM.
- **<code_size>**, ***float, required field***, represents the percentage of the number of classes to be used to create the code book. A number between 0 and 1 will require fewer classifiers than one-vs-the-rest. A number greater than 1 will require more classifiers than one-vs-the-rest.

17.3.7.4 Naive Bayes

Naive Bayes methods are a set of supervised learning algorithms based on applying Bayes' theorem with the "naive" assumption of independence between every pair of features. Given a class variable y and a dependent feature vector x_1 through x_n, Bayes' theorem states the following relationship:

$$P(y \mid x_1, \dots, x_n) = \frac{P(y) P(x_1, \dots x_n \mid y)}{P(x_1, \dots, x_n)} \tag{15}$$

Using the naive independence assumption that

$$P(x_i | y, x_1, \dots, x_{i-1}, x_{i+1}, \dots, x_n) = P(x_i | y), \tag{16}$$

for all i, this relationship is simplified to

$$P(y \mid x_1, \dots, x_n) = \frac{P(y) \prod_{i=1}^{n} P(x_i \mid y)}{P(x_1, \dots, x_n)} \tag{17}$$

Since $P(x_1, \dots, x_n)$ is constant given the input, we can use the following classification rule:

$$P(y \mid x_1, \dots, x_n) \propto P(y) \prod_{i=1}^{n} P(x_i \mid y) \Downarrow \quad (18)$$

$$\hat{y} = \arg\max_y P(y) \prod_{i=1}^{n} P(x_i \mid y), \quad (19)$$

and we can use Maximum A Posteriori (MAP) estimation to estimate $P(y)$ and $P(x_i \mid y)$; the former is then the relative frequency of class y in the training set. The different naive Bayes classifiers differ mainly by the assumptions they make regarding the distribution of $P(x_i \mid y)$.

In spite of their apparently over-simplified assumptions, naive Bayes classifiers have worked quite well in many real-world situations, famously document classification and spam filtering. They require a small amount of training data to estimate the necessary parameters. (For theoretical reasons why naive Bayes works well, and on which types of data it does, see the references below.) Naive Bayes learners and classifiers can be extremely fast compared to more sophisticated methods. The decoupling of the class conditional feature distributions means that each distribution can be independently estimated as a one dimensional distribution. This in turn helps to alleviate problems stemming from the curse of dimensionality.

On the flip side, although naive Bayes is known as a decent classifier, it is known to be a bad estimator, so the probability outputs from predict_proba are not to be taken too seriously. In the following, all the Naive Bayes available in RAVEN are reported.

17.3.7.4.1 Gaussian Naive Bayes

The *Gaussian Naive Bayes strategy* implements the Gaussian Naive Bayes algorithm for classification. The likelihood of the features is assumed to be Gaussian:

$$P(x_i \mid y) = \frac{1}{\sqrt{2\pi\sigma_y^2}} \exp\left(-\frac{(x_i - \mu_y)^2}{2\sigma_y^2}\right) \quad (20)$$

The parameters σ_y and μ_y are estimated using maximum likelihood.

In order to use the *Gaussian Naive Bayes strategy*, the user needs to set the sub-node:

<SKLtype>`naiveBayes|GaussianNB`**</SKLtype>**.

There are no additional sub-nodes available for this method.

It is important to NOTE that RAVEN uses a Z-score normalization of the training data before constructing the *GaussianNB* ROM:

$$\mathbf{X}' = \frac{(\mathbf{X} - \mu)}{\sigma} \quad (21)$$

17.3.7.4.2 Multinomial Naive Bayes

The *Multinomial Naive Bayes* implements the naive Bayes algorithm for multinomially distributed data, and is one of the two classic naive Bayes variants used in text classification (where the data is typically represented as word vector counts, although tf-idf vectors are also known to work well in practice). The distribution is parametrized by vectors $\theta_y = (\theta_{y1}, \ldots, \theta_{yn})$ for each class y, where n is the number of features (in text classification, the size of the vocabulary) and θ_{yi} is the probability $P(x_i \mid y)$ of feature i appearing in a sample belonging to class y. The parameters θ_y are estimated by a smoothed version of maximum likelihood, i.e. relative frequency counting:

$$\hat{\theta}_{yi} = \frac{N_{yi} + \alpha}{N_y + \alpha n} \tag{22}$$

where $N_{yi} = \sum_{x \in T} x_i$ is the number of times feature i appears in a sample of class y in the training set T, and $N_y = \sum_{i=1}^{|T|} N_{yi}$ is the total count of all features for class y. The smoothing priors $\alpha \geq 0$ account for features not present in the learning samples and prevents zero probabilities in further computations. Setting $\alpha = 1$ is called Laplace smoothing, while $\alpha < 1$ is called Lidstone smoothing. In order to use the *Multinomial Naive Bayes strategy*, the user needs to set the sub-node:

<SKLtype>naiveBayes|MultinomialNB**</SKLtype>**.

In addition to this XML node, several others are available:

- **<alpha>**, ***float, optional field***, specifies an additive (Laplace/Lidstone) smoothing parameter (0 for no smoothing).
  Default: 1.0
- **<fit_prior>**, ***boolean, required field***, determines whether to learn class prior probabilities or not. If false, a uniform prior will be used.
  Default: True
- **<class_prior>**, ***array-like float (n_classes), optional field***, specifies prior probabilities of the classes. If specified, the priors are not adjusted according to the data.
  Default: None

It is important to NOTE that RAVEN does not pre-normalize the training data before constructing the *MultinomialNB* ROM.

17.3.7.4.3 Bernoulli Naive Bayes

The *Bernoulli Naive Bayes* implements the naive Bayes training and classification algorithms for data that is distributed according to multivariate Bernoulli distributions; i.e., there may be multiple features but each one is assumed to be a binary-valued (Bernoulli, boolean) variable. Therefore,

this class requires samples to be represented as binary-valued feature vectors; if handed any other kind of data, a *Bernoulli Naive Bayes* instance may binarize its input (depending on the binarize parameter). The decision rule for Bernoulli naive Bayes is based on

$$P(x_i \mid y) = P(i \mid y)x_i + (1 - P(i \mid y))(1 - x_i) \tag{23}$$

which differs from multinomial NB's rule in that it explicitly penalizes the non-occurrence of a feature i that is an indicator for class y, where the multinomial variant would simply ignore a non-occurring feature. In the case of text classification, word occurrence vectors (rather than word count vectors) may be used to train and use this classifier. *Bernoulli Naive Bayes* might perform better on some datasets, especially those with shorter documents. It is advisable to evaluate both models, if time permits. In order to use the *Bernoulli Naive Bayes strategy*, the user needs to set the sub-node:

<SKLtype>naiveBayes|BernoulliNB**</SKLtype>**.

In addition to this XML node, several others are available:

- **<alpha>**, ***float, optional field***, specifies an additive (Laplace/Lidstone) smoothing parameter (0 for no smoothing).
  Default: 1.0
- **<binarize>**, ***float, optional field***, Threshold for binarizing (mapping to booleans) of sample features. If None, input is presumed to already consist of binary vectors.
  Default: 0.0
- **<fit_prior>**, ***boolean, required field***, determines whether to learn class prior probabilities or not. If false, a uniform prior will be used.
  Default: True
- **<class_prior>**, ***array-like float (n_classes), optional field***, specifies prior probabilities of the classes. If specified, the priors are not adjusted according to the data.
  Default: None

It is important to NOTE that RAVEN uses a Z-score normalization of the training data before constructing the *BernoulliNB* ROM:

$$\mathbf{X'} = \frac{(\mathbf{X} - \mu)}{\sigma} \tag{24}$$

17.3.7.5 Neighbors

The *Neighbors* class provides functionality for unsupervised and supervised neighbor-based learning methods. The unsupervised nearest neighbors method is the foundation of many other learning

methods, notably manifold learning and spectral clustering. Supervised neighbors-based learning comes in two flavors: classification for data with discrete labels, and regression for data with continuous labels.

The principle behind nearest neighbor methods is to find a predefined number of training samples closest in distance to the new point, and predict the label from these. The number of samples can be a user-defined constant (k-nearest neighbor learning), or vary based on the local density of points (radius-based neighbor learning). The distance can, in general, be any metric measure: standard Euclidean distance is the most common choice. Neighbor-based methods are known as non-generalizing machine learning methods, since they simply "remember" all of its training data (possibly transformed into a fast indexing structure such as a Ball Tree or KD Tree.).

It is important to NOTE that RAVEN uses a Z-score normalization of the training data before constructing the *Neighbors-based* ROM:

$$\mathbf{X}' = \frac{(\mathbf{X} - \mu)}{\sigma} \tag{25}$$

In the following, all the Neighbors' models available in RAVEN are reported.

17.3.7.5.1 K Neighbors Classifier

The *K Neighbors Classifier* is a type of instance-based learning or non-generalizing learning: it does not attempt to construct a general internal model, but simply stores instances of the training data. Classification is computed from a simple majority vote of the nearest neighbors of each point: a query point is assigned the data class which has the most representatives within the nearest neighbors of the point. It implements learning based on the k nearest neighbors of each query point, where k is an integer value specified by the user.

In order to use the *K Neighbors Classifier*, the user needs to set the sub-node:

<SKLtype>neighbors|KNeighborsClassifier**</SKLtype>**.

In addition to this XML node, several others are available:

- **<n_neighbors>**, ***integer, optional field***, specifies the number of neighbors to use by default for 'k_neighbors' queries.
  Default: 5
- **<weights>**, ***string, optional field***, specifies the weight function used in prediction. Possible values:
  - *uniform* : uniform weights. All points in each neighborhood are weighted equally;

- *distance* : weight points by the inverse of their distance. In this case, closer neighbors of a query point will have a greater influence than neighbors which are further away.

Default: uniform

- **<algorithm>**, ***string, optional field***, specifies the algorithm used to compute the nearest neighbors:
  - *ball_tree* will use BallTree.
  - *kd_tree* will use KDtree.
  - *brute* will use a brute-force search.
  - *auto* will attempt to decide the most appropriate algorithm based on the values passed to fit method.

  Note: Fitting on sparse input will override the setting of this parameter, using brute force.
  Default: auto

- **<leaf_size>**, ***integer, optional field***, sets the leaf size passed to the BallTree or KDTree. This can affect the speed of the construction and query, as well as the memory required to store the tree. The optimal value depends on the nature of the problem.
  Default: 30

- **<metric>**, ***string, optional field***, sets the distance metric to use for the tree. The Minkowski metric with p=2 is equivalent to the standard Euclidean metric.
  Default: minkowski

- **<p>**, ***integer, optional field***, is a parameter for the Minkowski metric. When $p = 1$, this is equivalent to using manhattan distance (L1), and euclidean distance (L2) for $p = 2$. For arbitrary p, minkowski distance (L_p) is used.
  Default: 2

17.3.7.5.2 Radius Neighbors Classifier

The *Radius Neighbors Classifier* is a type of instance-based learning or non-generalizing learning: it does not attempt to construct a general internal model, but simply stores instances of the training data. Classification is computed from a simple majority vote of the nearest neighbors of each point: a query point is assigned the data class which has the most representatives within the nearest neighbors of the point. It implements learning based on the number of neighbors within a fixed radius r of each training point, where r is a floating-point value specified by the user.

In order to use the *Radius Neighbors Classifier*, the user needs to set the sub-node:

<SKLtype>`neighbors|RadiusNeighbors`**</SKLtype>**.

In addition to this XML node, several others are available:

- **<radius>**, ***float, optional field***, specifies the range of parameter space to use by default for 'radius_neighbors' queries.
  Default: 1.0

- **<weights>**, ***string, optional field***, specifies the weight function used in prediction. Possible values:

  – *uniform* : uniform weights. All points in each neighborhood are weighted equally;

  – *distance* : weight points by the inverse of their distance. In this case, closer neighbors of a query point will have a greater influence than neighbors which are further away.

  Default: uniform

- **<algorithm>**, ***string, optional field***, specifies the algorithm used to compute the nearest neighbors:

  – *ball_tree* will use BallTree.

  – *kd_tree* will use KDtree.

  – *brute* will use a brute-force search.

  – *auto* will attempt to decide the most appropriate algorithm based on the values passed to fit method.

  Note: Fitting on sparse input will override the setting of this parameter, using brute force.
  Default: auto

- **<leaf_size>**, ***integer, optional field***, sets the leaf size passed to the BallTree or KDTree. This can affect the speed of the construction and query, as well as the memory required to store the tree. The optimal value depends on the nature of the problem.
  Default: 30

- **<metric>**, ***string, optional field***, sets the distance metric to use for the tree. The Minkowski metric with p=2 is equivalent to the standard Euclidean metric.
  Default: minkowski

- **<p>**, ***integer, optional field***, is a parameter for the Minkowski metric. When $p = 1$, this is equivalent to using manhattan distance (L1), and euclidean distance (L2) for $p = 2$. For arbitrary p, minkowski distance (L_p) is used.
  Default: 2

- **<outlier_label>**, ***integer, optional field***, is a label, which is given for outlier samples (samples with no neighbors on a given radius). If set to None, ValueError is raised, when an outlier is detected.
  Default: None

17.3.7.5.3 K Neighbors Regressor

The *K Neighbors Regressor* can be used in cases where the data labels are continuous rather than discrete variables. The label assigned to a query point is computed based on the mean of the labels of its nearest neighbors. It implements learning based on the k nearest neighbors of each query point, where k is an integer value specified by the user.

In order to use the *K Neighbors Regressor*, the user needs to set the sub-node:

<SKLtype>neighbors|KNeighborsRegressor**</SKLtype>**.

In addition to this XML node, several others are available:

- **<n_neighbors>**, ***integer, optional field***, specifies the number of neighbors to use by default for 'k_neighbors' queries.
  Default: 5

- **<weights>**, ***string, optional field***, specifies the weight function used in prediction. Possible values:

  - *uniform* : uniform weights. All points in each neighborhood are weighted equally;
  - *distance* : weight points by the inverse of their distance. In this case, closer neighbors of a query point will have a greater influence than neighbors which are further away.

  Default: uniform

- **<algorithm>**, ***string, optional field***, specifies the algorithm used to compute the nearest neighbors:

  - *ball_tree* will use BallTree.
  - *kd_tree* will use KDtree.
  - *brute* will use a brute-force search.
  - *auto* will attempt to decide the most appropriate algorithm based on the values passed to fit method.

  Note: Fitting on sparse input will override the setting of this parameter, using brute force.
  Default: auto

- **<leaf_size>**, ***integer, optional field***, sets the leaf size passed to the BallTree or KDTree. This can affect the speed of the construction and query, as well as the memory required to store the tree. The optimal value depends on the nature of the problem.
  Default: 30

- **<metric>**, ***string, optional field***, sets the distance metric to use for the tree. The Minkowski metric with p=2 is equivalent to the standard Euclidean metric.
  Default: minkowski

- **<p>**, ***integer, optional field***, is a parameter for the Minkowski metric. When $p = 1$, this is equivalent to using manhattan distance (L1), and euclidean distance (L2) for $p = 2$. For arbitrary p, minkowski distance (L_p) is used.
  Default: 2

17.3.7.5.4 Radius Neighbors Regressor

The *Radius Neighbors Regressor* can be used in cases where the data labels are continuous rather than discrete variables. The label assigned to a query point is computed based on the mean of the labels of its nearest neighbors. It implements learning based on the neighbors within a fixed radius r of the query point, where r is a floating-point value specified by the user.

In order to use the *Radius Neighbors Regressor*, the user needs to set the sub-node:

<SKLtype>`neighbors|RadiusNeighborsRegressor`**</SKLtype>**.

In addition to this XML node, several others are available:

- **<radius>**, ***float, optional field***, specifies the range of parameter space to use by default for 'radius_neighbors' queries.
  Default: 1.0

- **<weights>**, ***string, optional field***, specifies the weight function used in prediction. Possible values:
  - *uniform* : uniform weights. All points in each neighborhood are weighted equally;
  - *distance* : weight points by the inverse of their distance. In this case, closer neighbors of a query point will have a greater influence than neighbors which are further away.

  Default: uniform

- **<algorithm>**, ***string, optional field***, specifies the algorithm used to compute the nearest neighbors:
  - *ball_tree* will use BallTree.
  - *kd_tree* will use KDtree.
  - *brute* will use a brute-force search.
  - *auto* will attempt to decide the most appropriate algorithm based on the values passed to fit method.

Note: Fitting on sparse input will override the setting of this parameter, using brute force.
Default: auto

- **<leaf_size>**, ***integer, optional field***, sets the leaf size passed to the BallTree or KDTree. This can affect the speed of the construction and query, as well as the memory required to store the tree. The optimal value depends on the nature of the problem.
  Default: 30
- **<metric>**, ***string, optional field***, sets the distance metric to use for the tree. The Minkowski metric with p=2 is equivalent to the standard Euclidean metric.
  Default: minkowski
- **<p>**, ***integer, optional field***, is a parameter for the Minkowski metric. When $p = 1$, this is equivalent to using manhattan distance (L1), and euclidean distance (L2) for $p = 2$. For arbitrary p, minkowski distance (L_p) is used.
  Default: 2

17.3.7.5.5 Nearest Centroid Classifier

The *Nearest Centroid classifier* is a simple algorithm that represents each class by the centroid of its members. It also has no parameters to choose, making it a good baseline classifier. It does, however, suffer on non-convex classes, as well as when classes have drastically different variances, as equal variance in all dimensions is assumed.

In order to use the *Nearest Centroid Classifier*, the user needs to set the sub-node:

<SKLtype>`neighbors|NearestCentroid`**</SKLtype>**.

In addition to this XML node, several others are available:

- **<shrink_threshold>**,***float, optional field***, defines the threshold for shrinking centroids to remove features.
  Default: None

The *Quadratic Discriminant Analysis* is a classifier with a quadratic decision boundary, generated by fitting class conditional densities to the data and using Bayes' rule. The model fits a Gaussian density to each class.

In order to use the *Quadratic Discriminant Analysis Classifier*, the user needs to set the sub-node:

<SKLtype>`qda|QDA`**</SKLtype>**.

In addition to this XML node, several others are available:

- **<priors>**, ***array-like (n_classes), optional field***, specifies the priors on the classes.
  Default: None

- **<reg_param>**,***float, optional field***, regularizes the covariance estimate as (1-reg_param)*Sigma + reg_param*Identity(n_features).
  Default: 0.0

It is important to NOTE that RAVEN does not pre-normalize the training data before constructing the *QDA* ROM.

17.3.7.6 Tree

Decision Trees (DTs) are a non-parametric supervised learning method used for classification and regression. The goal is to create a model that predicts the value of a target variable by learning simple decision rules inferred from the data features.

- Some advantages of decision trees are:
- Simple to understand and to interpret. Trees can be visualized.
- Requires little data preparation. Other techniques often require data normalization, dummy variables need to be created and blank values to be removed. Note however, that this module does not support missing values.
- The cost of using the tree (i.e., predicting data) is logarithmic in the number of data points used to train the tree.
- Able to handle both numerical and categorical data. Other techniques are usually specialized in analyzing datasets that have only one type of variable.
- Able to handle multi-output problems.
- Uses a white box model. If a given situation is observable in a model, the explanation for the condition is easily explained by boolean logic. By contrast, in a black box model (e.g., in an artificial neural network), results may be more difficult to interpret.
- Possible to validate a model using statistical tests. That makes it possible to account for the reliability of the model.
- Performs well even if its assumptions are somewhat violated by the true model from which the data were generated.

The disadvantages of decision trees include:

- Decision-tree learners can create over-complex trees that do not generalise the data well. This is called overfitting. Mechanisms such as pruning (not currently supported), setting the minimum number of samples required at a leaf node or setting the maximum depth of the tree are necessary to avoid this problem.

- Decision trees can be unstable because small variations in the data might result in a completely different tree being generated. This problem is mitigated by using decision trees within an ensemble.

- The problem of learning an optimal decision tree is known to be NP-complete under several aspects of optimality and even for simple concepts. Consequently, practical decision-tree learning algorithms are based on heuristic algorithms such as the greedy algorithm where locally optimal decisions are made at each node. Such algorithms cannot guarantee to return the globally optimal decision tree. This can be mitigated by training multiple trees in an ensemble learner, where the features and samples are randomly sampled with replacement.

- There are concepts that are hard to learn because decision trees do not express them easily, such as XOR, parity or multiplexer problems.

- Decision tree learners create biased trees if some classes dominate. It is therefore recommended to balance the dataset prior to fitting with the decision tree.

It is important to NOTE that RAVEN uses a Z-score normalization of the training data before constructing the *tree-based* ROM:

$$\mathbf{X}' = \frac{(\mathbf{X} - \mu)}{\sigma} \tag{26}$$

In the following, all the tree-based algorithms available in RAVEN are reported.

17.3.7.6.1 Decision Tree Classifier

The *Decision Tree Classifier* is a classifier that is based on the decision tree logic.

In order to use the *Decision Tree Classifier*, the user needs to set the sub-node:

<SKLtype>`tree|DecisionTreeClassifier`**</SKLtype>**.

In addition to this XML node, several others are available:

- **<criterion>**, ***string, optional field***, specifies the function used to measure the quality of a split. Supported criteria are “gini” for the Gini impurity and “entropy” for the information gain.
  Default: gini

- **<splitter>**, ***string, optional field***, specifies the strategy used to choose the split at each node. Supported strategies are "best" to choose the best split and "random" to choose the best random split.
  Default: best

- **<max_features>**, ***int, float or string, optional field***, sets the number of features to consider when looking for the best split:

  - If int, then consider max_features features at each split.
  - If float, then max_features is a percentage and int(max_features * n_features) features are considered at each split.
  - If "auto," then max_features=sqrt(n_features).
  - If "sqrt," then max_features=sqrt(n_features).
  - If "log2," then max_features=log2(n_features).
  - If None, then max_features=n_features.

  Note: The search for a split does not stop until at least one valid partition of the node samples is found, even if it requires to effectively inspect more than max_features features.
  Default: None

- **<max_depth>**, ***integer, optional field***, determines the maximum depth of the tree. If None, then nodes are expanded until all leaves are pure or until all leaves contain less than min_samples_split samples. Ignored if max_samples_leaf is not None.
  Default: None

- **<min_samples_split>**, ***integer, optional field***, sets the minimum number of samples required to split an internal node.
  Default: 2

- **<min_samples_leaf>**, ***integer, optional field***, sets the minimum number of samples required to be at a leaf node.
  Default: 1

- **<max_leaf_nodes>**, ***integer, optional field***, grow a tree with max_leaf_nodes in best-first fashion. Best nodes are defined by relative reduction in impurity. If None then unlimited number of leaf nodes. If not None then max_depth will be ignored.
  Default: None

17.3.7.6.2 Decision Tree Regressor

The *Decision Tree Regressor* is a Regressor that is based on the decision tree logic. In order to use the *Decision Tree Regressor*, the user needs to set the sub-node:

<SKLtype>tree|DecisionTreeRegressor**</SKLtype>**.

In addition to this XML node, several others are available:

- **<criterion>**, ***string, optional field***, specifies the function used to measure the quality of a split. The only supported criterion is "mse" for mean squared error.
  Default: mse

- **<splitter>**, ***string, optional field***, specifies the strategy used to choose the split at each node. Supported strategies are "best" to choose the best split and "random" to choose the best random split.
  Default: best

- **<max_features>**, ***int, float or string, optional field***, sets the number of features to consider when looking for the best split:

  - If int, then consider max_features features at each split.
  - If float, then max_features is a percentage and int(max_features * n_features) features are considered at each split.
  - If "auto," then max_features=sqrt(n_features).
  - If "sqrt," then max_features=sqrt(n_features).
  - If "log2," then max_features=log2(n_features).
  - If None, then max_features=n_features.

  Note: The search for a split does not stop until at least one valid partition of the node samples is found, even if it requires to effectively inspect more than max_features features.
  Default: None

- **<max_depth>**, ***integer, optional field***, determines the maximum depth of the tree. If None, then nodes are expanded until all leaves are pure or until all leaves contain less than min_samples_split samples. Ignored if max_samples_leaf is not None.
  Default: None

- **<min_samples_split>**, ***integer, optional field***, sets the minimum number of samples required to split an internal node.
  Default: 2

- **<min_samples_leaf>**, ***integer, optional field***, sets the minimum number of samples required to be at a leaf node.
  Default: 1

- **<max_leaf_nodes>**, ***integer, optional field***, grow a tree with max_leaf_nodes in best-first fashion. Best nodes are defined by relative reduction in impurity. If None then unlimited number of leaf nodes. If not None then max_depth will be ignored.
  Default: None

17.3.7.6.3 Extra Tree Classifier

The *Extra Tree Classifier* is an extremely randomized tree classifier. Extra-trees differ from classic decision trees in the way they are built. When looking for the best split to separate the samples of a node into two groups, random splits are drawn for each of the max_features randomly selected features and the best split among those is chosen. When max_features is set 1, this amounts to building a totally random decision tree.

In order to use the *Extra Tree Classifier*, the user needs to set the sub-node:

<SKLtype>`tree|ExtraTreeClassifier`**</SKLtype>**.

In addition to this XML node, several others are available:

- **<criterion>**, ***string, optional field***, specifies the function used to measure the quality of a split. Supported criteria are "gini" for the Gini impurity and "entropy" for the information gain.
  Default: gini

- **<splitter>**, ***string, optional field***, specifies the strategy used to choose the split at each node. Supported strategies are "best" to choose the best split and "random" to choose the best random split.
  Default: random

- **<max_features>**, ***int, float or string, optional field***, sets the number of features to consider when looking for the best split:

  - If int, then consider max_features features at each split.
  - If float, then max_features is a percentage and int(max_features * n_features) features are considered at each split.
  - If "auto," then max_features=sqrt(n_features).
  - If "sqrt," then max_features=sqrt(n_features).
  - If "log2," then max_features=log2(n_features).
  - If None, then max_features=n_features.

  Note: The search for a split does not stop until at least one valid partition of the node samples is found, even if it requires to effectively inspect more than max_features features.
  Default: auto

- **<max_depth>**, ***integer, optional field***, determines the maximum depth of the tree. If None, then nodes are expanded until all leaves are pure or until all leaves contain less than min_samples_split samples. Ignored if max_samples_leaf is not None.
  Default: None

- **<min_samples_split>**, ***integer, optional field***, sets the minimum number of samples required to split an internal node.
  Default: 2

- **<min_samples_leaf>**, ***integer, optional field***, sets the minimum number of samples required to be at a leaf node.
  Default: 1

- **<max_leaf_nodes>**, ***integer, optional field***, grow a tree with max_leaf_nodes in best-first fashion. Best nodes are defined by relative reduction in impurity. If None then unlimited number of leaf nodes. If not None then max_depth will be ignored.
  Default: None

17.3.7.6.4 Extra Tree Regressor

The *Extra Tree Regressor* is an extremely randomized tree regressor. Extra-trees differ from classic decision trees in the way they are built. When looking for the best split to separate the samples of a node into two groups, random splits are drawn for each of the max_features randomly selected features and the best split among those is chosen. When max_features is set 1, this amounts to building a totally random decision tree.

In order to use the *Extra Tree Regressor*, the user needs to set the sub-node:

<SKLtype>`tree|ExtraTreeRegressor`**</SKLtype>**.

In addition to this XML node, several others are available:

- **<criterion>**, ***string, optional field***, specifies the function used to measure the quality of a split. The only supported criterion is "mse" for mean squared error.
  Default: mse

- **<splitter>**, ***string, optional field***, specifies the strategy used to choose the split at each node. Supported strategies are "best" to choose the best split and "random" to choose the best random split.
  Default: random

- **<max_features>**, ***int, float or string, optional field***, sets the number of features to consider when looking for the best split:
  - If int, then consider max_features features at each split.
  - If float, then max_features is a percentage and int(max_features * n_features) features are considered at each split.
  - If "auto," then max_features=sqrt(n_features).

- If "sqrt," then max_features=sqrt(n_features).
- If "log2," then max_features=log2(n_features).
- If None, then max_features=n_features.

Note: The search for a split does not stop until at least one valid partition of the node samples is found, even if it requires to effectively inspect more than max_features features.
Default: auto

- **<max_depth>**, ***integer, optional field***, determines the maximum depth of the tree. If None, then nodes are expanded until all leaves are pure or until all leaves contain less than min_samples_split samples. Ignored if max_samples_leaf is not None.
Default: None

- **<min_samples_split>**, ***integer, optional field***, sets the minimum number of samples required to split an internal node.
Default: 2

- **<min_samples_leaf>**, ***integer, optional field***, sets the minimum number of samples required to be at a leaf node.
Default: 1

- **<max_leaf_nodes>**, ***integer, optional field***, grow a tree with max_leaf_nodes in best-first fashion. Best nodes are defined by relative reduction in impurity. If None then unlimited number of leaf nodes. If not None then max_depth will be ignored.
Default: None

17.3.7.7 Gaussian Process

Gaussian Processes for Machine Learning (GPML) is a generic supervised learning method primarily designed to solve regression problems. The advantages of Gaussian Processes for Machine Learning are:

- The prediction interpolates the observations (at least for regular correlation models).
- The prediction is probabilistic (Gaussian) so that one can compute empirical confidence intervals and exceedance probabilities that might be used to refit (online fitting, adaptive fitting) the prediction in some region of interest.
- Versatile: different linear regression models and correlation models can be specified. Common models are provided, but it is also possible to specify custom models provided they are stationary.

The disadvantages of Gaussian Processes for Machine Learning include:

- It is not sparse. It uses the whole samples/features information to perform the prediction.
- It loses efficiency in high dimensional spaces – namely when the number of features exceeds a few dozens. It might indeed give poor performance and it loses computational efficiency.
- Classification is only a post-processing, meaning that one first needs to solve a regression problem by providing the complete scalar float precision output y of the experiment one is attempting to model.

In order to use the *Gaussian Process Regressor*, the user needs to set the sub-node:

<SKLtype>GaussianProcess|GaussianProcess**</SKLtype>**.

In addition to this XML node, several others are available:

- **<regr>**, ***string, optional field***, is a regression function returning an array of outputs of the linear regression functional basis. The number of observations n_samples should be greater than the size p of this basis. Available built-in regression models are 'constant,' 'linear,' and 'quadratic.'
  Default: constant
- **<corr>**, ***string, optional field***, is a stationary autocorrelation function returning the autocorrelation between two points x and x'. Default assumes a squared-exponential autocorrelation model. Built-in correlation models are 'absolute_exponential,' 'squared_exponential,' 'generalized_exponential,' 'cubic,' and 'linear.'
  Default: squared_exponential
- **<beta0>**,***float, array-like, optional field***, specifies the regression weight vector to perform Ordinary Kriging (OK).
  Default: None
- **<storage_mode>**, ***string, optional field***, specifies whether the Cholesky decomposition of the correlation matrix should be stored in the class (storage_mode = 'full') or not (storage_mode = 'light').
  Default: full
- **<verbose>**, ***boolean, optional field***, use verbose mode when fitting the model.
  Default: False
- **<theta0>**, ***float, array-like, optional field***, is an array with shape (n_features,) or (1,). This represents the parameters in the autocorrelation model. If thetaL and thetaU are also specified, theta0 is considered as the starting point for the maximum likelihood estimation of the best set of parameters.
  Default: [1e-1]

- **<thetaL>**, ***float, array-like, optional field***, is an array with shape matching that defined by **<theta0>**. Lower bound on the autocorrelation parameters for maximum likelihood estimation.
  Default: None

- **<thetaU>**, ***float, array-like, optional field***, is an array with shape matching that defined by **<theta0>**. Upper bound on the autocorrelation parameters for maximum likelihood estimation.
  Default: None

- **<normalize>**, ***boolean, optional field***, if True, the input X and observations y are centered and reduced w.r.t. means and standard deviations estimated from the n_samples observations provided.
  Default: True

- **<nugget>**, ***float, optional field***,specifies a nugget effect to allow smooth predictions from noisy data. The nugget is added to the diagonal of the assumed training covariance. In this way it acts as a Tikhonov regularization in the problem. In the special case of the squared exponential correlation function, the nugget mathematically represents the variance of the input values.
  *Default: 10 * MACHINE_EPSILON*

- **<optimizer>**, ***string, optional field***, specifies the optimization algorithm to be used. Available optimizers are: 'fmin_cobyla', 'Welch'.
  Default: fmin_cobyla

- **<random_start>**, ***integer, optional field***, sets the number of times the Maximum Likelihood Estimation should be performed from a random starting point. The first MLE always uses the specified starting point (theta0), the next starting points are picked at random according to an exponential distribution (log-uniform on [thetaL, thetaU]).
  Default: 1

- **<random_state>**, ***integer, optional field***, is the seed of the internal random number generator.
  Default: None

It is important to NOTE that RAVEN does not pre-normalize the training data before constructing the *GaussianProcess* ROM.

Example:

```
<Simulation>
  ...
  <Models>
```

```
    ...
   <ROM name='aUserDefinedName' subType='SciKitLearn'>
     <Features>var1,var2,var3</Features>
     <Target>result1</Target>
     <SKLtype>linear_model|LinearRegression</SKLtype>
     <fit_intercept>True</fit_intercept>
     <normalize>False</normalize>
   </ROM>
    ...
  </Models>
  ...
</Simulation>
```

17.3.8 ARMA

The ARMA sub-type contains a single ROM type, based on an autoregressive moving average time series model. ARMA is a type of time dependent model that characterizes the autocorrelation between time series data. The mathematic description of ARMA is given as

$$x_t = \sum_{i=1}^{p} \phi_i x_{t-i} + \alpha_t + \sum_{j=1}^{q} \theta_j \alpha_{t-j},$$

where x is a vector of dimension n, and ϕ_i and θ_j are both n by n matrices. When $q = 0$, the above is autoregressive (AR); when $p = 0$, the above is moving average (MA). The user is allowed to provide upper and lower limits for p and q (see below), and the training process will choose the optimal p and q that fall into the user-specified range. When training ARMA, the input needs to be a synchronized HistorySet. For unsynchronized data, use PostProcessor methods to synchronize the data before training ARMA.

The ARMA model implemented allows an option to use Fourier series to detrend the time series before fitting to ARMA model to train. The Fourier trend will be stored in the trained ARMA model for data generation. The following equation describes the detrend process.

$$x_t = y_t - \sum_{m} \left\{ \sum_{k=1}^{K_m} a_k \sin(2\pi k f_m t) + \sum_{k=1}^{K_m} b_k \cos(2\pi k f_m t) \right\},$$

where K_m and f_m are user-defined parameters.

In order to use this Reduced Order Model, the **<ROM>** attribute **subType** needs to be **'ARMA'** (see the example below). This model can be initialized with the following child:

- **<pivotParameter>**, ***string, optional field***, defines the pivot variable (e.g., time) that is non-decreasing in the input HistorySet.
  Default: Time

- **<Features>**, ***string, required field***, defines the features (e.g., scaling). Note that only one feature is allowed for **'ARMA'** and in current implementation this is used for evaluation only.

- **<Target>**, ***string, required field***, defines the variables of the time series.

- **<reseedCopies>**, ***boolean, optional field***, if True then whenever the ARMA is copied, a random reseeding will be performed to ensure different histories.
  Default: True

- **<Pmax>**, ***integer, optional field***, defines the maximum value of p.
  Default: 3

- **<Pmin>**, ***integer, optional field***, defines the minimum value of p.
  Default: 0

- **<Qmax>**, ***integer, optional field***, defines the maximum value of q.
  Default: 3

- **<Qmin>**, ***integer, optional field***, defines the minimum value of q.
  Default: 0

- **<Fourier>**, ***integers, optional field***, must be positive integers. This defines the based period (with unit of second) that would be used for Fourier detrending, i.e., this field defines $1/f_m$ in the above equation. When this filed is not specified, the ARMA considers no Fourier detrend.

- **<FourierOrder>**, ***integers, optional field***, must be positive integers. The number of integers specified in this field should be exactly same as the number of base periods specified in the node **<Fourier>**. This field defines K_m in the above equation.

- **<outTruncation>**, ***string, optional field***, defines whether and how the output time series is truncated. Currently available options are: positive, negative.
  Default: None

Example:

```
<Simulation>
  ...
  <Models>
    ...
    <ROM name='aUserDefinedName' subType='ARMA'>
```

```
      <pivotParameter>Time</pivotParameter>
      <Features>scaling</Features>
      <Target>Speed1,Speed2</Target>
      <Pmax>5</Pmax>
      <Pmin>1</Pmin>
      <Qmax>4</Qmax>
      <Qmin>1</Qmin>
      <Fourier>604800,86400</Fourier>
      <FourierOrder>2, 4</FourierOrder>
     </ROM>
    ...
  </Models>
  ...
</Simulation>
```

17.4 External Model

As the name suggests, an external model is an entity that is embedded in the RAVEN code at run time. This object allows the user to create a python module that is going to be treated as a predefined internal model object. In other words, the **External Model** is going to be treated by RAVEN as a normal external Code (e.g. it is going to be called in order to compute an arbitrary quantity based on arbitrary input).

The specifications of an External Model must be defined within the XML block **<ExternalModel>**. This XML node needs to contain the attributes:

- **name**, ***required string attribute***, user-defined name of this External Model. **Note:** As with the other objects, this is the name that can be used to refer to this specific entity from other input blocks in the XML.
- **subType**, ***required string attribute***, must be kept empty.
- **ModuleToLoad**, ***required string attribute***, file name with its absolute or relative path. **Note:** If a relative path is specified, the code first checks relative to the working directory, then it checks with respect to where the user runs the code. Using the relative path with respect to where the code is run is not recommended.

In order to make the RAVEN code aware of the variables the user is going to manipulate/use in her/his own python Module, the variables need to be specified in the **<ExternalModel>** input block. The user needs to input, within this block, only the variables that RAVEN needs to be aware of (i.e. the variables are going to directly be used by the code) and not the local variables that the user does not want to, for example, store in a RAVEN internal object. These variables are specified within a **<variables>** block:

- **<variables>**, ***string, required parameter***. Comma-separated list of variable names. Each variable name needs to match a variable used/defined in the external python model.

In addition, if the user wants to use the alias system, the following XML block can be inputted:

- **<alias>** ***string, optional field*** specifies alias for any variable of interest in the input or output space for the ExternalModel. These aliases can be used anywhere in the RAVEN input to refer to the ExternalModel variables. In the body of this node the user specifies the name of the variable that the model is going to use (during its execution). The actual alias, usable throughout the RAVEN input, is instead defined in the **variable** attribute of this tag.
  The user can specify aliases for both the input and the output space. As sanity check, RAVEN requires an additional required attribute **type**. This attribute can be either "input" or "output". **Note:** The user can specify as many aliases as needed.
  Default: None

When the external function variables are defined, at run time, RAVEN initializes them and tracks their values during the simulation. Each variable defined in the **<ExternalModel>** block is available in the module (each method implemented) as a python "self."

In the External Python module, the user can implement all the methods that are needed for the functionality of the model, but only the following methods, if present, are called by the framework:

- **def _readMoreXML**, ***OPTIONAL METHOD***, can be implemented by the user if the XML input that belongs to this External Model needs to be extended to contain other information. The information read needs to be stored in "self" in order to be available to all the other methods (e.g. if the user needs to add a couple of newer XML nodes with information needed by the algorithm implemented in the "run" method).
- **def initialize**, ***OPTIONAL METHOD***, can implement all the actions need to be performed at the initialization stage.
- **def createNewInput**, ***OPTIONAL METHOD***, creates a new input with the information coming from the RAVEN framework. In this function the user can retrieve the information coming from the RAVEN framework, during the employment of a calculation flow, and use them to construct a new input that is going to be transferred to the "run" method.
- **def run**, ***REQUIRED METHOD***, is the actual location where the user needs to implement the model action (e.g. resolution of a set of equations, etc.). This function is going to receive the Input (or Inputs) generated either by the External Model "createNewInput" method or the internal RAVEN one.

In the following sub-sections, all the methods are going to be analyzed in detail.

17.4.1 Method: def _readMoreXML

As already mentioned, the **readMoreXML** method can be implemented by the user if the XML input that belongs to this External Model needs to be extended to contain other information. The information read needs to be stored in "self" in order to be available to all the other methods (e.g. if the user needs to add a couple of newer XML nodes with information needed by the algorithm implemented in the "run" method). If this method is implemented in the **External Model**, RAVEN is going to call it when the node **<ExternalModel>** is found parsing the XML input file. The method receives from RAVEN an attribute of type "xml.etree.ElementTree", containing all the sub-nodes and attribute of the XML block **<ExternalModel>**.

Example XML:

```
<Simulation>
  ...
  <Models>
     ...
    <ExternalModel name='AnExtModule' subType=''
       ModuleToLoad='path_to_external_module'>
       <variables>sigma,rho,outcome</variables>
       <!--
          here we define other XML nodes RAVEN does not read
             automatically.
          We need to implement, in the external module
             'AnExtModule' the readMoreXML method
        -->
        <newNodeWeNeedToRead>
            whatNeedsToBeRead
        </newNodeWeNeedToRead>
    </ExternalModel>
     ...
  </Models>
  ...
</Simulation>
```

Corresponding Python function:

```
def _readMoreXML(self,xmlNode):
  # the xmlNode is passed in by RAVEN framework
  # <newNodeWeNeedToRead> is unknown (in the RAVEN framework)
  # we have to read it on our own
  # get the node
  ourNode = xmlNode.find('newNodeWeNeedToRead')
  # get the information in the node
```

```
self.ourNewVariable = ourNode.text
# end function
```

17.4.2 Method: `def initialize`

The **initialize** method can be implemented in the **External Model** in order to initialize some variables needed by it. For example, it can be used to compute a quantity needed by the "run" method before performing the actual calculation). If this method is implemented in the **External Model**, RAVEN is going to call it at the initialization stage of each "Step" (see section 20. RAVEN will communicate, thorough a set of method attributes, all the information that are generally needed to perform a initialization:

- runInfo, a dictionary containing information regarding how the calculation is set up (e.g. number of processors, etc.). It contains the following attributes:
  - `DefaultInputFile` – default input file to use
  - `SimulationFiles` – the xml input file
  - `ScriptDir` – the location of the pbs script interfaces
  - `FrameworkDir` – the directory where the framework is located
  - `WorkingDir` – the directory where the framework should be running
  - `TempWorkingDir` – the temporary directory where a simulation step is run
  - `NumMPI` – the number of mpi process by run
  - `NumThreads` – number of threads by run
  - `numProcByRun` – total number of core used by one run (number of threads by number of mpi)
  - `batchSize` – number of contemporaneous runs
  - `ParallelCommand` – the command that should be used to submit jobs in parallel (mpi)
  - `numNode` – number of nodes
  - `procByNode` – number of processors by node
  - `totalNumCoresUsed` – total number of cores used by driver
  - `queueingSoftware` – queueing software name
  - `stepName` – the name of the step currently running
  - `precommand` – added to the front of the command that is run
  - `postcommand` – added after the command that is run

- `delSucLogFiles` – if a simulation (code run) has not failed, delete the relative log file (if True)
- `deleteOutExtension` – if a simulation (code run) has not failed, delete the relative output files with the listed extension (comma separated list, for example: 'e,r,txt')
- `mode` – running mode, curently the only mode supported is mpi (but custom modes can be created)
- *expectedTime* – how long the complete input is expected to run
- *logfileBuffer* – logfile buffer size in bytes

- inputs, a list of all the inputs that have been specified in the "Step" using this model.

In the following an example is reported:

```
def initialize(self,runInfo,inputs):
 # Let's suppose we just need to initialize some variables
  self.sigma = 10.0
  self.rho   = 28.0
  # end function
```

17.4.3 Method: `def createNewInput`

The **createNewInput** method can be implemented by the user to create a new input with the information coming from the RAVEN framework. In this function, the user can retrieve the information coming from the RAVEN framework, during the employment of a calculation flow, and use them to construct a new input that is going to be transferred to the "run" method. The new input created needs to be returned to RAVEN (i.e. "return NewInput").
This method expects that the new input is returned in a Python "dictionary". RAVEN communicates, thorough a set of method attributes, all the information that are generally needed to create a new input:

- `inputs`, ***python list***, a list of all the inputs that have been defined in the "Step" using this model.
- `samplerType`, ***string***, the type of Sampler, if a sampling strategy is employed; will be None otherwise.
- `Kwargs`, ***dictionary***, a dictionary containing several pieces of information (that can change based on the "Step" type). If a sampling strategy is employed, this dictionary contains another dictionary identified by the keyword "SampledVars", in which the variables perturbed by the sampler are reported.

Note: If the "Step" that is using this Model has as input(s) an object of main class type "DataObjects" (see Section 14), the internal "createNewInput" method is going to convert it in a dictionary of values.

Here we present an example:

```
def createNewInput(self,inputs,samplerType,**Kwargs):
  # in here the actual createNewInput of the
  # model is implemented
  if samplerType == 'MonteCarlo':
    avariable = inputs['something']*inputs['something2']
  else:
    avariable = inputs['something']/inputs['something2']
  return avariable*Kwargs['SampledVars']['aSampledVar']
```

17.4.4 Method: `def run`

As stated previously, the only method that *must* be present in an External Module is the **run** function. In this function, the user needs to implement the algorithm that RAVEN will execute. The **run** method is generally called after having inquired the "createNewInput" method (either the internal or the user-implemented one). The only attribute this method is going to receive is a Python list of inputs (the inputs coming from the `createNewInput` method). If the user wants RAVEN to collect the results of this method, the outcomes of interest need to be stored in "self." **Note:** RAVEN is trying to collect the values of the variables listed only in the **<ExternalModel>** XML block.

In the following an example is reported:

```
def run(self,Input):
  # in here the actual run of the
  # model is implemented
  input = Input[0]
  self.outcome = self.sigma*self.rho*input[``whatEver'']
```

17.5 PostProcessor

A Post-Processor (PP) can be considered as an action performed on a set of data or other type of objects. Most of the post-processors contained in RAVEN, employ a mathematical operation on the data given as "input". RAVEN supports several different types of PPs.

Currently, the following types are available in RAVEN:

- **BasicStatistics**
- **ComparisonStatistics**
- **ImportanceRank**
- **SafestPoint**
- **LimitSurface**
- **LimitSurfaceIntegral**
- **External**
- **TopologicalDecomposition**
- **RavenOutput**
- **DataMining**
- **Metric**
- **CrossValidation**

The specifications of these types must be defined within the XML block **<PostProcessor>**. This XML node needs to contain the attributes:

- **name**, ***required string attribute***, user-defined identifier of this post-processor. **Note:** As with other objects, this is the name that can be used to refer to this specific entity from other input XML blocks.
- **subType**, ***required string attribute***, defines which of the post-processors needs to be used, choosing among the previously reported types. This choice conditions the subsequent required and/or optional **<PostProcessor>** sub nodes.

As already mentioned, all the types and meaning of the remaining sub-nodes depend on the post-processor type specified in the attribute **subType**. In the following sections the specifications of each type are reported.

17.5.1 BasicStatistics

The **BasicStatistics** post-processor is the container of the algorithms to compute many of the most important statistical quantities. It is important to notice that this post-processor can accept as input both ***PointSet*** and ***HistorySet*** data objects, depending on the type of statistics the user wants to compute:

- ***PointSet***: Static Statistics;
- ***HistorySet***: Dynamic Statistics. Depending on a "pivot parameter" (e.g. time) the post-processor is going to compute the statistics for each value of it (e.g. for each time step). In case an **HistorySet** is provided as Input, the Histories needs to be synchronized (use ***Interfaced*** post-processor of type **HistorySetSync**).

In order to use the *BasicStatistics post-processor* PP, the user needs to set the **subType** of a **<PostProcessor>** node:

<PostProcessor subType='BasicStatistics'/>.

Several sub-nodes are available:

- **<(metric)>**, ***comma separated string or node list, required field***, specifications for the metric to be calculated. The name of each node is the requested metric. There are two forms for specifying the requested parameters of the metric. For scalar values such as **<expectedValue>** and **<variance>**, the text of the node is a comma-separated list of the parameters for which the metric should be calculated. For matrix values such as **<sensitivty>** and **<covariance>**, the matrix node requires two sub-nodes, **<targets>** and **<features>**, each of which is a comma-separated list of the targets for which the metric should be calculated, and the features for which the metric should be calculated for that target. See the example below.

  Note: When defining the metrics to use, it is possible to have multiple nodes with the same name. For example, if a problem has inputs W, X, Y, and Z, and the responses are A, B, and C, it is possible that the desired metrics are the **<sensitivity>** of A and B to X and Y, as well as the **<sensitivity>** of C to W and Z, but not the sensitivity of A to W. In this event, two copies of the **<sensitivity>** node are added to the input. The first has targets A, B and features X, Y, while the second node has target C and features W, Z. This could reduce some computation effort in problems with many responses or inputs. An example of this is shown below.
  Currently the scalar quantities available for request are:

  – **expectedValue**: expected value or mean
  – **minimum**: The minimum value of the samples.
  – **maximum**: The maximum value of the samples.
  – **median**: median
  – **variance**: variance
  – **sigma**: standard deviation
  – **percentile**: the percentile. If this quantity is inputted as *percentile* the 5% and 95% percentile(s) are going to be computed. Otherwise the user can specify this quantity with a parameter *percent='X'*, where the *X* represents the requested percentile (a floating point value between 0.0 and 100.0)
  – **variationCoefficient**: coefficient of variation, i.e. **sigma/expectedValue**. **Note:** If the **expectedValue** is zero, the **variationCoefficient** will be **INF**.
  – **skewness**: skewness
  – **kurtosis**: excess kurtosis (also known as Fisher's kurtosis)

The matrix quantities available for request are:

- **sensitivity**: matrix of sensitivity coefficients, computed via linear regression method.
- **covariance**: covariance matrix
- **pearson**: matrix of correlation coefficients
- **NormalizedSensitivity**: matrix of normalized sensitivity coefficients. **Note:** It is the matrix of normalized VarianceDependentSensitivity
- **VarianceDependentSensitivity**: matrix of sensitivity coefficients dependent on the variance of the variables
- **samples**: the number of samples in the data set used to determine the statistics.

If all the quantities need to be computed, this can be done through the `<all>` node, which requires the `<targets>` and `<features>` sub-nodes.

Note: If the weights are present in the system then weighted quantities are calculated automatically. In addition, if a matrix quantity is requested (e.g. Covariance matrix, etc.), only the weights in the output space are going to be used for both input and output space (the computation of the joint probability between input and output spaces is not implemented yet).
Note: Certain ROMs provide their own statistical information (e.g., those using the sparse grid collocation sampler such as: `'GaussPolynomialRom'` and `'HDMRRom'`) which can be obtained by printing the ROM to file (xml). For these ROMs, computing the basic statistics on data generated from one of these sampler/ROM combinations may not provide the information that the user expects.

- `<pivotParameter>`, ***string, optional field***, name of the parameter that needs to be used for the computation of the Dynamic BasicStatistics (e.g. time). This node needs to be inputted just in case an **HistorySet** is used as Input. It represents the reference monotonic variable based on which the statistics is going to be computed (e.g. time-dependent statistical moments).
Default: None

- `<biased>`, ***string (boolean), optional field***, if *True* biased quantities are going to be calculated, if *False* unbiased.
Default: False

- `<methodsToRun>`, ***comma separated string, optional field***, specifies the method names of an external Function that need to be run before computing any of the predefined quantities. If this XML node is specified, the `<Function>` node must be present.
Default: None

- **Assembler Objects** This object is required in case the `<methodsToRun>` node is specified. The object must be listed with a rigorous syntax that, except for the xml node tag, is

common among all the objects. Each of these nodes must contain 2 attributes that are used to map those within the simulation framework:

- `class`, ***required string attribute***, it is the main "class" the listed object is from;
- `type`, ***required string attribute***, it is the object identifier or sub-type.

The **BasicStatistics** post-processor approach optionally accepts the following object type:

- **<Function>**, ***string, required field***, The body of this xml block needs to contain the name of an External Function defined within the **<Functions>** main block (see section 18). This object needs to contain the methods listed in the node **<methodsToRun>**.

Example (Static Statistics): This example demonstrates how to request the expected value of `'x01'` and `'x02'`, along with the sensitivity of both `'x01'` and `'x02'` to `'a'` and `'b'`.

```
<Simulation>
  ...
  <Models>
    ...
    <PostProcessor name='aUserDefinedName'
       subType='BasicStatistics' verbosity='debug'>
      <expectedValue>x01,x02</expectedValue>
      <sensitivity>
        <targets>x01,x02</targets>
        <features>a,b</features>
      </sensitivity>
      <methodsToRun>failureProbability</methodsToRun>
    </PostProcessor>
    ...
  </Models>
  ...
</Simulation>
```

Example (Static using <all>): This example is similar to the one above, but shows using the **<all>** node.

```
<Simulation>
  ...
  <Models>
    ...
    <PostProcessor name='aUserDefinedName'
       subType='BasicStatistics' verbosity='debug'>
      <all>
        <targets>x01,x02</targets>
```

```
        <features>a,b</features>
      </all>
    </PostProcessor>
    ...
  </Models>
  ...
</Simulation>
```

Example (Static, multiple matrix nodes): This example shows how multiple nodes can specify particular metrics multiple times to include different target/feature combinations. This postprocessor calculates the expected value of A, B, and C, as well as the sensitivity of both A and B to X and Y as well as the sensitivity of C to W and Z.

```
<Simulation>
  ...
  <Models>
    ...
    <PostProcessor name='aUserDefinedName'
       subType='BasicStatistics' verbosity='debug'>
      <expectedValue>A,B,C</expectedValue>
      <sensitivity>
        <targets>A,B</targets>
        <features>x,y</features>
      </sensitivity>
      <sensitivity>
        <targets>C</targets>
        <features>w,z</features>
      </sensitivity>
    </PostProcessor>
    ...
  </Models>
  ...
</Simulation>
```

Example (Dynamic Statistics):

```
<Simulation>
  ...
  <Models>
    ...
    <PostProcessor name='aUserDefinedNameForDynamicPP'
       subType='BasicStatistics' verbosity='debug'>
      <expectedValue>x01,x02</expectedValue>
```

```
      <sensitivity>
        <targets>x01,x02</targets>
        <features>a,b</features>
      </sensitivity>
      <methodsToRun>failureProbability</methodsToRun>
      <pivotParameter>time</pivotParameter>
    </PostProcessor>
    ...
  </Models>
  ...
</Simulation>
```

17.5.2 ComparisonStatistics

The **ComparisonStatistics** post-processor computes statistics for comparing two different dataObjects. This is an experimental post-processor, and it will definitely change as it is further developed.

There are four nodes that are used in the post-processor.

- <kind>: specifies information to use for comparing the data that is provided. This takes either uniformBins which makes the bin width uniform or equalProbability which makes the number of counts in each bin equal. It can take the following attributes:
  - numBins which takes a number that directly specifies the number of bins
  - binMethod which takes a string that specifies the method used to calculate the number of bins. This can be either square-root or sturges.
- <compare>: specifies the data to use for comparison. This can either be a normal distribution or a dataObjects:
  - <data>: This will specify the data that is used. The different parts are separated by |'s.
  - <reference>: This specifies a reference distribution to be used. It takes distribution to use that is defined in the distributions block. A name parameter is used to tell which distribution is used.
- <fz>: If the text is true, then extra comparison statistics for using the f_z function are generated. These take extra time, so are not on by default.
- <interpolation>: This switches the interpolation used for the cdf and the pdf functions between the default of quadratic or linear.

The **ComparisonStatistics** post-processor generates a variety of data. First for each data provided, it calculates bin boundaries, and counts the numbers of data points in each bin. From the numbers in each bin, it creates a cdf function numerically, and from the cdf takes the derivative to generate a pdf. It also calculates statistics of the data such as mean and standard deviation. The post-processor can generate a CSV file only.

The post-processor uses the generated pdf and cdf function to calculate various statistics. The first is the cdf area difference which is:

$$cdf_area_difference = \int_{-\infty}^{\infty} \|CDF_a(x) - CDF_b(x)\| dx \tag{27}$$

This given an idea about how far apart the two pieces of data are, and it will have units of x.

The common area between the two pdfs is calculated. If there is perfect overlap, this will be 1.0, if there is no overlap, this will be 0.0. The formula used is:

$$pdf_common_area = \int_{-\infty}^{\infty} \min(PDF_a(x), PDF_b(x)) dx \tag{28}$$

The difference pdf between the two pdfs is calculated. This is calculated as:

$$f_Z(z) = \int_{-\infty}^{\infty} f_X(x) f_Y(x - z) dx \tag{29}$$

This produces a pdf that contains information about the difference between the two pdfs. The mean can be calculated as (and will be calculated only if fz is true):

$$\bar{z} = \int_{-\infty}^{\infty} z f_Z(z) dz \tag{30}$$

The mean can be used to get an signed difference between the pdfs, which shows how their means compare.

The variance of the difference pdf can be calculated as (and will be calculated only if fz is true):

$$var = \int_{-\infty}^{\infty} (z - \bar{z})^2 f_Z(z) dz \tag{31}$$

The sum of the difference function is calculated if fz is true, and is:

$$sum = \int_{-\infty}^{\infty} f_z(z) dz \tag{32}$$

This should be 1.0, and if it is different that points to approximations in the calculation.

Example:

```
<Simulation>
   ...
   <Models>
      ...
      <PostProcessor name="stat_stuff"
         subType="ComparisonStatistics">
      <kind binMethod='sturges'>uniformBins</kind>
      <compare>
        <data>OriData|Output|tsin_TEMPERATURE</data>
        <reference name='normal_410_2' />
      </compare>
      <compare>
        <data>OriData|Output|tsin_TEMPERATURE</data>
        <data>OriData|Output|tsout_TEMPERATURE</data>
      </compare>
      </PostProcessor>
      <PostProcessor name="stat_stuff2"
         subType="ComparisonStatistics">
        <kind numBins="6">equalProbability</kind>
        <compare>
          <data>OriData|Output|tsin_TEMPERATURE</data>
        </compare>
        <Distribution class='Distributions'
           type='Normal'>normal_410_2</Distribution>
      </PostProcessor>
      ...
   </Models>
   ...
   <Distributions>
      <Normal name='normal_410_2'>
         <mean>410.0</mean>
         <sigma>2.0</sigma>
      </Normal>
   </Distributions>
</Simulation>
```

17.5.3 ImportanceRank

The **ImportanceRank** post-processor is specifically used to compute sensitivity indices and importance indices with respect to input parameters associated with multivariate normal distributions.

In addition, the user can also request the transformation matrix and the inverse transformation matrix when the PCA reduction is used. In order to use the *ImportanceRank* PP, the user needs to set the **subType** of a **<PostProcessor>** node:

<PostProcessor subType='ImportanceRank'/>.

Several sub-nodes are available:

- **<what>**, ***comma separated string, required field***, List of quantities to be computed. Currently the quantities available are:
  - **'SensitivityIndex'**: used to measure the impact of sensitivities on the model.
  - **'ImportanceIndex'**: used to measure the impact of sensitivities and input uncertainties on the model.
  - **'PCAIndex'**: the indices of principal component directions, used to measure the impact of principal component directions on input covariance matrix. **Note:** **'PCAIndex'** can be only requested when subnode **<latent>** is defined in **<features>**.
  - **'transformation'**: the transformation matrix used to map the latent variables to the manifest variables in the original input space.
  - **'InverseTransformation'**: the inverse transformation matrix used to map the manifest variables to the latent variables in the transformed space.
  - **'ManifestSensitivity'**: the sensitivity coefficients of **<target>** with respect to **<manifest>** variables defined in **<features>**.

    Note: In order to request **'transformation'** matrix or **'InverseTransformation'** matrix or **'ManifestSensitivity'**, the subnodes **<latent>** and **<manifest>** under **<features>** are required (more details can be found in the following).

  If all the quantities need to be computed, the user can input in the body of **<what>** the string **'all'**. **Note:** **'all'** equivalent to **'SensitivityIndex, ImportanceIndex, PCAIndex'**.

  Since the transformation and InverseTransformation matrix can be very large, they are not printed with option **'all'**. In order to request the transformation matrix (or inverse transformation matrix) from this post processor, the user need to specify **'transformation'** or **'InverseTransformation'** in **<what>**. In addition, both **<manifest>** and **<latent>** subnodes are required and should be defined in node **<features>**. For example, let $\mathbf{L}, \mathbf{P}$ represent the transformation and inverse transformation matrices, respectively. We will define vectors $\mathbf{x}$ as manifest variables and vectors $\mathbf{y}$ as latent variables. If a absolute covariance matrix is used in given distribution, the following equation will be used:

  $\delta\mathbf{x} = \mathbf{L} * \mathbf{y}$

  $\mathbf{y} = \mathbf{P} * \delta\mathbf{x}$

If a relative covariance matrix is used in given distribution, the following equation will be used:

$\frac{\delta \mathbf{x}}{\mu} = \mathbf{L} * \mathbf{y}$

$\mathbf{y} = \mathbf{P} * \frac{\delta \mathbf{x}}{\mu}$

where $\delta\mathbf{x}$ denotes the changes in the input vector $\mathbf{x}$, and μ denotes the mean values of the input vector $\mathbf{x}$.

- **<features>**, ***XML node, required parameter***, used to specify the information for the input variables. In this xml-node, the following xml sub-nodes need to be specified:
  - **<manifest>**,***XML node, optional parameter***, used to indicate the input variables belongs to the original input space. It can accept the following child node:
    * **<variables>**,***comma separated string, required field***, lists manifest variables.
    * **<dimensions>**, ***comma separated integer, optional field***, lists the dimensions corresponding to the manifest variables. If not provided, the dimensions are determined by the order indices of given manifest variables.
  - **<latent>**,***XML node, optional parameter***, used to indicate the input variables belongs to the transformed space. It can accept the following child node:
    * **<variables>**,***comma separated string, required field***, lists latent variables.
    * **<dimensions>**, ***comma separated integer, optional field***, lists the dimensions corresponding to the latent variables. If not provided, the dimensions are determined by the order indices of given latent variables.

  Note: At least one of the subnodes, i.e. **<manifest>** and **<latent>** needs to be specified.
- **<targets>**, ***comma separated string, required field***, lists output responses.
- **<mvnDistribution>**, ***string, required field***, specifies the multivariate normal distribution name. The **<MultivariateNormal>** node must be present.

Here is an example to show the user how to request the transformation matrix, the inverse transformation matrix, the manifest sensitivities and other quantities.

Example:

```
<Simulation>
  ...
  <Models>
    ...
    <PostProcessor name='aUserDefinedName'
       subType='ImportanceRank'>
```

```xml
      <what>SensitivityIndex,ImportanceIndex,Transformation,
         InverseTransformation,ManifestSensitivity</what>
      <features>
        <manifest>
          <variables>x1,x2</variables>
          <dimensions>1,2</dimensions>
        </manifest>
        <latent>
          <variables>latent_1, latent_2</variables>
          <dimensions>1,2</dimensions>
        </latent>
      </features>
      <targets>y1,y2</targets>
      <mvnDistribution>MVN</mvnDistribution>
    </PostProcessor>
    ...
  </Models>
  ...
</Simulation>
```

17.5.4 SafestPoint

The **SafestPoint** post-processor provides the coordinates of the farthest point from the limit surface that is given as an input. The safest point coordinates are expected values of the coordinates of the farthest points from the limit surface in the space of the "controllable" variables based on the probability distributions of the "non-controllable" variables.

The term "controllable" identifies those variables that are under control during the system operation, while the "non-controllable" variables are stochastic parameters affecting the system behaviour randomly.

The "SafestPoint" post-processor requires the set of points belonging to the limit surface, which must be given as an input. The probability distributions as "Assembler Objects" are required in the "Distribution" section for both "controllable" and "non-controllable" variables.

The sampling method used by the "SafestPoint" is a "value" or "CDF" grid. At present only the "equal" grid type is available.

In order to use the *Safest Point* PP, the user needs to set the **subType** of a **<PostProcessor>** node:

<PostProcessor subType='SafestPoint'/>.

Several sub-nodes are available:

- **<Distribution>**, ***Required***, represents the probability distributions of the "controllable" and "non-controllable" variables. These are **Assembler Objects**, each of these nodes must contain 2 attributes that are used to identify those within the simulation framework:
  - **class**, ***required string attribute***, is the main "class" the listed object is from.
  - **type**, ***required string attribute***, is the object identifier or sub-type.
- **<controllable>** lists the controllable variables. Each variable is associated with its name and the two items below:
  - **<distribution>** names the probability distribution associated with the controllable variable.
  - **<grid>** specifies the **type**, **steps**, and tolerance of the sampling grid.
- **<non-controllable>** lists the non-controllable variables. Each variable is associated with its name and the two items below:
  - **<distribution>** names the probability distribution associated with the non-controllable variable.
  - **<grid>** specifies the **type**, **steps**, and tolerance of the sampling grid.

Example:

```
<Simulation>
 ...
  <Models>
  ...
  <PostProcessor name='SP' subType='SafestPoint'>
    <Distribution  class='Distributions'
       type='Normal'>x1_dst</Distribution>
    <Distribution  class='Distributions'
       type='Normal'>x2_dst</Distribution>
    <Distribution  class='Distributions'
       type='Normal'>gammay_dst</Distribution>
    <controllable>
      <variable name='x1'>
        <distribution>x1_dst</distribution>
        <grid type='value' steps='20'>1</grid>
      </variable>
      <variable name='x2'>
        <distribution>x2_dst</distribution>
```

```
          <grid type='value' steps='20'>1</grid>
        </variable>
      </controllable>
      <non-controllable>
        <variable name='gammay'>
          <distribution>gammay_dst</distribution>
          <grid type='value' steps='20'>2</grid>
        </variable>
      </non-controllable>
    </PostProcessor>
    ...
  </Models>
  ...
</Simulation>
```

17.5.5 LimitSurface

The **LimitSurface** post-processor is aimed to identify the transition zones that determine a change in the status of the system (Limit Surface).

In order to use the *LimitSurface* PP, the user needs to set the `subType` of a `<PostProcessor>` node:

`<PostProcessor subType='LimitSurface'/>`.

Several sub-nodes are available:

- `<parameters>`, ***comma separated string, required field***, lists the parameters that define the uncertain domain and from which the LS needs to be computed.
- `<tolerance>`, ***float, optional field***, sets the absolute value (in CDF) of the convergence tolerance. This value defines the coarseness of the evaluation grid.
  Default: 1.0e-4
- `<side>`, ***string, optional field***, in this node the user can specify which side of the limit surface needs to be computed. Three options are available:
  negative, Limit Surface corresponding to the goal function value of "-1";
  positive, Limit Surface corresponding to the goal function value of "1";
  both, either positive and negative Limit Surface is going to be computed.
  Default: negative
- **Assembler Objects** These objects are either required or optional depending on the functionality of the Adaptive Sampler. The objects must be listed with a rigorous syntax that, except

for the xml node tag, is common among all the objects. Each of these nodes must contain 2 attributes that are used to map those within the simulation framework:

- `class`, ***required string attribute***, is the main "class" of the listed object. For example, it can be "Models," "Functions," etc.
- `type`, ***required string attribute***, is the object identifier or sub-type. For example, it can be "ROM," "External," etc.

The **LimitSurface** post-processor requires or optionally accepts the following objects' types:

- `<ROM>`, ***string, optional field***, body of this xml node must contain the name of a ROM defined in the `<Models>` block (see section 17.3).
- `<Function>`, ***string, required field***, the body of this xml block needs to contain the name of an External Function defined within the `<Functions>` main block (see section 18). This object represents the boolean function that defines the transition boundaries. This function must implement a method called *__residuumSign(self)*, that returns either -1 or 1, depending on the system conditions (see section 18).

Example:

```
<Simulation>
 ...
 <Models>
  ...
    <PostProcessor name="computeLimitSurface"
       subType='LimitSurface' verbosity='debug'>
      <parameters>x0,y0</parameters>
      <ROM class='Models' type='ROM'>Acc</ROM>
      <!-- Here, you can add a ROM defined in Models block.
           If it is not Present, a nearest neighbor algorithm
           will be used.
       -->
      <Function class='Functions' type='External'>
        goalFunctionForLimitSurface
      </Function>
    </PostProcessor>
    ...
  </Models>
  ...
</Simulation>
```

17.5.6 LimitSurfaceIntegral

The **LimitSurfaceIntegral** post-processor is aimed to compute the likelihood (probability) of the event, whose boundaries are represented by the Limit Surface (either from the LimitSurface post-processor or Adaptive sampling strategies). The inputted Limit Surface needs to be, in the **Post-Process** step, of type **PointSet** and needs to contain both boundary sides (-1.0, +1.0).
The **LimitSurfaceIntegral** post-processor accepts as outputs both files (CSV) and/or **PointSet**s.

In order to use the *LimitSurfaceIntegral* PP, the user needs to set the **subType** of a **<PostProcessor>** node:

<PostProcessor subType='LimitSurfaceIntegral'/>.

Several sub-nodes are available:

- **<variable>**, ***XML node, required parameter*** will specify one attribute:
  - **name**, ***required string attribute***, user-defined name of this variable.

  This **<variable>** recognizes the following child node:
  - **<distribution>**, ***string, optional field***, name of the distribution that is associated to this variable. Its name needs to be contained in the **<Distributions>** block explained in Section 11. If this node is not present, the **<lowerBound>** and **<upperBound>** XML nodes must be inputted.
  - **<lowerBound>**, ***float, optional field***, lower limit of integration domain for this dimension (variable). If this node is not present, the **<distribution>** XML node must be inputted.
  - **<upperBound>**, ***float, optional field***, upper limit of integration domain for this dimension (variable). If this node is not present, the **<distribution>** XML node must be inputted.
- **<tolerance>**, ***float, optional field***, specifies the tolerance for numerical integration confidence.
  Default: 1.0e-4
- **<integralType>**, ***string, optional field***, specifies the type of integrations that need to be used. Currently only MonteCarlo integration is available
  Default: MonteCarlo
- **<seed>**, ***integer, optional field***, specifies the random number generator seed.
  Default: 20021986

- <target>, ***string, optional field***, specifies the target name that represents the $f(\bar{x})$ that needs to be integrated.
  Default: last output found in the inputted PointSet

Example:

```
<Simulation>
 ...
 <Models>
  ...
    <PostProcessor name="LimitSurfaceIntegralDistributions"
       subType='LimitSurfaceIntegral'>
        <tolerance>0.0001</tolerance>
        <integralType>MonteCarlo</integralType>
        <seed>20021986</seed>
        <target>goalFunctionOutput</target>
        <variable name='x0'>
          <distribution>x0_distrib</distribution>
        </variable>
        <variable name='y0'>
          <distribution>y0_distrib</distribution>
        </variable>
    </PostProcessor>
    <PostProcessor name="LimitSurfaceIntegralLowerUpperBounds"
       subType='LimitSurfaceIntegral'>
        <tolerance>0.0001</tolerance>
        <integralType>MonteCarlo</integralType>
        <seed>20021986</seed>
        <target>goalFunctionOutput</target>
        <variable name='x0'>
          <lowerBound>-2.0</lowerBound>
          <upperBound>12.0</upperBound>
        </variable>
        <variable name='y0'>
            <lowerBound>-1.0</lowerBound>
            <upperBound>11.0</upperBound>
        </variable>
    </PostProcessor>
    ...
  </Models>
  ...
</Simulation>
```

17.5.7 External

The **External** post-processor will execute an arbitrary python function defined externally using the *Functions* interface (see Section 18 for more details).

In order to use the *External* PP, the user needs to set the `subType` of a `<PostProcessor>` node:

`<PostProcessor subType='External'/>`.

Several sub-nodes are available:

- `<method>`, ***comma separated string, required field***, lists the method names of an external Function that will be computed (each returning a post-processing value). The name of the method represents a new variable that can be stored in a new ***DataObjects*** entity.
- `<Function>`, ***xml node, required string field***, specifies the name of a Function where the *methods* listed above are defined. **Note:** This name should match one of the Functions defined in the `<Functions>` block of the input file. The objects must be listed with a rigorous syntax that, except for the XML node tag, is common among all the objects. Each of these sub-nodes must contain 2 attributes that are used to map them within the simulation framework:
  - `class`, ***required string attribute***, is the main "class" the listed object is from, the only acceptable class for this post-processor is `'Functions'`;
  - `type`, ***required string attribute***, is the object identifier or sub-type, the only acceptable type for this post-processor is `'External'`.

This Post-Processor accepts as Input/Output both `'PointSet'` and `'HistorySet'`:

- If a `'PointSet'` is used as Input, the parameters are passed in the external `'Function'` as numpy arrays. The methods' return type must be either a new array or a scalar. In the following it is reported an example with two methods, one that returns a scalar and the other one that returns an array:

```
import numpy as np
def sum(self):
  return np.sum(self.aParameterInPointSet)

def sumTwoArraysAndReturnAnotherone(self):
  return self.aParamInPointSet1+self.aParamInPointSet2
```

- If a **'HistorySet'** is used as Input, the parameters are passed in the external **'Function'** as a list of numpy arrays. The methods' return type must be either a new list of arrays (if the Output is another **'HistorySet'**), a scalar or a single array (if the Output is **'PointSet'** . In the following it is reported an example with two methods, one that returns a new list of arrays (Output = HistorySet) and the other one that returns an array (Output = PointSet):

```
import numpy as np
def newHistorySetParameter(self):
  x = []*len(self.time)
  for history in range(len(self.time)):
    for ts in range(len(self.time[history])):
      if self.time[history][ts] >= 0.001: break
    x[history] = self.x[history][ts:]
  return x

def aNewPointSetParameter(self):
  x = []*len(self.time)
  for history in range(len(self.time)):
    x[history] = self.x[history][-1]
  return x
```

Example:

```
<Simulation>
  ...
  <Models>
    ...
    <PostProcessor name="externalPP" subType='External'
       verbosity='debug'>
      <method>Delta,Sum</method>
      <Function class='Functions'
         type='External'>operators</Function>
        <!-- Here, you can add a Function defined in the
             Functions block. This should be present or
             else RAVEN will not know where to find the
             defined methods. -->
    </PostProcessor>
    ...
  </Models>
  ...
</Simulation>
```

17.5.8 TopologicalDecomposition

The **TopologicalDecomposition** post-processor will compute an approximated hierarchical Morse-Smale complex which will add two columns to a dataset, namely `minLabel` and `maxLabel` that can be used to decompose a dataset.

The topological post-processor can also be run in 'interactive' mode, that is by passing the keyword `interactive` to the command line of RAVEN's driver. In this way, RAVEN will initiate an interactive UI that allows one to explore the topological hierarchy in real-time and adjust the simplification setting before adjusting a dataset. Use in interactive mode will replace the parameter **<simplification>** described below with whatever setting is set in the UI upon exiting it.

In order to use the **TopologicalDecomposition** post-processor, the user needs to set the attribute **subType**: **<PostProcessor subType='TopologicalDecomposition'>**. The following is a list of acceptable sub-nodes:

- **<graph>** ***, string, optional field***, specifies the type of neighborhood graph used in the algorithm, available options are:
  - `beta skeleton`
  - `relaxed beta skeleton`
  - `approximate knn`

  Default: *`beta skeleton`*
- **<gradient>**, ***string, optional field***, specifies the method used for estimating the gradient, available options are:
  - `steepest`

  Default: *`steepest`*
- **<beta>**, ***float in the range: (0,2], optional field***, is only used when the **<graph>** is set to `beta skeleton` or `relaxed beta skeleton`.
  Default: 1.0
- **<knn>**, ***integer, optional field***, is the number of neighbors when using the **'approximate knn'** for the **<graph>** sub-node and used to speed up the computation of other graphs by using the approximate knn graph as a starting point for pruning. -1 means use a fully connected graph.
  Default: -1

- **<weighted>**, ***boolean, optional***, a flag that specifies whether the regression models should be probability weighted.
  Default: False

- **<interactive>**, if this node is present *and* the user has specified the keyword `interactive` at the command line, then this will initiate a graphical interface for exploring the different simplification levels of the topological hierarchy. Upon exit of the graphical interface, the specified simplification level will be updated to use the last value of the graphical interface before writing any "output" results.

- **<persistence>**, ***string, optional field***, specifies how to define the hierarchical simplification by assigning a value to each local minimum and maximum according to the one of the strategy options below:

  - `difference` - The function value difference between the extremum and its closest-valued neighboring saddle.
  - `probability` - The probability integral computed as the sum of the probability of each point in a cluster divided by the count of the cluster.
  - `count` - The count of points that flow to or from the extremum.

  Default: `difference`

- **<simplification>**, ***float, optional field***, specifies the amount of noise reduction to apply before returning labels.
  Default: 0

- **<parameters>**, ***comma separated string, required field***, lists the parameters defining the input space.

- **<response>**, ***string, required field***, is a single variable name defining the scalar output space.

Example:

```
<Simulation>
  ...
  <Models>
    ...
    <PostProcessor name="***" subType='TopologicalDecomposition'>
      <graph>beta skeleton</graph>
      <gradient>steepest</gradient>
      <beta>1</beta>
      <knn>8</knn>
      <normalization>None</normalization>
```

```
      <parameters>X,Y</parameters>
      <response>Z</response>
      <weighted>true</weighted>
      <simplification>0.3</simplification>
      <persistence>difference</persistence>
    </PostProcessor>
    ...
  <Models>
  ...
<Simulation>
```

17.5.9 DataMining

Knowledge discovery in databases (KDD) is the process of discovering useful knowledge from a collection of data. This widely used data mining technique is a process that includes data preparation and selection, data cleansing, incorporating prior knowledge on data sets and interpreting accurate solutions from the observed results. Major KDD application areas include marketing, fraud detection, telecommunication and manufacturing.

DataMining is the analysis step of the KDD process. The overall of the data mining process is to extract information from a data set and transform it into an understandable structure for further use. The actual data mining task is the automatic or semi-automatic analysis of large quantities of data to extract previously unknown, interesting patterns such as groups of data records (cluster analysis), unusual records (anomaly detection), and dependencies (association rule mining).
In order to use the **DataMining** post-processor, the user needs to set the attribute **subType**:

<PostProcessor subType= 'DataMining'>.

The following is a list of acceptable sub-nodes:

- **<KDD>** ***string,required field***, the subnodes specifies the necessary information for the algorithm to be used in the postprocessor. The **<KDD>** has the required attribute: **lib**, the name of the library the algorithm belongs to. Current algorithms applied in the KDD model is based on SciKit-Learn library. Thus currently there is only one library:

  - **'SciKitLearn'**

  The **<KDD>** has the optional attribute: **labelFeature**, the name associated to labels or dimensions generated by the **DataMining** post-processor. The default name depends on the type of algorithm employed. For clustering and mixture models it is the name of the Post-Processor followed by "Labels" (e.g., if the name of a clustering PostProcessor is "kMeans"

then the default name associated to the labels is "kMeansLabels" if not specified in the attribute `labelFeature`). For decomposition and manifold models, the default names are the name of the PostProcessor followed by "Dimension" and an integer identifier beginning with 1. (e.g., if the name of a dimensionality reduction PostProcessor is "dr" and the user specifies 3 components, then the output dataObject will have three new outputs named "drDimension1," "drDimension2," and "drDimension3.").

17.5.9.1 SciKitLearn

'SciKitLearn' is based on algorithms in SciKit-Learn library, and it performs data mining over PointSet and HistorySet. Note that for HistorySet's **'SciKitLearn'** performs the task given in **<SKLType>** (see below) for each time step, and so only synchronized HistorySet can be used as input to this model. For unsynchronized HistorySet, use **'HistorySetSync'** method in **'Interfaced'** post-processor to synchronize the input data before using **'SciKitLearn'**. The rest of this subsection and following subsection is dedicated to the **'SciKitLearn'** library.

The temporal variable for a HistorySet **'SciKitLearn'** is specified in the **<pivotParameter>** node:

- **<pivotParameter>**, ***string, optional parameter*** specifies the pivot variable (e.g., time, etc) in the input HistorySet.
  Default: None.

The algorithm for the dataMining is chosen by the subnode **<SKLType>** under the parent node **<KDD>**. The format is same as in 17.3.7. However, for the completeness sake, it is repeated here.

The data that are used in the training of the **DataMining** postprocessor are suplied with subnode **<Features>** in the parent node **<KDD>**.

- **<SKLtype>**, ***vertical bar (|) separated string, required field***, contains a string that represents the data mining algorithm to be used. As mentioned, its format is:
  <SKLtype>`mainSKLclass|algorithm`**</SKLtype>** where the first word (before the "`|`" symbol) represents the main class of algorithms, and the second word (after the "`|`" symbol) represents the specific algorithm.
- **<Features>**, ***string, required field***, defines the data to be used for training the data mining algorithm. It can be:
  - the name of the variable in the defined dataObject entity
  - the location (i.e. input or output). In this case the data mining is applied to all the variables in the defined space.

The **<KDD>** node can have either optional or required subnodes depending on the dataMining algorithm used. The possible subnodes will be described separately for each algorithm below. The time dependent clustering data mining algorithms have a **<reOrderStep>** option that will try and keep the same labels on the clusters. The higher the number, the longer the history that the clustering algorithm will look through to maintain the same labeling between time steps.

All the available algorithms are described in the following sections.

17.5.9.2 Gaussian mixture models

A Gaussian mixture model is a probabilistic model that assumes all the data points are generated from a mixture of a finite number of Gaussian distributions with unknown parameters.
Scikit-learn implements different classes to estimate Gaussian mixture models, that correspond to different estimation strategies, detailed below.

17.5.9.2.1 GMM classifier

The GMM object implements the expectation-maximization (EM) algorithm for fitting mixture-of-Gaussian models. The GMM comes with different options to constrain the covariance of the difference classes estimated: spherical, diagonal, tied or full covariance.

In order to use the *Gaussian Mixture Model*, the user needs to set the sub-node:

<SKLtype>mixture|GMM**</SKLtype>**.

In addition to this XML node, several others are available:

- **<n_components>**, ***integer, optional field*** Number of mixture components.
  Default: 1
- **<covariance_type>**, ***string, optional field***, describes the type of covariance parameters to use. Must be one of 'spherical', 'tied', 'diag', 'full'.
  Default: diag
- **<random_state>**, ***integer seed or random number generator instance, optional field***, A random number generator instance
  Default: 0 or None
- **<min_covar>**, ***float, optional field***, Floor on the diagonal of the covariance matrix to prevent overfitting.
  Default: 1e-3.

- **<thresh>**, ***float, optional field***, convergence threshold.
  Default: 0.01

- **<n_iter>**, ***integer, optional field***, Number of EM iterations to perform.
  Default: 100

- **<n_init>**, ***integer, optional***, Number of initializations to perform. the best results is kept.
  Default: 1

- **<params>**, ***string, optional field***, Controls which parameters are updated in the training process. Can contain any combination of 'w' for weights, 'm' for means, and 'c' for covars.

  Default: 'wmc'

- **<init_params>**, ***string, optional field***, Controls which parameters are updated in the initialization process. Can contain any combination of 'w' for weights, 'm' for means, and 'c' for covars.
  Default: 'wmc'.

Example:

```
<Simulation>
  ...
  <Models>
    ...
      <PostProcessor name='PostProcessorName'
         subType='DataMining'>
          <KDD lib='SciKitLearn'>
              <Features>variableName</Features>
              <SKLtype>mixture|GMM</SKLtype>
              <n_components>2</n_components>
              <covariance_type>spherical</covariance_type>
          </KDD>
      </PostProcessor>
    ...
  <Models>
  ...
<Simulation>
```

17.5.9.2.2 Variational GMM Classifier (VBGMM)

The VBGMM object implements a variant of the Gaussian mixture model with variational inference algorithms. The API is identical to GMM.

In order to use the *Variational Gaussian Mixture Model*, the user needs to set the sub-node:

<SKLtype>mixture|VBGMM**</SKLtype>**.

In addition to this XML node, several others are available:

- **<n_components>**, ***integer, optional field*** Number of mixture components.
  Default: 1
- **<covariance_type>**, ***string, optional field***, describes the type of covariance parameters to use. Must be one of 'spherical', 'tied', 'diag', 'full'.
  Default: diag
- **<alpha>**, ***float, optional field***, represents the concentration parameter of the dirichlet process. Intuitively, the Dirichlet Process is as likely to start a new cluster for a point as it is to add that point to a cluster with alpha elements. A higher alpha means more clusters, as the expected number of clusters is $\alpha * log(N)$.
  Default: 1.

17.5.9.3 Clustering

Clustering of unlabeled data can be performed with this subType of the DataMining PostProcessor.

An overwiev of the different clustering algorithms is given in Table3.

Table 3. Overview of Clustering Methods

Method name	Parameters	Scalability	Usecase	Geometry (metric used)
K-Means	number of clusters	Very large n_samples, medium n_clusters with MiniBatch code	General-purpose, even cluster size, flat geometry, not too many clusters	Distances between points
Affinity propagation	damping, sample preference	Not scalable with n_samples	Many clusters, uneven cluster size, non-flat geometry	Graph distance (e.g. nearest-neighbor graph)
Mean-shift	bandwidth	Not scalable with n_samples	Many clusters, uneven cluster size, non-flat geometry	Distances between points
Spectral clustering	number of clusters	Medium n_samples, small n_clusters	Few clusters, even cluster size, non-flat geometry	Graph distance (e.g. nearest-neighbor graph)
Ward hierarchical clustering	number of clusters	Large n_samples and n_clusters	Many clusters, possibly connectivity constraints	Distances between points
Agglomerative clustering	number of clusters, linkage type, distance	Large n_samples and n_clusters	Many clusters, possibly connectivity constraints, non Euclidean distances	Any pairwise distance
DBSCAN	neighborhood size	Very large n_samples, medium n_clusters	Non-flat geometry, uneven cluster sizes	Distances between nearest points
Gaussian mixtures	many	Not scalable	Flat geometry, good for density estimation	Mahalanobis distances to centers

17.5.9.3.1 K-Means Clustering

The KMeans algorithm clusters data by trying to separate samples in n groups of equal variance, minimizing a criterion known as the inertia or within-cluster sum-of-squares. This algorithm requires the number of clusters to be specified. It scales well to large number of samples and has been used across a large range of application areas in many different fields

In order to use the *K-Means Clustering*, the user needs to set the sub-node:

`<SKLtype>cluster|KMeans</SKLtype>`.

In addition to this XML node, several others are available:

- **<n_clusters>**, ***integer, optional field*** The number of clusters to form as well as the number of centroids to generate.
  Default: 8
- **<max_iter>**, ***integer, optional field***, Maximum number of iterations of the k-means algorithm for a single run.
  Default: 300
- **<n_init>**, ***integer, optional field***, Number of time the k-means algorithm will be run with different centroid seeds. The final results will be the best output of n_init consecutive runs in terms of inertia.
  Default: 3
- **<init>**, ***string, optional***, Method for initialization, k-means++', 'random' or an ndarray:
  - 'k-means++' : selects initial cluster centers for k-mean clustering in a smart way to speed up convergence.
  - 'random': choose k observations (rows) at random from data for the initial centroids.
  - If an ndarray is passed, it should be of shape (n_clusters, n_features) and gives the initial centers.
- **<precompute_distances>**, ***boolean, optional field***, Precompute distances (if true faster but takes more memory).
  Default: true
- **<tol>**, ***float, optional field***, Relative tolerance with regards to inertia to declare convergence.
  Default: 1e-4
- **<n_jobs>**, ***integer, optional field***, The number of jobs to use for the computation. This works by breaking down the pairwise matrix into n jobs even slices and computing them in parallel. If -1 all CPUs are used. If 1 is given, no parallel computing code is used at all,

which is useful for debugging. For n_jobs below -1, (n_cpus + 1 + n_jobs) are used. Thus for n_jobs = -2, all CPUs but one are used.
Default: 1

- **<random_state>**, ***integer or numpy.RandomState, optional field*** The generator used to initialize the centers. If an integer is given, it fixes the seed.
Default: the global numpy random number generator.

Example:

```
<Simulation>
  ...
  <Models>
    ...
      <PostProcessor name='PostProcessorName'
         subType='DataMining'>
          <KDD lib='SciKitLearn'>
              <Features>variableName</Features>
              <SKLtype>cluster|KMeans</SKLtype>
              <n_clusters>2</n_clusters>
              <tol>0.0001</tol>
              <init>random</init>
          </KDD>
      </PostProcessor>
    ...
  <Models>
  ...
<Simulation>
```

17.5.9.3.2 Mini Batch K-Means

The MiniBatchKMeans is a variant of the KMeans algorithm which uses mini-batches to reduce the computation time, while still attempting to optimise the same objective function. Mini-batches are subsets of the input data, randomly sampled in each training iteration.

MiniBatchKMeans converges faster than KMeans, but the quality of the results is reduced. In practice this difference in quality can be quite small.

In order to use the *Mini Batch K-Means Clustering*, the user needs to set the sub-node:

`<SKLtype>cluster|MiniBatchKMeans</SKLtype>`.

In addition to this XML node, several others are available:

- **<n_clusters>**, ***integer, optional field*** The number of clusters to form as well as the number of centroids to generate.
  Default: 8

- **<max_iter>**, ***integer, optional field***, Maximum number of iterations of the k-means algorithm for a single run.
  Default: 100

- **<max_no_improvement>**, ***integer, optional firld***, Control early stopping based on the consecutive number of mini batches that does not yield an improvement on the smoothed inertia. To disable convergence detection based on inertia, set max_no_improvement to None.

  Default: 10

- **<tol>**, ***float, optional field***, Control early stopping based on the relative center changes as measured by a smoothed, variance-normalized of the mean center squared position changes. This early stopping heuristics is closer to the one used for the batch variant of the algorithms but induces a slight computational and memory overhead over the inertia heuristic. To disable convergence detection based on normalized center change, set tol to 0.0 (default).
  Default: 0.0

- **<batch_size>**, ***integer, optional field***, Size of the mini batches.
  Default: 100

- init_size, ***integer, optional field***, Number of samples to randomly sample for speeding up the initialization (sometimes at the expense of accuracy): the only algorithm is initialized by running a batch KMeans on a random subset of the data. *This needs to be larger than k.*,
  *Default: 3 * **<batch_size>***

- **<init>**, ***string, optional***, Method for initialization, k-means++', 'random' or an ndarray:
  - 'k-means++' : selects initial cluster centers for k-mean clustering in a smart way to speed up convergence.
  - 'random': choose k observations (rows) at random from data for the initial centroids.
  - If an ndarray is passed, it should be of shape (n_clusters, n_features) and gives the initial centers.

- **<precompute_distances>**, ***boolean, optional field***, Precompute distances (if true faster but takes more memory).
  Default: true

- **<n_init>**, ***integer, optional field***, Number of time the k-means algorithm will be run with different centroid seeds. The final results will be the best output of n_init consecutive runs in terms of inertia.
  Default: 3

- **<compute_labels>**, ***boolean, optional field***, Compute label assignment and inertia for the complete dataset once the minibatch optimization has converged in fit.
  Default: True

- **<random_state>**, ***integer or numpy.RandomState, optional field*** The generator used to initialize the centers. If an integer is given, it fixes the seed.
  Default: the global numpy random number generator.

- reassignment_ratio, **<float, optional field>**, Control the fraction of the maximum number of counts for a center to be reassigned. A higher value means that low count centers are more easily reassigned, which means that the model will take longer to converge, but should converge in a better clustering.
  Default: 0.01

17.5.9.3.3 Affinity Propagation

AffinityPropagation creates clusters by sending messages between pairs of samples until convergence. A dataset is then described using a small number of exemplars, which are identified as those most representative of other samples. The messages sent between pairs represent the suitability for one sample to be the exemplar of the other, which is updated in response to the values from other pairs. This updating happens iteratively until convergence, at which point the final exemplars are chosen, and hence the final clustering is given.

In order to use the *AffinityPropogation Clustering*, the user needs to set the sub-node:

<SKLtype>cluster|AffinityPropogation**</SKLtype>**.

In addition to this XML node, several others are available:

- **<damping>**, ***float, optional field***, Damping factor between 0.5 and 1.
  Default: 0.5

- **<convergence_iter>**, ***integer, optional field***, Number of iterations with no change in the number of estimated clusters that stops the convergence.
  Default: 15

- **<max_iter>**, ***integer, optional field***, Maximum number of iterations.
  Default: 200

- **<copy>**, ***boolean, optional field***, Make a copy of input data or not.
  Default: True

- **<preference>**, ***array-like, shape (n_samples,) or float, optional field***, Preferences for each point - points with larger values of preferences are more likely to be chosen as exem-

plars. The number of exemplars, ie of clusters, is influenced by the input preferences value.
Default: If the preferences are not passed as arguments, they will be set to the median of the input similarities.

- **<affinity>**, ***string, optional field***,Which affinity to use. At the moment precomputed and euclidean are supported. euclidean uses the negative squared euclidean distance between points.
Default: "euclidean"

- **<verbose>**, ***boolean, optional field***, Whether to be verbose.
Default: False

17.5.9.3.4 Mean Shift

MeanShift clustering aims to discover blobs in a smooth density of samples. It is a centroid based algorithm, which works by updating candidates for centroids to be the mean of the points within a given region. These candidates are then filtered in a post-processing stage to eliminate near-duplicates to form the final set of centroids.

In order to use the *Mean Shift Clustering*, the user needs to set the sub-node:

<SKLtype>cluster|MeanShift**</SKLtype>**.

In addition to this XML node, several others are available:

- **<bandwidth>**, ***float, optional field***, Bandwidth used in the RBF kernel. If not given, the bandwidth is estimated using *sklearn.cluster.estimate_bandwidth*; see the documentation for that function for hints on scalability.

- **<seeds>**, ***array, shape=[n_samples, n_features], optional field***, Seeds used to initialize kernels. If not set, the seeds are calculated by *clustering.get_bin_seeds* with bandwidth as the grid size and default values for other parameters.

- **<bin_seeding>**, ***boolean, optional field***, If true, initial kernel locations are not locations of all points, but rather the location of the discretized version of points, where points are binned onto a grid whose coarseness corresponds to the bandwidth. Setting this option to True will speed up the algorithm because fewer seeds will be initialized.
Default: False Ignored if seeds argument is not None.

- **<min_bin_freq>**, ***integer, optional field***, To speed up the algorithm, accept only those bins with at least min_bin_freq points as seeds.
Default: 1.

- **<cluster_all>**, ***boolean, optional field***, If true, then all points are clustered, even those orphans that are not within any kernel. Orphans are assigned to the nearest kernel. If false,

then orphans are given cluster label -1.
Default: True

17.5.9.3.5 Spectral clustering

SpectralClustering does a low-dimension embedding of the affinity matrix between samples, followed by a *KMeans* in the low dimensional space. It is especially efficient if the affinity matrix is sparse and the pyamg module is installed.

In order to use the *Spectral Clustering*, the user needs to set the sub-node:

<SKLtype>cluster|Spectral**</SKLtype>**.

In addition to this XML node, several others are available:

- **<n_clusters>**, ***integer, optional field***, The dimension of the projection subspace.
  Default: 8
- **<affinity>**, ***string, array-like or callable, optional field***, If a string, this may be one of:
  - 'nearest_neighbors',
  - 'precomputed',
  - 'rbf' or
  - one of the kernels supported by *sklearn.metrics.pairwise_kernels*.

  Only kernels that produce similarity scores (non-negative values that increase with similarity) should be used. This property is not checked by the clustering algorithm.
  Default: 'rbf'
- **<gamma>**, ***float, optional field***, Scaling factor of RBF, polynomial, exponential chi^2 and sigmoid affinity kernel. Ignored for $affinity =' nearest_neighbors'$.
  Default: 1.0
- **<degree>**,***float, optional field***, Degree of the polynomial kernel. Ignored by other kernels.

  Default: 3
- **<coef0>**, ***float, optional field***, Zero coefficient for polynomial and sigmoid kernels. Ignored by other kernels.
  Default: 1
- **<n_neighbors>**, ***integer, optional field***, Number of neighbors to use when constructing the affinity matrix using the nearest neighbors method. Ignored for affinity='rbf'.
  Default: 10

- **<eigen_solver>** ***string, optional field***, The eigenvalue decomposition strategy to use:
  - None,
  - 'arpack',
  - 'lobpcg', or
  - 'amg'

  Note: AMG requires pyamg to be installed. It can be faster on very large, sparse problems, but may also lead to instabilities
- **<random_state>**, ***integer seed, RandomState instance, or None, optional field***, A pseudo random number generator used for the initialization of the lobpcg eigen vectors decomposition when $eigen_solver == `amg'$ and by the K-Means initialization.
  Default: None
- **<n_init>**, ***integer, optional field***, Number of time the k-means algorithm will be run with different centroid seeds. The final results will be the best output of n_init consecutive runs in terms of inertia.
  Default: 10
- **<eigen_tol>**, ***float, optional field***, Stopping criterion for eigendecomposition of the Laplacian matrix when using arpack eigen_solver.
  Default: 0.0
- **<assign_labels>**, ***string, optional field***, The strategy to use to assign labels in the embedding space. There are two ways to assign labels after the laplacian embedding:
  - 'kmeans',
  - 'discretize'

  k-means can be applied and is a popular choice. But it can also be sensitive to initialization. Discretization is another approach which is less sensitive to random initialization.
  Default: 'kmeans'
- **<kernel_params>**, ***dictionary of string to any, optional field***, Parameters (keyword arguments) and values for kernel passed as callable object. Ignored by other kernels.
  Default: None

Notes

If you have an affinity matrix, such as a distance matrix, for which 0 means identical elements, and high values means very dissimilar elements, it can be transformed in a similarity matrix that is well suited for the algorithm by applying the Gaussian (RBF, heat) kernel:

$$np.exp(-X**2/(2.*delta**2)) \tag{33}$$

Another alternative is to take a symmetric version of the k nearest neighbors connectivity matrix of the points. If the *pyamg* package is installed, it is used: this greatly speeds up computation.

17.5.9.3.6 DBSCAN Clustering

The Density-Based Spatial Clustering of Applications with Noise (DBSCAN) algorithm views clusters as areas of high density separated by areas of low density. Due to this rather generic view, clusters found by DBSCAN can be any shape, as opposed to k-means which assumes that clusters are convex shaped.

In order to use the *DBSCAN Clustering*, the user needs to set the sub-node:

<SKLtype>`cluster|DBSCAN`**</SKLtype>**.

In addition to this XML node, several others are available:

- **<eps>**, ***float, optional field***, The maximum distance between two samples for them to be considered as in the same neighborhood.
  Default: 0.5
- **<min_samples>**, ***integer, optional field***, The number of samples in a neighborhood for a point to be considered as a core point.
  Default: 5
- **<metric>**, ***string, or callable, optional field*** The metric to use when calculating distance between instances in a feature array. If metric is a string or callable, it must be one of the options allowed by *metrics.pairwise.calculate_distance* for its metric parameter. If metric is "precomputed", X is assumed to be a distance matrix and must be square.
  Default: 'euclidean'
- **<random_state>**, ***numpy.RandomState, optional field***, The generator used to initialize the centers.
  Default: numpy.random.

17.5.9.3.7 Agglomerative Clustering

Hierarchical clustering is a general family of clustering algorithms that build nested clusters by merging or splitting them successively. This hierarchy of clusters is represented as a tree (or dendrogram). The root of the tree is the unique cluster that gathers all of the samples, the leaves being the clusters with only one sample. The AgglomerativeClustering object performs a hierarchical clustering using a bottom up approach: each observation starts in its own cluster, and clusters are successively merged together. The linkage criteria determines the metric used for the merge strategy:

- Ward: it minimizes the sum of squared differences within all clusters. It is a variance-minimizing approach and in this sense is similar to the k-means objective function but tackled with an agglomerative hierarchical approach.

- Maximum or complete linkage: it minimizes the maximum distance between observations of pairs of clusters.
- Average linkage: it minimizes the average of the distances between all observations of pairs of clusters.

AgglomerativeClustering can also scale to large number of samples when it is used jointly with a connectivity matrix, but is computationally expensive when no connectivity constraints are added between samples: it considers at each step all of the possible merges.

In order to use the *Agglomerative Clustering*, the user needs to set the sub-node:

<SKLtype>cluster|Agglomerative**</SKLtype>**.

In addition to this XML node, several others are available:

- **<n_clusters>**, ***int, optional field***, The number of clusters to find.
  Default: 2
- **<connectivity>**, ***array like or callable, optional field***, Connectivity matrix. Defines for each sample the neighboring samples following a given structure of the data. This can be a connectivity matrix itself or a callable that transforms the data into a connectivity matrix, such as derived from kneighbors graph. Default is None, i.e, the hierarchical clustering algorithm is unstructured.
  Default: None
- **<affinity>**, ***string or callable, optional field***, Metric used to compute the linkage. Can be "euclidean", "$l1$", "$l2$", "manhattan","cosine", or "precomputed". If linkage is "ward", only "euclidean" is accepted.
  Default: euclidean
- **<n_components>**, ***int, optional field***, Number of connected components. If None the number of connected components is estimated from the connectivity matrix. NOTE: This parameter is now directly determined from the connectivity matrix and will be removed in 0.18.
- **<linkage>**, ***ward,complete,average, optional field***, Which linkage criterion to use. The linkage criterion determines which distance to use between sets of observation. The algorithm will merge the pairs of cluster that minimize this criterion. Ward minimizes the variance of the clusters being merged. Average uses the average of the distances of each observation of the two sets. Complete or maximum linkage uses the maximum distances between all observations of the two sets..
  Default: ward

17.5.9.3.8 Clustering performance evaluation

Evaluating the performance of a clustering algorithm is not as trivial as counting the number of errors or the precision and recall of a supervised classification algorithm. In particular any evaluation metric should not take the absolute values of the cluster labels into account but rather if this clustering define separations of the data similar to some ground truth set of classes or satisfying some assumption such that members belong to the same class are more similar that members of different classes according to some similarity metric.

If the ground truth labels are not known, evaluation must be performed using the model itself. The **Silhouette Coefficient** is an example of such an evaluation, where a higher Silhouette Coefficient score relates to a model with better defined clusters. The Silhouette Coefficient is defined for each sample and is composed of two scores:

1. The mean distance between a sample and all other points in the same class.
2. The mean distance between a sample and all other points in the next nearest cluster.

The Silhoeutte Coefficient s for a single sample is then given as:

$$s = \frac{b - a}{max(a, b)} \quad (34)$$

The Silhouette Coefficient for a set of samples is given as the mean of the Silhouette Coefficient for each sample. In normal usage, the Silhouette Coefficient is applied to the results of a cluster analysis.

Advantages

- The score is bounded between -1 for incorrect clustering and +1 for highly dense clustering. Scores around zero indicate overlapping clusters.
- The score is higher when clusters are dense and well separated, which relates to a standard concept of a cluster.

Drawbacks

The Silhouette Coefficient is generally higher for convex clusters than other concepts of clusters, such as density based clusters like those obtained through DBSCAN.

17.5.9.4 Decomposing signals in components (matrix factorization problems)

17.5.9.4.1 Principal component analysis (PCA)

- **Exact PCA and probabilistic interpretation**
  Linear Dimensionality reduction using Singular Value Decomposition of the data and keeping only the most significant singular vectors to project the data to a lower dimensional space. In order to use the *Exact PCA*, the user needs to set the sub-node:

  `<SKLtype>decomposition|PCA</SKLtype>`.

  In addition to this XML node, several others are available:

  – `<n_components>`, ***integer, None or String, optional field***, Number of components to keep. if

  – `<n_components>` is not set all components are kept,
  Default: all components

  – `<copy>`, ***boolean, optional field***, If False, data passed to fit are overwritten and running fit(X).transform(X) will not yield the expected results, use fit_transform(X) instead.
  Default: True

  – `<whiten>`, ***boolean, optional field***, When True the components_ vectors are divided by n_samples times singular values to ensure uncorrelated outputs with unit component-wise variances. Whitening will remove some information from the transformed signal (the relative variance scales of the components) but can sometime improve the predictive accuracy of the downstream estimators by making there data respect some hard-wired assumptions.
  Default: False

  Example:

```
<Simulation>
  ...
  <Models>
    ...
      <PostProcessor name='PostProcessorName'
         subType='DataMining'>
          <KDD lib='SciKitLearn'>
              <Features>variable1,variable2,variable3,
                 variable4,variable5</Features>
              <SKLtype>decomposition|PCA</SKLtype>
              <n_components>2</n_components>
          </KDD>
      </PostProcessor>
    ...
  <Models>
  ...
<Simulation>
```

- **Randomized (Approximate) PCA**
  Linear Dimensionality reduction using Singular Value Decomposition of the data and keeping only the most significant singular vectors to project the data to a lower dimensional space. In order to use the *Randomized PCA*, the user needs to set the sub-node:

  <SKLtype>decomposition|RandomizedPCA</SKLtype>.

  In addition to this XML node, several others are available:

  – <n_components>, ***interger, None or String, optional field***, Number of components to keep. if n_components is not set all components are kept.
  Default: all components

  – <copy>, ***boolean, optional field***, If False, data passed to fit are overwritten and running fit(X).transform(X) will not yield the expected results, use fit_transform(X) instead.
  Default: True

  – <iterated_power>, ***integer, optional field***, Number of iterations for the power method.
  Default: 3

  – <whiten>, ***boolean, optional field***, When True the components_ vectors are divided by n_samples times singular values to ensure uncorrelated outputs with unit component-wise variances. Whitening will remove some information from the transformed signal (the relative variance scales of the components) but can sometime improve the predictive accuracy of the downstream estimators by making there data respect some hard-wired assumptions.
  Default: False

  – <random_state>, ***int, or Random State instance or None, optional field***, Pseudo Random Number generator seed control. If None, use the numpy.random singleton.
  Default: None

- **Kernel PCA**
  Non-linear dimensionality reduction through the use of kernels. In order to use the *Kernel PCA*, the user needs to set the sub-node:

  <SKLtype>decomposition|KernelPCA</SKLtype>.

  In addition to this XML node, several others are available:

  – <n_components>, ***interger, None or String, optional field***, Number of components to keep. if n_components is not set all components are kept.
  Default: all components

  – <kernel>, ***string, optional field***, name of the kernel to be used, options are:

    * linear
    * poly

* rbf
* sigmoid
* cosine
* precomputed

Default: linear **<degree>**, ***integer, optional field***, Degree for poly kernels, ignored by other kernels.
Default: 3 **<gamma>**, ***float, optional field***, Kernel coefficient for rbf and poly kernels, ignored by other kernels.
Default: 1/n_features

- **<coef0>**, ***float, optional field***, independent term in poly and sigmoig kernels, ignored by other kernels.
- **<kernel_params>**, ***mapping of string to any, optional field***, Parameters (keyword arguments) and values for kernel passed as callable object. Ignored by other kernels.
  Default: 3
- alpha, ***int, optional field***, Hyperparameter of the ridge regression that learns the inverse transform (when fit_inverse_transform=True).
  Default: 1.0
- **<fit_inverse_transform>**, ***bool, optional field***, Learn the inverse transform for non-precomputed kernels. (i.e. learn to find the pre-image of a point)
  Default: False
- **<eigen_solver>**, ***string, optional field***, Select eigensolver to use. If n_components is much less than the number of training samples, arpack may be more efficient than the dense eigensolver. Options are:
  * auto
  * dense
  * arpack

  Default: False
- tol, ***float, optional field***, convergence tolerance for arpack.
  Default: 0 (optimal value will be chosen by arpack)
- max_iter, ***int, optional field***, maximum number of iterations for arpack.
  Default: None (optimal value will be chosen by arpack)
- **<remove_zero_eig>**, ***boolean, optional field***, If True, then all components with zero eigenvalues are removed, so that the number of components in the output may be ¡ n_components (and sometimes even zero due to numerical instability). When n_components is None, this parameter is ignored and components with zero eigenvalues are removed regardless.
  Default: True

- **Sparse PCA**
  Finds the set of sparse components that can optimally reconstruct the data. The amount of sparseness is controllable by the coefficient of the L1 penalty, given by the parameter alpha. In order to use the *Sparse PCA*, the user needs to set the sub-node:

  `<SKLtype>decomposition|SparsePCA</SKLtype>`.

  In addition to this XML node, several others are available:

  – **<n_components>**, ***integer, optional field***, Number of sparse atoms to extract.
  Default: None

  – **<alpha>**, ***float, optional field***, Sparsity controlling parameter. Higher values lead to sparser components.
  Default: 1.0

  – **<ridge_alpha>**, ***float, optional field***, Amount of ridge shrinkage to apply in order to improve conditioning when calling the transform method.
  Default: 0.01

  – **<max_iter>**, ***float, optional field***, maximum number of iterations to perform.
  Default: 1000

  – **<tol>**, ***float, optional field***, convergence tolerance.
  Default: 1E-08

  – **<method>**, ***string, optional field***, method to use, options are:

    * lars: uses the least angle regression method to solve the lasso problem (linear_model.lars_path)
    * cd: uses the coordinate descent method to compute the Lasso solution (linear_model.Lasso)

  Lars will be faster if the estimated components are sparse.
  Default: lars

  – **<n_jobs>**, ***int, optional field***, number of parallel runs to run.
  Default: 1

  – **<U_init>**, ***array of shape (n_samples, n_components) , optional field***, Initial values for the loadings for warm restart scenarios
  Default: None

  – **<V_init>**, ***array of shape (n_components, n_features), optional field***, Initial values for the components for warm restart scenarios
  Default: None

  – verbose, ***boolean, optional field***, Degree of verbosity of the printed output.
  Default: False

  – random_state, ***int or Random State, optional field***, Pseudo number generator state used for random sampling.
  Default: None

- **Mini Batch Sparse PCA**
  Finds the set of sparse components that can optimally reconstruct the data. The amount of sparseness is controllable by the coefficient of the L1 penalty, given by the parameter alpha. In order to use the *Mini Batch Sparse PCA*, the user needs to set the sub-node:

  <SKLtype>decomposition|MiniBatchSparsePCA**</SKLtype>**.

  In addition to this XML node, several others are available:

  - **<n_components>**, ***integer, optional field***, Number of sparse atoms to extract.
    Default: None
  - **<alpha>**, ***float, optional field***, Sparsity controlling parameter. Higher values lead to sparser components.
    Default: 1.0
  - **<ridge_alpha>**, ***float, optional field***, Amount of ridge shrinkage to apply in order to improve conditioning when calling the transform method.
    Default: 0.01
  - **<n_iter>**, ***float, optional field***, number of iterations to perform per mini batch.
    Default: 100
  - **<callback>**, ***callable, optional field***, callable that gets invoked every five iterations.

    Default: None
  - **<batch_size>**, ***int, optional field***, the number of features to take in each mini batch.

    Default: 3
  - **<verbose>**, ***boolean, optional field***, Degree of verbosity of the printed output.
    Default: False
  - **<shuffle>**, ***boolean, optional field***, whether to shuffle the data before splitting it in batches.
    Default: True
  - **<n_jobs>**, ***integer, optional field***, Parameters (keyword arguments) and values for kernel passed as callable object. Ignored by other kernels.
    Default: 3
  - **<metho>**, ***string, optional field***, method to use, options are:
    * lars: uses the least angle regression method to solve the lasso problem (linear_model.lars_path),
    * cd: uses the coordinate descent method to compute the Lasso solution (linear_model.Lasso)

    Lars will be faster if the estimated components are sparse.
    Default: lars
  - **<random_state>**, ***integer or Random State, optional field***, Pseudo number generator state used for random sampling.
    Default: None

17.5.9.4.2 Truncated singular value decomposition

Dimensionality reduction using truncated SVD (aka LSA). In order to use the *Truncated SVD*, the user needs to set the sub-node:

<SKLtype>decomposition|TruncatedSVD**</SKLtype>**.

In addition to this XML node, several others are available:

- **<n_components>**, ***integer, optional field***, Desired dimensionality of output data. Must be strictly less than the number of features. The default value is useful for visualisation. For LSA, a value of 100 is recommended.
  Default: 2
- **<algorithm>**, ***string, optional field***, SVD solver to use:
  - Randomized: randomized algorithm
  - Arpack: ARPACK wrapper in.

  Default: Randomized
- **<n_iter>**, ***float, optional field***, number of iterations andomized SVD solver. Not used by ARPACK.
  Default: 5
- **<random_state>**, ***int or Random State, optional field***, Pseudo number generator state used for random sampling. If not given, the numpy.random singleton is used.
  Default: None
- **<tol>**, ***float, optional field***, Tolerance for ARPACK. 0 means machine precision. Ignored by randomized SVD solver.
  Default: 0.0

17.5.9.4.3 Fast ICA

A fast algorithm for Independent Component Analysis. In order to use the *Fast ICA*, the user needs to set the sub-node:

<SKLtype>decomposition|FastICA**</SKLtype>**.

In addition to this XML node, several others are available:

- **<n_components>**, ***integer, optional field***, Number of components to use. If none is passed, all are used.
  Default: None

- **<algorithm>**, ***string, optional field***, algorithm used in FastICA:
  - parallel,
  - deflation.

  Default: parallel

- **<fun>**, ***string or function, optional field***, The functional form of the G function used in the approximation to neg-entropy. Could be either:
  - logcosh,
  - exp, or
  - cube.

  One can also provide own function. It should return a tuple containing the value of the function, and of its derivative, in the point.
  Default: logcosh

- **<fun_args>**, ***dictionary, optional field***, Arguments to send to the functional form. If empty and if fun='logcosh', fun_args will take value 'alpha' : 1.0.
  Default: None

- **<max_iter>**, ***float, optional field***, maximum number of iterations during fit.
  Default: 200

- **<tol>**, ***float, optional field***, Tolerance on update at each iteration.
  Default: 0.0001

- **<w_init>**, ***None or an (n_components, n_components) ndarray, optional field***, The mixing matrix to be used to initialize the algorithm.
  Default: None

- **<randome_state>**, ***int or Random State, optional field***, Pseudo number generator state used for random sampling.
  Default: None

17.5.9.5 Manifold learning

A manifold is a topological space that resembles a Euclidean space locally at each point. Manifold learning is an approach to non-linear dimensionality reduction. It assumes that the data of interest lie on an embedded non-linear manifold within the higher-dimensional space. If this manifold is of low dimension, data can be visualized in the low-dimensional space. Algorithms for this task are based on the idea that the dimensionality of many data sets is only artificially high.

17.5.9.5.1 Isomap

Non-linear dimensionality reduction through Isometric Mapping (Isomap). In order to use the *Isometric Mapping*, the user needs to set the sub-node:

<SKLtype>manifold|Isomap**</SKLtype>**.

In addition to this XML node, several others are available:

- **<n_neighbors>**, ***integer, optional field***, Number of neighbors to consider for each point.
  Default: 5

- **<n_components>**, ***integer, optional field***, Number of coordinates to manifold.
  Default: 2

- **<eigen_solver>**, ***string, optional field***, eigen solver to use:
  - auto: Attempt to choose the most efficient solver for the given problem,
  - arpack: Use Arnoldi decomposition to find the eigenvalues and eigenvectors
  - dense: Use a direct solver (i.e. LAPACK) for the eigenvalue decomposition

  Default: auto

- **<tol>**, ***float, optional field***, Convergence tolerance passed to arpack or lobpcg. not used if eigen_solver is 'dense'.
  Default: 0.0

- **<max_iter>**, ***float, optional field***, Maximum number of iterations for the arpack solver. not used if eigen_solver == 'dense'.
  Default: None

- **<path_method>**, ***string, optional field***, Method to use in finding shortest path. Could be either:
  - Auto: attempt to choose the best algorithm
  - FW: Floyd-Warshall algorithm
  - D: Dijkstra algorithm with Fibonacci Heaps

  Default: auto

- **<neighbors_algorithm>**, ***string, optional field***, Algorithm to use for nearest neighbors search, passed to neighbors.NearestNeighbors instance.
  - auto,

- brute
- kd_tree
- ball_tree

Default: auto

Example:

```
<Simulation>
  ...
  <Models>
    ...
      <PostProcessor name='PostProcessorName'
         subType='DataMining'>
          <KDD lib='SciKitLearn'>
              <Features>input</Features>
              <SKLtype>manifold|Isomap</SKLtype>
              <n_neighbors>5</n_neighbors>
              <n_components>3</n_components>
              <eigen_solver>arpack</eigen_solver>
              <neighbors_algorithm>kd_tree</neighbors_algorithm>
            </KDD>
      </PostProcessor>
    ...
  <Models>
  ...
<Simulation>
```

17.5.9.5.2 Locally Linear Embedding

In order to use the *Locally Linear Embedding*, the user needs to set the sub-node:

```
<SKLtype>manifold|LocallyLinearEmbedding</SKLtype>.
```

In addition to this XML node, several others are available:

- <n_neighbors>, ***integer, optional field***, Number of neighbors to consider for each point. *Default: 5*
- <n_components>, ***integer, optional field***, Number of coordinates to manifold. *Default: 2*

- **<reg>**,*float, optional field*, regularization constant, multiplies the trace of the local covariance matrix of the distances.
  Default: 0.01

- **<eigen_solver>**, *string, optional field*, eigen solver to use:

  – auto: Attempt to choose the most efficient solver for the given problem,

  – arpack: use arnoldi iteration in shift-invert mode.

  – dense: use standard dense matrix operations for the eigenvalue

  Default: auto

- **<tol>**,*float, optional field*, Convergence tolerance passed to arpack. not used if eigen_solver is 'dense'.
  Default: 1E-06

- **<max_iter>**, *int, optional field*, Maximum number of iterations for the arpack solver. not used if eigen_solver == 'dense'.
  Default: 100

- **<method>**, *string, optional field*, Method to use. Could be either:

  – Standard: use the standard locally linear embedding algorithm

  – hessian: use the Hessian eigenmap method

  – itsa: use local tangent space alignment algorithm

  Default: standard

- **<hessian_tol>**,*float, optional field*, Tolerance for Hessian eigenmapping method. Only used if method == 'hessian'
  Default: 0.0001

- **<modified_tol>**,*float, optional field*, Tolerance for modified LLE method. Only used if method == 'modified'
  Default: 0.0001

- **<neighbors_algorithm>**, *string, optional field*, Algorithm to use for nearest neighbors search, passed to neighbors.NearestNeighbors instance.

  – auto,

  – brute

  – kd_tree

- ball_tree

Default: auto

- **<random_state>**, ***int or numpy random state, optional field***, the generator or seed used to determine the starting vector for arpack iterations.
  Default: None

17.5.9.5.3 Spectral Embedding

Spectral embedding for non-linear dimensionality reduction, it forms an affinity matrix given by the specified function and applies spectral decomposition to the corresponding graph laplacian. The resulting transformation is given by the value of the eigenvectors for each data point In order to use the *Spectral Embedding*, the user needs to set the sub-node:

<SKLtype>`manifold|SpectralEmbedding`**</SKLtype>**.

In addition to this XML node, several others are available:

- **<n_components>**, ***integer, optional field***, the dimension of projected sub-space.
  Default: 2
- **<eigen_solver>**, ***string, optional field***, the eigen value decomposition strategy to use:
  - none,
  - arpack.
  - lobpcg,
  - amg

  Default: none
- **<random_state>**, ***integer or numpy random state, optional field***, A pseudo random number generator used for the initialization of the lobpcg eigen vectors decomposition when eigen_solver == 'amg.
  Default: None
- **<affinity>**, ***string or callable, optional field***, How to construct the affinity matrix:
  - *nearest_neighbors* : construct affinity matrix by knn graph
  - *rbf* : construct affinity matrix by rbf kernel
  - *precomputed* : interpret X as precomputed affinity matrix

- *callable* : use passed in function as affinity the function takes in data matrix (n_samples, n_features) and return affinity matrix (n_samples, n_samples).

Default: nearest_neighbor

- **<gamma>**, ***float, optional field***, Kernel coefficient for rbf kernel.
  Default: None
- **<n_neighbors>**, ***int, optional field***, Number of nearest neighbors for nearest_neighbors graph building.
  Default: None

17.5.9.5.4 Multi-dimensional Scaling (MDS)

In order to use the *Multi Dimensional Scaling*, the user needs to set the sub-node:

<SKLtype>`manifold|MDS`**</SKLtype>**.

In addition to this XML node, several others are available:

- **<metric>**, ***boolean, optional field***, compute metric or nonmetric SMACOF (Scaling by Majorizing a Complicated Function) algorithm
  Default: True
- **<n_components>**, ***integer, optional field***, number of dimension in which to immerse the similarities overridden if initial array is provided.
  Default: 2
- **<n_init>**, ***integer, optional field***, Number of time the smacof algorithm will be run with different initialisation. The final results will be the best output of the n_init consecutive runs in terms of stress.
  Default: 4
- **<max_iter>**, ***integer, optional field***, Maximum number of iterations of the SMACOF algorithm for a single run
  Default: 300
- **<verbose>**, ***integer, optional field***, level of verbosity
  Default: 0
- **<eps>**, ***float, optional field***, relative tolerance with respect to stress to declare converge
  Default: 1E-06

- **<n_jobs>**, ***integer, optional field***, The number of jobs to use for the computation. This works by breaking down the pairwise matrix into n_jobs even slices and computing them in parallel. If -1 all CPUs are used. If 1 is given, no parallel computing code is used at all, which is useful for debugging. For n_jobs below -1, (n_cpus + 1 + n_jobs) are used. Thus for n_jobs = -2, all CPUs but one are used.
  Default: 1
- **<random_state>**, **<integer or numpy random state, optional field>**, The generator used to initialize the centers. If an integer is given, it fixes the seed. Defaults to the global numpy random number generator.
  Default: None
- **<dissimilarity>**, ***string, optional field***, Which dissimilarity measure to use. Supported are 'euclidean' and 'precomputed'.
  Default: euclidean

17.5.9.6 Scipy

'Scipy' provides a Hierarchical clustering that performs clustering over PointSet and HistorySet. This algorithm also automatically generates a dendrogram in .pdf format (i.e., dendrogram.pdf).

- **<SCIPYtype>**, ***string, required field***, SCIPY algorithm to be employed.
- **<Features>**, ***string, required field***, defines the data to be used for training the data mining algorithm. It can be:
  - the name of the variable in the defined dataObject entity
  - the location (i.e. input or output). In this case the data mining is applied to all the variables in the defined space.
- **<method>**, ***string, required field***, The linkage algorithm to be used
  Default: single, complete, weighted, centroids, median, ward.
- **<metric>**, ***string, required field***, The distance metric to be used
  Default: 'braycurtis', 'canberra', 'chebyshev', 'cityblock', 'correlation', 'cosine', 'dice', 'euclidean', 'hamming', 'jaccard', 'kulsinski', 'mahalanobis', 'matching', 'minkowski', 'rogerstanimoto', 'russellrao', 'seuclidean', 'sokalmichener', 'sokalsneath', 'sqeuclidean', 'yule'.
- **<level>**, ***float, required field***, Clustering distance level where actual clusters are formed.
- **<criterion>**, ***string, required field***, The criterion to use in forming flat clusters. This can be any of the following values:

– "inconsistent" : If a cluster node and all its descendants have an inconsistent value less than or equal to 't' then all its leaf descendants belong to the same flat cluster. When no non-singleton cluster meets this criterion, every node is assigned to its own cluster. (Default)
– "distance" : Forms flat clusters so that the original observations in each flat cluster have no greater a cophenetic distance than t.
– "maxclust" : Finds a minimum threshold "r" so that the cophenetic distance between any two original observations in the same flat cluster is no more than "r" and no more than t flat clusters are formed.
– "monocrit" : Forms a flat cluster from a cluster node c with index i when $monocrit[j] <= t$.
– "maxclust_monocrit" : Forms a flat cluster from a non-singleton cluster node "c" when $monocrit[i] <= r$ for all cluster indices "i" below and including "c". "r" is minimized such that no more than "t" flat clusters are formed. monocrit must be monotonic.

- **<dendrogram>**, ***boolean, required field***, If True the dendrogram is actually created.
- **<truncationMode>**, ***string, required field***, The dendrogram can be hard to read when the original observation matrix from which the linkage is derived is large. Truncation is used to condense the dendrogram. There are several modes:
  – "None": No truncation is performed (Default).
  – "lastp": The last p non-singleton formed in the linkage are the only non-leaf nodes in the linkage; they correspond to rows $Z[n - p - 2 : end]$ in Z. All other non-singleton clusters are contracted into leaf nodes.
  – "level"/"mtica": No more than p levels of the dendrogram tree are displayed. This corresponds to Mathematica behavior.
- **<p>**, ***int, required field***, The p parameter for truncationMode.
- **<leafCounts>**, ***boolean, required field***, When True the cardinality non singleton nodes contracted into a leaf node is indiacted in parenthesis.
- **<showContracted>**, ***boolean, required field***, When True the heights of non singleton nodes contracted into a leaf node are plotted as crosses along the link connecting that leaf node.
- **<annotatedAbove>**, ***float, required field***, Clustering level above which the branching level is annotated.

Example:

```xml
<Simulation>
  ...
  <Models>
    ...
    <PostProcessor name="hierarchical" subType="DataMining"
       verbosity="quiet">
      <KDD lib="Scipy" labelFeature='labels'>
        <SCIPYtype>cluster|Hierarchical</SCIPYtype>
        <Features>output</Features>
        <method>single</method>
        <metric>euclidean</metric>
        <level>75</level>
        <criterion>distance</criterion>
        <dendrogram>true</dendrogram>
        <truncationMode>lastp</truncationMode>
        <p>20</p>
        <leafCounts>True</leafCounts>
        <showContracted>True</showContracted>
        <annotatedAbove>10</annotatedAbove>
      </KDD>
    </PostProcessor>
    ...
  <Models>
  ...
<Simulation>
```

17.5.10 Interfaced

The **Interfaced** post-processor is a Post-Processor that allows the user to create its own Post-Processor. While the External Post-Processor (see Section 17.5.7 allows the user to create case-dependent Post-Processors, with this new class the user can create new general purpose Post-Processors.

In order to use the *Interfaced* PP, the user needs to set the `subType` of a `<PostProcessor>` node:

`<PostProcessor subType='Interfaced'/>`.

Several sub-nodes are available:

- `<method>`, ***comma separated string, required field***, lists the method names of a method

that will be computed (each returning a post-processing value). All available methods need to be included in the "/raven/framework/PostProcessorFunctions/" folder

Example:

```xml
<Simulation>
  ...
  <Models>
    ...
    <PostProcessor name="example"
       subType='InterfacedPostProcessor'verbosity='debug'>
       <method>testInterfacedPP</method>
       <!--Here, the xml nodes required by the chosen method
          have to be
       included.
        -->
    </PostProcessor>
    ...
  </Models>
  ...
</Simulation>
```

All the **Interfaced** post-processors need to be contained in the "/raven/framework/PostProcessorFunctions/" folder. In fact, once the **Interfaced** post-processor is defined in the RAVEN input file, RAVEN search that the method of the post-processor is located in such folder.

The class specified in the **Interfaced** post-processor has to inherit the PostProcessorInterfaceBase class and the user must specify this set of methods:

- initialize: in this method, the internal parameters of the post-processor are initialized. Mandatory variables that needs to be specified are the following:
  - self.inputFormat: type of dataObject expected in input
  - self.outputFormat: type of dataObject generated in output
- readMoreXML: this method is in charge of reading the PostProcessor xml node, parse it and fill the PostProcessor internal variables.
- run: this method performs the desired computation of the dataObject.

```python
from PostProcessorInterfaceBaseClass import PostProcessorInterfaceBase
class testInterfacedPP(PostProcessorInterfaceBase):
  def initialize(self)
```

```
  def readMoreXML(self,xmlNode)
  def run(self,inputDic)
```

17.5.10.1 Data Format

The user is not allowed to modify directly the DataObjects, however the content of the DataObjects is available in the form of a python dictionary. Both the dictionary give in input and the one generated in the output of the PostProcessor are structured as follows:

```
inputDict = {'data':{}, 'metadata':{}}
```

where:

```
inputDict['data'] = {'input':{}, 'output':{}}
```

In the input dictonary, each input variable is listed as a dictionary that contains a numpy array with its own values as shown below for a simplified example

```
inputDict['data']['input'] = {'inputVar1': array([ 1.,2.,3.]),
                              'inputVar2': array([4.,5.,6.])}
```

Similarly, if the dataObject is a PointSet then the output dictionary is structured as follows:

```
inputDict['data']['output'] = {'outputVar1': array([ .1,.2,.3]),
                               'outputVar2':array([.4,.5,.6])}
```

Howevers, if the dataObject is a HistorySet then the output dictionary is structured as follows:

```
inputDict['data']['output'] = {'hist1': {}, 'hist2':{}}
```

where

```
inputDict['output']['data'][hist1] = {'time': array([ .1,.2,.3]),
                              'outputVar1':array([ .4,.5,.6])}
inputDict['output']['data'][hist2] = {'time': array([ .1,.2,.3]),
                              'outputVar1':array([ .14,.15,.16])}
```

17.5.10.2 Method: HStoPSOperator

This Post-Processor performs the conversion from HistorySet to PointSet performing a projection of the output space.

In the <PostProcessor> input block, the following XML sub-nodes are available:

- <pivotParameter>, ***string, optional field***, ID of the temporal variable. Default is "time". **Note:** Used just in case the <pivotValue>-based operation is requested
- <operator>, ***string, optional field***, the operation to perform on the output space:
  - **min**, compute the minimum along each single history
  - **max**, compute the maximum along each single history
  - **min**, compute the average along each single history

  Note: This node can be inputted only if <pivotValue> and <row> are not present
- <pivotValue>, ***float, optional field***, the value of the pivotParameter with respect to the other outputs need to be extracted. **Note:** This node can be inputted only if <operator> and <row> are not present
- <pivotStrategy>, ***string, optional field***, The strategy to use for the pivotValue:
  - **nearest**, find the value that is the nearest with respect the <pivotValue>
  - **floor**, find the value that is the nearest with respect to the <pivotValue> but less then the <pivotValue>
  - **celing**, find the value that is the nearest with respect to the <pivotValue> but greater then the <pivotValue>
  - **interpolate**, if the exact <pivotValue> can not be found, interpolate using a linear approach

  Note: Valid just in case <pivotValue> is present
- <row>, ***int, optional field***, the row index at which the outputs need to be extracted. **Note:** This node can be inputted only if <operator> and <pivotValue> are not present

This example will show how the XML input block would look like:

```
<Simulation>
  ...
  <Models>
    ...
    <PostProcessor name="HStoPSperatorRows"
       subType="InterfacedPostProcessor">
      <method>HStoPSOperator</method>
      <row>-1</row>
    </PostProcessor>
```

```xml
    <PostProcessor name="HStoPSoperatorPivotValues"
       subType="InterfacedPostProcessor">
        <method>HStoPSOperator</method>
        <pivotParameter>time</pivotParameter>
        <pivotValue>0.3</pivotValue>
    </PostProcessor>
    <PostProcessor name="HStoPSoperatorOperatorMax"
       subType="InterfacedPostProcessor">
        <method>HStoPSOperator</method>
        <pivotParameter>time</pivotParameter>
        <operator>max</operator>
    </PostProcessor>
    <PostProcessor name="HStoPSoperatorOperatorMin"
       subType="InterfacedPostProcessor">
        <method>HStoPSOperator</method>
        <pivotParameter>time</pivotParameter>
        <operator>min</operator>
    </PostProcessor>
    <PostProcessor name="HStoPSoperatorOperatorAverage"
       subType="InterfacedPostProcessor">
        <method>HStoPSOperator</method>
        <pivotParameter>time</pivotParameter>
        <operator>average</operator>
    </PostProcessor>
    ...
  </Models>
  ...
</Simulation>
```

17.5.10.3 Method: HistorySetSampling

This Post-Processor performs the conversion from HistorySet to HistorySet The conversion is made so that each history H is re-sampled accordingly to a specific sampling strategy. It can be used to reduce the amount of space required by the HistorySet.

In the **<PostProcessor>** input block, the following XML sub-nodes are required, independent of the **subType** specified:

- **<samplingType>**, ***string, required field***, specifies the type of sampling method to be used (uniform, firstDerivative secondDerivative, filteredFirstDerivative or filteredSecondDerivative).

- **<numberOfSamples>**, ***integer, optional field***, number of samples (required only for the following sampling types: uniform, firstDerivative secondDerivative)
- **<pivotParameter>**, ***string, required field***, ID of the temporal variable
- **<interpolation>**, ***string, optional field***, type of interpolation to be employed for the history recostruction (required only for the following sampling types: uniform, firstDerivative secondDerivative). Valid types of interpolation to specified: linear, nearest, zero, slinear, quadratic, cubic, intervalAverage;
- **<tolerance>**, ***string, optional field***, tolerance level (required only for the following sampling types: filteredFirstDerivative or filteredSecondDerivative)

17.5.10.4 Method: HistorySetSync

This Post-Processor performs the conversion from HistorySet to HistorySet The conversion is made so that all histories are synchronized in time. It can be used to allow the histories to be sampled at the same time instant.

There are two possible synchronization methods, specified through the **<syncMethod>** node. If the **<syncMethod>** is **'grid'**, a **<numberOfSamples>** node is specified, which yields an equally-spaced grid of time points. The output values for these points will be linearly derived using nearest sampled time points, and the new HistorySet will contain only the new grid points.

The other methods are used by specifying **<syncMethod>** as **'all'**, **'min'**, or **'max'**. For **'all'**, the postprocessor will iterate through the existing histories, collect all the time points used in any of them, and use these as the new grid on which to establish histories, retaining all the exact original values and interpolating linearly where necessary. In the event of **'min'** or **'max'**, the postprocessor will find the smallest or largest time history, respectively, and use those time values as nodes to interpolate between.

In the **<PostProcessor>** input block, the following XML sub-nodes are required, independent of the subType specified:

- **<pivotParameter>**, ***string, required field***, ID of the temporal variable
- **<extension>**, ***string, required field***, type of extension when the sync process goes outside the boundaries of the history (zeroed or extended)
- **<syncMethod>**, ***string, required field***, synchronization strategy to employ (see description above). Options are **'grid'**, **'all'**, **'max'**, **'min'**.
- **<numberOfSamples>**, ***integer, optional field***, required if **<syncMethod>** is **'grid'**, number of new time samples

17.5.10.5 Method: HistorySetSnapShot

This Post-Processor performs the conversion from HistorySet to PointSet The conversion is made so that each history H is converted to a single point P. There are several methods that can be employed to choose the single point from the history:

- min: Take a time slice when the **<pivotVar>** is at its smallest value,
- max: Take a time slice when the **<pivotVar>** is at its largest value,
- average: Take a time slice when the **<pivotVar>** is at its time-weighted average value,
- value: Take a time slice when the **<pivotVar>** *first passes* its specified value,
- timeSlice: Take a time slice index from the sampled time instance space.

To demonstrate the timeSlice, assume that each history H is a dict of n output variables $x_1 = [...], x_n = [...]$, then the resulting point P is at time instant index t: $P = [x_1[t], ..., x_n[t]]$.

Choosing one the these methods for the **<type>** node will take a time slice for all the variables in the output space based on the provided parameters. Alternatively, a **'mixed'** type can be used, in which each output variable can use a different time slice parameter. In other words, you can take the max of one variable while taking the minimum of another, etc.

In the **<PostProcessor>** input block, the following XML sub-nodes are required, independent of the **subType** specified:

- **<type>**, ***string, required field***, type of operation: **'min'**, **'max'**, **'average'**, **'value'**, **'timeSlice'**, or **'mixed'**
- **<extension>**, ***string, required field***, type of extension when the sync process goes outside the boundaries of the history (zeroed or extended)
- **<pivotParameter>**, ***string, optional field***, name of the temporal variable. Required for the **'average'** and **'timeSlice'** methods.

If a **'timeSlice'** type is in use, the following nodes also are required:

- **<timeInstant>**, ***integer, required field***, required and only used in the **'timeSlice'** type. Location of the time slice (integer index)
- **<numberOfSamples>**, ***integer, required field***, number of samples

If instead a `'min'`, `'max'`, `'average'`, or `'value'` is used, the following nodes are also required:

- **<pivotVar>**, ***string, required field***, Name of the chosen indexing variable (the variable whose min, max, average, or value is used to determine the time slice)
- **<pivotVal>**, ***float, optional field***, required for `'value'` type, the value for the chosen variable

Lastly, if a `'mixed'` approach is used, the following nodes apply:

- **<max>**, ***string, optional field***, the names of variables whose output should be their own maximum value within the history.
- **<min>**, ***string, optional field***, the names of variables whose output should be their own minimum value within the history.
- **<average>**, ***string, optional field***, the names of variables whose output should be their own average value within the history. Note that a **<pivotParameter>** node is required to perform averages.
- **<value>**, ***string, optional field***, the names of variables whose output should be taken at a time slice determined by another variable. As with the non-mixed `'value'` type, the first time the `pivotVar` crosses the specified `pivotVal` will be the time slice taken. This node requires two attributes, if used:
  - `pivotVar`, ***string, required field***, the name of the variable on which the time slice will be performed. That is, if we want the value of y when $t = 0.245$, this attribute would be `'t'`.
  - `pivotVal`, ***float, required field***, the value of the `pivotVar` on which the time slice will be performed. That is, if we want the value of y when $t = 0.245$, this attribute would be `'0.245'`.

  Note that all the outputs of the **<DataObject>** output of this postprocessor must be listed under one of the `'mixed'` node types in order for values to be returned.

Example (mixed): This example will output the average value of x for x, the value of y at time$= 0.245$ for y, and the value of z at $x = 4.0$ for z.

```
<Simulation>
  ...
  <Models>
    ...
```

```xml
    <PostProcessor name="mampp2"
       subType="InterfacedPostProcessor">
      <method>HistorySetSnapShot</method>
      <type>mixed</type>
      <average>x</average>
      <value pivotVar="time" pivotVal="0.245">y</value>
      <value pivotVar="x" pivotVal="4.0">z</value>
      <pivotParameter>time</pivotParameter>
      <extension>zeroed</extension>
    </PostProcessor>
    ...
  </Models>
  ...
</Simulation>
```

17.5.10.6 Method: HSPS

This Post-Processor performs the conversion from HistorySet to PointSet The conversion is made so that each history H is converted to a single point P. Assume that each history H is a dict of n output variables $x_1 = [...], x_n = [...]$, then the resulting point P is as follows; $P = [x_1, ..., x_n]$ Note: it is here assumed that all histories have been sync so that they have the same length, start point and end point. If you are not sure, do a pre-processing the the original history set.

In the **<PostProcessor>** input block, the following XML sub-nodes are required, independent of the **subType** specified (min, max, avg and value case):

- **<pivotParameter>**, ***string, optional field***, ID of the temporal variable (only for avg)

17.5.10.7 Method: TypicalHistoryFromHistorySet

This Post-Processor performs a simplified procedure of [6] to form a "typical" time series from multiple time series. The input should be a HistorySet, with each history in the HistorySet synchronized. For HistorySet that is not synchronized, use Post-Processor method **HistorySetSync** to synchronize the data before running this method.

Each history in input HistorySet is first converted to multiple histories each has maximum time specified in **<outputLen>** (see below). Each converted history H_i is divided into a set of subsequences $\{H_i^j\}$, and the division is guided by the **<subseqLen>** node specified in the input XML. The value of **<subseqLen>** should be a list of positive numbers that specify the length

of each subsequence. If the number of subsequence for each history is more than the number of values given in **<subseqLen>**, the values in **<subseqLen>** would be reused.

For each variable x, the method first computes the empirical CDF (cumulative density function) by using all the data values of x in the HistorySet. This CDF is termed as long-term CDF for x. Then for each subsequence H_i^j, the method computes the empirical CDF by using all the data values of x in H_i^j. This CDF is termed as subsequential CDF. For the first interval window (i.e., $j = 1$), the method computes the Finkelstein-Schafer (FS) statistics [7] between the long term CDF and the subsequential CDF of H_i^1 for each i. The FS statistics is defined as following.

$$FS = \sum_x FS_x$$

$$FS_x = \frac{1}{N}\sum_{n=1}^{N} \delta_n$$

where N is the number of value reading in the empirical CDF and δ_n is the absolute difference between the long term CDF and the subsequential CDF at value x_n. The subsequence H_i^1 with minimal FS statistics will be selected as the typical subsequence for the interval window $j = 1$. Such process repeats for $j = 2, 3, \dots$ until all subsequences have been processed. Then all the typical subsequences will be concatenated to form a complete history.

In the **<PostProcessor>** input block, the following XML sub-nodes are required, independent of the **subType** specified:

- **<pivotParameter>**, ***string, optional field***, ID of the temporal variable
  Default: Time
- **<subseqLen>**, ***integers, required field***, length of the divided subsequence (see above)
- **<outputLen>**, ***integer, optional field***, maximum value of the temporal variable for the generated typical history
  Default: Maximum value of the variable with name of **<pivotParameter>**

For example, consider history of data collected over three years in one-second increments, where the user wants a single *typical year* extracted from the data. The user wants this data constructed by combining twelve equal *typical month* segments. In this case, the parameter **<outputLen>** should be `31536000` (the number of seconds in a year), while the parameter **<subseqLen>** should be `2592000` (the number of seconds in a month). Using a value for **<subseqLen>** that is either much, much smaller than **<outputLen>** or of equal size to **<outputLen>** might have unexpected results. In general, we recommend using a **<subseqLen>** that is roughly an order of magnitude smaller than **<outputLen>**.

17.5.10.8 Method: dataObjectLabelFilter

This Post-Processor allows to filter the portion of a dataObject, either PointSet or HistorySet, with a given clustering label. A clustering algorithm associates a unique cluster label to each element of the dataObject (PointSet or HistorySet). This cluster label is a natural number ranging from 0 (or 1 depending on the algorithm) to N where N is the number of obtained clusters. Recall that some clustering algorithms (e.g., K-Means) receive N as input while others (e.g., Mean-Shift) determine N after clustering has been performed. Thus, this Post-Processor is naturally employed after a data-mining clustering techniques has been performed on a dataObject so that each clusters can be analyzed separately.

In the **<PostProcessor>** input block, the following XML sub-nodes are required, independently of the **subType** specified:

- **<label>**, ***string, required field***, name of the clustering label
- **<clusterIDs>**, ***integers, required field***, ID of the selected clusters. Note that more than one ID can be provided as input

17.5.10.9 Method: HSPS

The **<HSPS>** Post-Processor performs a filtering of the dataObject. This particular filtering is based on the labels generated by any clustering algorithm. Given the selected label, this Post-Processor filters out all histories or points having a different label. In the **<PostProcessor>** input block, the following XML sub-nodes are required:

- **<dataType>**, ***string, required field***, type of dataObject (HistorySet or PointSet)
- **<label>**, ***string, required field***, varaiable which contains the cluster labels
- **<clusterIDs>**, ***int, required field***, cluster labels considered

17.5.10.10 Method: Discrete Risk Measures

This Post-Processor calculates a series of risk importance measures from a PointSet. This calculation if performed for a set of input paramteres given an output target.

The user is required to provide the following information:

- the set of input variables. For each variable the following need to be specified:

  - the set of values that imply a reliability value equal to 1 for the input variable
  - the set of values that imply a reliability value equal to 0 for the input variable

- the output target variable. For this variable it is needed to specify the values of the output target variable that defines the desired outcome.

The following variables are first determined for each input variable i:

- R_0 Probability of the outcome of the output target variable (nominal value)
- R_i^+ Probability of the outcome of the output target variable if reliability of the input variable is equal to 0
- R_i^- Probability of the outcome of the output target variable if reliability of the input variable is equal to 1

Available measures are:

- Risk Achievement Worth (RAW): $RAW = R_i^+/R_0$
- Risk Achievement Worth (RRW): $RRW = R_0/R_i^-$
- Fussell-Vesely (FV): $FV = (R_0 - R_i^-)/R_0$
- Birnbaum (B): $B = R_i^+ - R_i^-$

In the **<PostProcessor>** input block, the following XML sub-nodes are required, independent of the **subType** specified:

- **<measures>**, ***string, required field***, desired risk importance measures that have to be computed (RRW, RAW, FV, B)
- **<variable>**, ***string, required field***, ID of the input variable. This node is provided for each input variable. This nodes needs to contain also these attributes:
  - **R0values**, ***float, required field***, interval of values (comma separated values) that implies a reliability value equal to 0 for the input variable
  - **R1values**, ***float, required field***, interval of values (comma separated values) that implies a reliability value equal to 1 for the input variable
- **<target>**, ***string, required field***, ID of the output variable. This nodes needs to contain also the attribute **values**, ***string, required field***, interval of values of the output target variable that defines the desired outcome

Example: This example shows an example where it is desired to calculate all available risk importance measures for two input variables (i.e., pumpTime and valveTime) given an output target variable (i.e., Tmax). A value of the input variable pumpTime in the interval $[0, 240]$ implies a reliability value of the input variable pumpTime equal to 0. A value of the input variable valveTime in the interval $[0, 60]$ implies a reliability value of the input variable valveTime equal to 0. A value of the input variables valveTime and pumpTime in the interval $[1441, 2880]$ implies a reliability value of the input variables equal to 1. The desired outcome of the output variable Tmax occurs in the interval $[2200, 2500]$.

```
<Simulation>
  ...
  <Models>
    ...
    <PostProcessor name="riskMeasuresDiscrete"
       subType="InterfacedPostProcessor">
      <method>RiskMeasuresDiscrete</method>
      <measures>B,FV,RAW,RRW</measures>
      <variable R0values='0,240'
         R1values='1441,2880'>pumpTime</variable>
      <variable R0values='0,60'
         R1values='1441,2880'>valveTime</variable>
      <target   values='2200,2500'
         >Tmax</target>
    </PostProcessor>
    ...
  </Models>
  ...
</Simulation>
```

This Post-Processor allows the user to consider also multiple datasets (a data set for each initiating event) and calculate the global risk importance measures. This can be performed by:

- Including all datasets in the step

```
<Simulation>
  ...
  </Steps>
    ...
    <PostProcess name="PP">
      <Input   class="DataObjects"  type="PointSet"
         >outRun1</Input>
      <Input   class="DataObjects"  type="PointSet"
         >outRun2</Input>
```

```
      <Model    class="Models"         type="PostProcessor"
         >riskMeasuresDiscrete</Model>
      <Output  class="DataObjects"   type="PointSet"
         >outPPS</Output>
      <Output  class="OutStreams"    type="Print"
         >PrintPPS_dump</Output>
    </PostProcess>
  </Steps>
  ...
</Simulation>
```

- Adding in the Post-processor the frequency of the initiating event associated to each dataset

```
<Simulation>
  ...
  <Models>
    ...
    <PostProcessor name="riskMeasuresDiscrete"
       subType="InterfacedPostProcessor">
      <method>riskMeasuresDiscrete</method>
      <measures>FV,RAW</measures>
      <variable R1values='-0.1,0.1'
         R0values='0.9,1.1'>Astatus</variable>
      <variable R1values='-0.1,0.1'
         R0values='0.9,1.1'>Bstatus</variable>
      <variable R1values='-0.1,0.1'
         R0values='0.9,1.1'>Cstatus</variable>
      <variable R1values='-0.1,0.1'
         R0values='0.9,1.1'>Dstatus</variable>
      <target    values='0.9,1.1'>outcome</target>
      <data      freq='0.01'>outRun1</data>
      <data      freq='0.02'>outRun2</data>
    </PostProcessor>
    ...
  </Models>
  ...
</Simulation>
```

This post-processor can be made time dependent if a single HistorySet is provided among the other data objects. The HistorySet contains the temporal profiles of a subset of the input variables. This temporal profile can be only boolean, i.e., 0 (component offline) or 1 (component online). Note that the provided history set must contains a single History; multiple Histories are

not allowed. When this post-processor is in a dynamic configuration (i.e., time-dependent), the user is required to specify an xml node **<temporalID>** that indicates the ID of the temporal variable. For each time instant, this post-processor determines the temporal profiles of the desired risk importance measures. Thus, in this case, an HistorySet must be chosen as an output data object. An example is shown below:

```
<Simulation>
  ...
  <Models>
    ...
    <PostProcessor name="riskMeasuresDiscrete"
       subType="InterfacedPostProcessor">
      <method>riskMeasuresDiscrete</method>
      <measures>B,FV,RAW,RRW,R0</measures>
      <variable R1values='-0.1,0.1'
         R0values='0.9,1.1'>Astatus</variable>
      <variable R1values='-0.1,0.1'
         R0values='0.9,1.1'>Bstatus</variable>
      <variable R1values='-0.1,0.1'
         R0values='0.9,1.1'>Cstatus</variable>
      <target   values='0.9,1.1'>outcome</target>
      <data     freq='1.0'>outRun1</data>
      <temporalID>time</temporalID>
    </PostProcessor>
    ...
  </Models>
  ...
  <Steps>
    ...
    <PostProcess name="PP">
      <Input     class="DataObjects"  type="PointSet"
         >outRun1</Input>
      <Input     class="DataObjects"  type="HistorySet"
         >timeDepProfiles</Input>
      <Model     class="Models"       type="PostProcessor"
         >riskMeasuresDiscrete</Model>
      <Output    class="DataObjects"  type="HistorySet"
         >outHS</Output>
      <Output    class="OutStreams"   type="Print"
         >PrintHS</Output>
    </PostProcess>
    ...
  </Steps>
```

```
  ...
</Simulation>
```

17.5.11 RavenOutput

The **RavenOutput** post-processor is specifically used to gather data from RAVEN output files and generate a PointSet suitable for plotting or other analysis. It can do this in two modes: static and dynamic. In static mode, the PostProcessor reads from from several static XML output files produced by RAVEN. In dynamic mode, the PostProcessor reads from a single dynamic XML output file and builds a PointSet where the pivot parameter (e.g. time) is the input and the requested values are returned for each of the pivot parameter values (e.g. points in time). The name for the pivot parameter will be taken directly from the XML structure. Note: by default the PostProcessor operates in static mode; to read a dynamic file, the **<dynamic>** node must be specified. In order to use the *RavenOutput* PP, the user needs to set the **subType** of a **<PostProcessor>** node:

```
<PostProcessor subType='RavenOutput'/>
```
.

Several sub-nodes are available:

- **<dynamic>**, ***string, optional field***, if included will trigger reading a single dynamic file instead of multiple static files, unless the text of this field is **'false'**, in which case it will return to the default (multiple static files).
  Default: (False)
- **<File>**, ***XML Node, required field*** For each file to be read by this postprocessor, an entry in the **<Files>** node must be added, and a **<File>** node must be added to the postprocessor input block. The **<File>** requires two identifying attributes:
  - **name**, ***string, required field***, the RAVEN-assigned name of the file,
  - **ID**, ***float, optional field***, the floating point ID that will be unique to this file. This will appear as an entry in the output **<DataObject>** and the corresponding column are the values extracted from this file. If not specified, RAVEN will attempt to find a suitable integer ID to use, and a warning will be raised.

  Each value that needs to be extracted from the file needs to be specified by one of the following **<output>** nodes within the **<File>** node:

  - **<output>**, ***—-separated string, required field***, the specification of the output to extract from the file. RAVEN uses `xpath` as implemented in Python's `xml.etree` module to specify locations in XML. For example, to search tags, use a path separated by forward slash characters ("/"), starting under the root; this means the root node

should not be included in the path. See the example. For more details on xpath options available, see https://docs.python.org/2/library/xml.etree.elementtree.html#xpath-support. The <output> node requires the following attribute:

* name, ***string, required field***, specifies the entry in the Data Object that this value should be stored under.

Example (Static): Using an example, let us have two input files, named *in1.xml* and *in2.xml*. They appear as follows. Note that the name of the variables we want changes slightly between the XML; this is fine.

in1.xml

```
<BasicStatistics>
  <ans>
    <val1>6</val1>
    <val2>7</val2>
  </ans>
</BasicStatistics>
```

in2.xml

```
<ROM>
  <ans>
    <first>6.1</first>
    <second>7.1</second>
  </ans>
</BasicStatistics>
```

The RAVEN input to extract this information would appear as follows:

```
<Simulation>
 ...
 <Files>
   <Input name='in1'>inp1.xml</Input>
   <Input name='in2'>inp2.xml</Input>
 </Files>
 ...
 <Models>
   ...
   <PostProcessor name='pp' subType='RavenOut'>
     <File name='in1' ID='1'>
       <output name='first'>ans/val1</output>
```

```
        <output name='second'>ans/val2</output>
      </File>
      <File name='in2' ID='2'>
        <output name='first'>ans/first</output>
        <output name='second'>ans/second</output>
      </File>
    </PostProcessor>
    ...
  </Models>
  ...
</Simulation>
```

Example (Dynamic): For a dynamic example, consider this time-evolution of values example. *inFile.xml* is a RAVEN dynamic XML output.

in1.xml

```
<BasicStatistics type='Dynamic'>
  <time value='0.0'>
    <ans>
      <val1>6</val1>
      <val2>7</val2>
    </ans>
  <\time>
  <time value='1.0'>
    <ans>
      <val1>9</val1>
      <val2>10</val2>
    </ans>
  <\time>
</BasicStatistics>
```

The RAVEN input to extract this information would appear as follows:

```
<Simulation>
  ...
  <Files>
    <Input name='inFile'>inFile.xml</Input>
  </Files>
  ...
  <Models>
    ...
    <PostProcessor name='pp' subType='RavenOut'>
```

```
      <dynamic>true</dynamic>
      <File name='inFile'>
        <output name='first'>ans|val1</output>
      </File>
    </PostProcessor>
    ...
  </Models>
  ...
</Simulation>
```

The resulting PointSet has *time* as an input and *first* as an output.

17.5.12 ETImporter

The **ETImporter** post-processor has been designed to import Event-Tree (ET) object into RAVEN. This is performed by saving the structure of the ET (from file) as a **PointSet** (only **PointSet** are allowed). Since an ET is a static Boolean logic structure, the **PointSet** is structured as follows:

- Input variables of the **PointSet** are the branching conditions of the ET. The value of each input variable can be:
  - 0: event did occur (typically upper branch)
  - 1: event did not occur (typically lower branch)
  - -1: event is not queried (no branching occured)
- Output variables of the **PointSet** are the ID of each branch of the ET (i.e., positive integers greater than 0)

Since several ET file formats are available, as of now only the OpenPSA format (see https://open-psa.github.io/joomla1.5/index.php.html) is supported. The ETImporter PP supports also:

- links to sub-trees
- by-pass branches
- symbolic definition of outcomes: typically outcomes are defined as either 0 (upper branch) or 1 (lower branch). If instead the ET uses the success/failure labels, then they are converted into 0/1 labels
- symbolic/numerical definition of sequences: if the ET contains a symbolic sequence then a .xml file is generated. This file contains the mapping between the sequences defined in

the ET and the numerical IDs created by RAVEN. The file name is the concatenation of the ET name and "_mapping". As an example the file "eventTree_mapping.xml" generated by RAVEN:

```
<map Tree="eventTree">
  <sequence ID="0">seq_1</sequence>
  <sequence ID="1">seq_2</sequence>
  <sequence ID="2">seq_3</sequence>
  <sequence ID="3">seq_4</sequence>
</map>
```

contains the mapping of four sequences defined in the ET (seq_1,seq_2,seq_3,seq_4) with the IDs generated by RAVEN (0,1,2,3). Note that if the sequences defined in the ET are both numerical and symbolic then they are all mapped.

The **<collect-formula>** are not considered since this node is used to connect the Boolean formulae generated by the Fault-Trees to the branch (i.e., fork) point.

In order to use the *ETImporter* PP, the user needs to set the **subType** of a **<PostProcessor>** node:

<PostProcessor subType='ETImporter'/>.

Several sub-nodes are available:

- **<fileFormat>**, ***string, required field***, specifies the format of the file that contains the ET structure (supported format: OpenPSA).

Example:

```
<Simulation>
 ...
  <Models>
    ...
    <PostProcessor name="ETImporter" subType="ETImporter">
      <fileFormat>OpenPSA</fileFormat>
    </PostProcessor>
    ...
  </Models>
 ...
  <Steps>
    ...
    <PostProcess name="import">
```

```
      <Input    class="Files"         type=""
        >eventTreeTest</Input>
      <Model    class="Models"        type="PostProcessor"
        >ETImporter</Model>
      <Output   class="DataObjects"   type="PointSet"
        >ET_PS</Output>
    </PostProcess>
    ...
</Simulation>
```

17.5.13 Metric

The **Metric** post-processor is specifically used to calculate the distance values among points and histories, while the **Metrics** block (See Chapter 19) allows the user to specify the similarity/dissimilarity metrics to be used in this post-processor. It is important to notice that this post-processor currently can only accept **PointSet** data object and does not accept **HistorySet** data. If the name of the variable is unique, it can be used, otherwise the variable can be specified with DataObjectName—InputOrOutput—variablename like other places in RAVEN. Some of the Metrics also accept distributions to calculate the distance against. These are specified by using the name of the distribution. In order to use the *Metric* PP, the user needs to set the **subType** of a **<PostProcessor>** node:

`<PostProcessor subType='Metric'/>`.

Several sub-nodes are available:

- **<Features>**, ***comma separated string, required field***, specifies the names of the features. This xml-node accepts the following attribute:
  - **type**, ***required string attribute***, the type of provided features. Currently only accept 'variable'.
- **<Targets>**, ***comma separated string, required field***, contains a comma separated list of the targets. **Note:** Each target is paired with a feature listed in xml node **<Features>**. In this case, the number of targets should be equal to the number of features. This xml-node accepts the following attribute:
  - **type**, ***required string attribute***, the type of provided features. Currently only accept 'variable'.
- **<Metric>**, ***string, required field***, specifies the **Metric** name that is defined via **Metrics** entity. In this xml-node, the following xml attributes need to be specified:

- `class`, ***required string attribute***, the class of this metric (e.g. Metrics)
- `type`, ***required string attribute***, the sub-type of this Metric (e.g. SKL, Minkowski)

Example:

```
<Simulation>
 ...
  <Models>
    ...
    <PostProcessor name="pp1" subType="Metric">
      <Features type="variable">ans</Features>
      <Targets type="variable">ans2</Targets>
      <Metric class="Metrics" type="SKL">euclidean</Metric>
      <Metric class="Metrics" type="SKL">rbf</Metric>
      <Metric class="Metrics" type="SKL">poly</Metric>
      <Metric class="Metrics" type="SKL">sigmoid</Metric>
      <Metric class="Metrics" type="SKL">polynomial</Metric>
      <Metric class="Metrics" type="SKL">linear</Metric>
      <Metric class="Metrics" type="SKL">cosine</Metric>
      <Metric class="Metrics" type="SKL">cityblock</Metric>
      <Metric class="Metrics" type="SKL">l1</Metric>
      <Metric class="Metrics" type="SKL">l2</Metric>
      <Metric class="Metrics" type="SKL">manhattan</Metric>
      <Metric class="Metrics" type="SKL">laplacian</Metric>
    </PostProcessor>
    ...
  </Models>
 ...
</Simulation>
```

17.5.14 CrossValidation

The **CrossValidation** post-processor is specifically used to evaluate estimator (i.e. ROMs) performance. Cross-validation is a statistical method of evaluating and comparing learning algorithms by dividing data into two portions: one used to 'train' a surrogate model and the other used to validate the model, based on specific scoring metrics. In typical cross-validation, the training and validation sets must crossover in successive rounds such that each data point has a chance of being validated against the various sets. The basic form of cross-validation is k-fold cross-validation. Other forms of cross-validation are special cases of k-fold or involve repeated rounds of k-fold cross-validation. **Note:** It is important to notice that this post-processor currently can only accept

PointSet data object. In order to use the *CrossValidation* PP, the user needs to set the **subType** of a **<PostProcessor>** node:

<PostProcessor subType='CrossValidation'/>.

Several sub-nodes are available:

- **<SciKitLearn>**, ***string, required field***, the subnodes specifies the necessary information for the algorithm to be used in the post-processor. 'SciKitLearn' is based on algorithms in SciKit-Learn library, and currently it performs cross-validation over **PointSet** only.
- **<Metric>**, ***string, required field***, specifies the **Metric** name that is defined via **Metrics** entity. In this xml-node, the following xml attributes need to be specified:
  - **class**, ***required string attribute***, the class of this metric (e.g. Metrics)
  - **type**, ***required string attribute***, the sub-type of this Metric (e.g. SKL, Minkowski)

  Note: Currently, cross-validation post-processor only accepts **<SKL>** metrics with **<metricType>** **'mean_absolute_error'**, **'explained_variance_score'**, **'r2_score'**, **'mean_squared_e**
  and **'median_absolute_error'**.

17.5.14.1 SciKitLearn

The algorithm for cross-validation is chosen by the subnode **<SKLtype>** under the parent node **<SciKitLearn>**. In addition, a special subnode **<average>** can be used to obtain the average cross validation results.

- **<SKLtype>**, ***string, required field***, contains a string that represents the cross-validation algorithm to be used. As mentioned, its format is:

  <SKLtype>algorithm**</SKLtype>**.
- **<average>**, ***boolean, optional field***, if 'True', dump the average cross validation results into the output files.

Based on the **<SKLtype>** several different algorithms are available. In the following paragraphs a brief explanation and the input requirements are reported for each of them.

17.5.14.2 K-fold

KFold divides all the samples in k groups of samples, called folds (if $k = n$, this is equivalent to the **Leave One Out** strategy), of equal sizes (if possible). The prediction function is learned using

$k - 1$ folds, and fold left out is used for test. In order to use this algorithm, the user needs to set the subnode: **<SKLtype>**KFold**</SKLtype>**. In addition to this XML node, several others are available:

- **<n_splits>**, ***integer, optional field***, number of folds, must be at least 2.
  Default: 3
- **<shuffle>**, ***boolean, optional field***, whether to shuffle the data before splitting into batches.
- **<random_state>**, ***None, integer or RandomState, optional field***, when shuffle=True, pseudo-random number generator state used for shuffling. If None, use default numpy RNG for shuffling.

17.5.14.3 Stratified k-fold

StratifiedKFold is a variation of *k-fold* which returns stratified folds: each set contains approximately the same percentage of samples of each target class as the complete set. In order to use this algorithm, the user needs to set the subnode:

<SKLtype>StratifiedKFold**</SKLtype>**.

In addition to this XML node, several others are available:

- **<y>**, ***array-like, [n_samples], required field***, samples to split in K folds.
- **<n_splits>**, ***integer, optional field***, number of folds, must be at least 2.
  Default: 3
- **<shuffle>**, ***boolean, optional field***, whether to shuffle the data before splitting into batches.
- **<random_state>**, ***None, integer or RandomState, optional field***, when shuffle=True, pseudo-random number generator state used for shuffling. If None, use default numpy RNG for shuffling.

17.5.14.4 Label k-fold

LabelKFold is a variation of *k-fold* which ensures that the same label is not in both testing and training sets. This is necessary for example if you obtained data from different subjects and you want to avoid over-fitting (i.e., learning person specific features) by testing and training on different subjects. In order to use this algorithm, the user needs to set the subnode:

<SKLtype>LabelKFold**</SKLtype>**.

In addition to this XML node, several others are available:

- **<labels>**, ***array-like with shape (n_samples,), required field***, contains a label for each sample. The folds are built so that the same label does not appear in two different folds.
- **<n_splits>**, ***integer, optional field***, number of folds, must be at least 2.
  Default: 3

17.5.14.5 Leave-One-Out - LOO

LeaveOneOut (or LOO) is a simple cross-validation. Each learning set is created by taking all the samples except one, the test set being the sample left out. Thus, for n samples, we have n different training sets and n different tests set. This is cross-validation procedure does not waste much data as only one sample is removed from the training set. In order to use this algorithm, the user needs to set the subnode:

<SKLtype>LeaveOneOut**</SKLtype>**.

17.5.14.6 Leave-P-Out - LPO

LeavePOut is very similar to **LeaveOneOut** as it creates all the possible training/test sets by removing p samples from the complete set. For n samples, this produces $\binom{n}{p}$ train-test pairs. Unlike **LeaveOneOut** and **KFold**, the test sets will overlap for $p > 1$. In order to use this algorithm, the user needs to set the subnode:

<SKLtype>LeavePOut**</SKLtype>**.

In addition to this XML node, several others are available:

- **<p>**, ***integer, required field***, size of the test sets

17.5.14.7 Leave-One-Label-Out - LOLO

LeaveOneLabelOut (LOLO) is a cross-validation scheme which holds out the samples according to a third-party provided array of integer labels. This label information can be used to encode arbitrary domain specific pre-defined cross-validation folds. Each training set is thus constituted by all samples except the ones related to a specific label. In order to use this algorithm, the user needs to set the subnode:

<SKLtype>LeaveOneLabelOut**</SKLtype>**.

In addition to this XML node, several others are available:

- **<labels>**, ***array-like of integer with shape (n_samples,), required field***, arbitrary domain-specific stratificatioin of the data to be used to draw the splits.

17.5.14.8 Leave-P-Label-Out

LeavePLabelOut is imilar as *Leave-One-Label-Out*, but removes samples related to P labels for each training/test set. In order to use this algorithm, the user needs to set the subnode:

<SKLtype>LeavePLabelOut**</SKLtype>**.

In addition to this XML node, several others are available:

- **<labels>**, ***array-like of integer with shape (n_samples,), required field***, arbitrary domain-specific stratificatioin of the data to be used to draw the splits.
- **<p>**, ***integer, optional field***, number of samples to leave out in the test split.

17.5.14.9 ShuffleSplit

ShuffleSplit iterator will generate a user defined number of independent train/test dataset splits. Samples are first shuffled and then split into a pair of train and test sets. it is possible to control the randomness for reproducibility of the results by explicitly seeding the **<random_state>** pseudo random number generator. In order to use this algorithm, the user needs to set the subnode:

<SKLtype>ShuffleSplit**</SKLtype>**.

In addition to this XML node, several others are available:

- **<n_iter>**, ***integer, optional field***, number of re-shuffling and splitting iterations
  Default: 10.
- **<test_size>**,***float, integer or None***, if float, should be between 0.0 and 1.0 and represent the proportion of the dataset to include in the test split.
  Default: 0.1 If integer, represents the absolute number of test samples. If None, the value is automatically set to the complement of the train size.
- **<train_size>**, ***float, integer or None***, if float, should be between 0.0 and 1.0 and represent the proportion of the dataset to include in the train split. If integer, represents the absolute number of train samples. If None, the value is automatically set to the complement

of the test size.
Default: None

- **<random_state>**, ***None, integer or RandomState, optional field***, when shuffle=True, pseudo-random number generator state used for shuffling. If None, use default numpy RNG for shuffling.

17.5.14.10 Label-Shuffle-Split

LabelShuffleSplit iterator behaves as a combination of **ShuffleSplit** and **LeavePLabelOut**, and generates a sequence of randomized partitions in which a subset of labels are held out for each split. In order to use this algorithm, the user needs to set the subnode:

<SKLtype>LabelShuffleSplit**</SKLtype>**.

In addition to this XML node, several others are available:

- **<labels>**, ***array, [n_samples]***, labels of samples.
- **<n_iter>**, ***integer, optional field***, number of re-shuffling and splitting iterations
Default: 10.
- **<test_size>**,***float, integer or None***, if float, should be between 0.0 and 1.0 and represent the proportion of the dataset to include in the test split.
Default: 0.1 If integer, represents the absolute number of test samples. If None, the value is automatically set to the complement of the train size.
- **<train_size>**, ***float, integer or None***, if float, should be between 0.0 and 1.0 and represent the proportion of the dataset to include in the train split. If integer, represents the absolute number of train samples. If None, the value is automatically set to the complement of the test size.
Default: None
- **<random_state>**, ***None, integer or RandomState, optional field***, when shuffle=True, pseudo-random number generator state used for shuffling. If None, use default numpy RNG for shuffling.

Example:

```
<Simulation>
 ...
  <Files>
    <Input name="output_cv" type="">output_cv.xml</Input>
```

```xml
    <Input name="output_cv.csv" type="">output_cv.csv</Input>
  </Files>
  <Models>
    ...
    <ROM name="surrogate" subType="SciKitLearn">
      <SKLtype>linear_model|LinearRegression</SKLtype>
      <Features>x1,x2</Features>
      <Target>ans</Target>
      <fit_intercept>True</fit_intercept>
      <normalize>True</normalize>
    </ROM>
    <PostProcessor name="pp1" subType="CrossValidation">
        <SciKitLearn>
            <SKLtype>KFold</SKLtype>
            <n_splits>3</n_splits>
            <shuffle>False</shuffle>
            <random_state>None</random_state>
        </SciKitLearn>
        <Metric class="Metrics" type="SKL">m1</Metric>
    </PostProcessor>
    ...
  </Models>
  <Metrics>
    <SKL name="m1">
      <metricType>mean_absolute_error</metricType>
    </SKL>
  </Metrics>
  <Steps>
    <PostProcess name="PP1">
        <Input class="DataObjects"
           type="PointSet">outputDataMC</Input>
        <Input class="Models" type="ROM">surrogate</Input>
        <Model class="Models" type="PostProcessor">pp1</Model>
        <Output class="Files" type="">output_cv</Output>
        <Output class="Files" type="">output_cv.csv</Output>
    </PostProcess>
  </Steps>
 ...
</Simulation>
```

17.6 EnsembleModel

As already mentioned, the **EnsembleModel** is able to combine **Code**(see 17.1), **ExternalModel**(see 17.4) and **ROM**(see 17.4) Models.
It is aimed to create a chain of Models (whose execution order is determined by the Input/Output relationships among them). If the relationships among the models evolve in a non-linear system, a Picard's Iteration scheme is employed.
Currently this model is able to share information (i.e. data) using both **PointSet** and **HistorySet**.

The specifications of a EnsembleModel must be defined within the XML block **<EnsembleModel>**. This XML node needs to contain the attributes:

- **name**, ***required string attribute***, user-defined name of this EnsembleModel. **Note:** As with the other objects, this is the name that can be used to refer to this specific entity from other input blocks in the XML.
- **subType**, ***required string attribute***, must be kept empty.

Within the **<EnsembleModel>** XML node, the multiple Models that constitute this EnsembleModel needs to be inputted. Each Model is specified within a **<Model>** block (**Note:** each model here specified need to be inputted in the**<Models>** main XML block) :

- **<Model>**, ***XML node, required parameter***. The text portion of this node needs to contain the name of the Model
  This XML node needs to contain the attributes:
  - **class**, ***required string attribute***, the class of this sub-model (e.g. Models)
  - **type**, ***required string attribute***, the sub-type of this Model (e.g. ExternalModel, ROM, Code)

  In addition the following XML sub-nodes need to be inputted (or optionally inputted):

  - **<TargetEvaluation>**, ***string, required field***, represents the container where the output of this Model are stored. From a practical point of view, this XML node must contain the name of a data object defined in the **<DataObjects>** block (see Section 14). Currently, the **<EnsembleModel>** accept "DataObjects" both of type "PointSet" and "HistorySet". **Note:** The **<TargetEvaluation>** is primary used for input-output identification. If the linked DataObject is not placed as additional output of the Step where the EnsembleModel is used, it will not be filled with the data coming from the calculation and it will be kept empty.
  - **<Input>**, ***string, required field***, represents the input entities that need to be passed to this sub-model The user can specify as many **<Input>** as required by the sub-model. **Note:** All the inputs here specified need to be listed in the Steps where the EnsembleModel is used.

- <Output>, ***string, optional field***, represents the output entities that need to be linked to this sub-model. **Note:** The <Output>s here specified are not part of the determination of the EnsembleModel execution but represent an additional storage of results from the sub-models. For example, if the <TargetEvaluation> is of type PointSet (since just scalar data needs to be transferred to other models) and the sub-model is able to also output history-type data, this Output can be of type HistorySet. Note that the structure of each Output dataObject must include only variables (either input or output) that are defined among the model. As an example, the Output dataObjects cannot contained variables that are defined at the Ensemble model level. The user can specify as many <Output> (s) as needed. The optional <Output>s can be of both classes "DataObjects" and "Databases" (e.g. PointSet, HistorySet, HDF5). **Note: The <Output> (s) here specified MUST be listed in the Step in which the EnsembleModel is used.**

It is important to notice that when the EnsembleModel detects a chain of models that evolve in a non-linear system, a Picard's Iteration scheme is activated. In this case, an additional XML sub-node within the main <EnsembleModel> XML node needs to be specified:

- <settings>, ***XML node, required parameter (if Picard's activated)***. The body of this sub-node contains the following XML sub-nodes:
  - <maxIterations>, ***integer, optional field***, maximum number of Picard's iteration to be performed (in case the iteration scheme does not previously converge).
    Default: 30;
  - <tolerance>, ***float, optional field***, convergence criterion. It represents the L2 norm residue below which the Picard's iterative scheme is considered converged.
    Default: 0.001;
  - <initialConditions>, ***XML node, required parameter (if Picard's activated)***, Within this sub-node, the initial conditions for the input variables (that are part of a loop) need to be specified in sub-nodes named with the variable name (e.g. <varName>). The body of the <varName> contains the value of the initial conditions (scalar or arrays, depending of the type of variable). If an array needs to be inputted, the user can specify the attribute repeat and the code is going to repeat for repeat-times the value inputted in the body.
  - <initialStartModels>, ***XML node, only required parameter when Picard's iteration is activated***, specifies the list of models that will be initially executed. **Note:** Do not input this node for non-Picard calculations, otherwise an error will be raised.

Example (Linear System):

```
<Simulation>
  ...
```

```xml
  <Models>
    ...
    <EnsembleModel name="heatTransferEnsembleModel" subType="">
      <Model class="Models" type="ExternalModel">
        thermalConductivityComputation
        <TargetEvaluation class="DataObjects" type="PointSet">
          thermalConductivityComputationContainer
        </TargetEvaluation>
        <Input class="DataObjects" type="PointSet">
          inputHolder
        </Input>
      </Model>
      <Model class="Models" type="ExternalModel" >
          heatTransfer
          <TargetEvaluation class="DataObjects" type="PointSet">
            heatTransferContainer
          </TargetEvaluation>
        <Input class="DataObjects" type="PointSet">
          inputHolder
        </Input>
        <Output class="DataObjects" type="HistorySet">
          thisModelLinkedOutput
        </Output>
        <Output class="Databases" type="HDF5">
          thisModelLinkedHDF5
        </Output>
      </Model>
    </EnsembleModel>
    ...
  </Models>
  ...
</Simulation>
```

Example (Non-Linear System):

```xml
<Simulation>
  ...
  <Models>
    ...
    <EnsembleModel name="heatTransferEnsembleModel" subType="">
      <settings>
        <maxIterations>8</maxIterations>
         <tolerance>0.01</tolerance>
```

```
        <initialConditions>
          <!-- the value 0.7 is going to be repeated 10 times
             in order to create  an array for var1 -->
          <var1 repeat="10">0.7</var1>
          <!-- an array for var2 has been inputted -->
          <var2> 0.5 0.3 0.4</var2>
          <!-- a scalar for var3 has been inputted -->
          <var3> 45.0</var3>
        </initialConditions>
      </settings>

      <Model class="Models" type="ExternalModel">
        thermalConductivityComputation
        <TargetEvaluation class="DataObjects" type="PointSet">
          thermalConductivityComputationContainer
        </TargetEvaluation>
        <Input class="DataObjects" type="PointSet">
          inputHolder
        </Input>
      </Model>
      <Model class="Models" type="ExternalModel" >
          heatTransfer
          <TargetEvaluation class="DataObjects" type="PointSet">
            heatTransferContainer
          </TargetEvaluation>
        <Input class="DataObjects" type="PointSet">
          inputHolder
        </Input>
      </Model>
    </EnsembleModel>
    ...
  </Models>
  ...
</Simulation>
```

17.7 HybridModel

The **HybridModel** is a new *Model* entity. This new Model is able to combine reduced order model (ROMs) and any other high-fidelity Model (i.e. Code, ExternalModel). The ROMs will be trained based on the results from the high-fidelity model. The accuracy of the ROMs will be evaluated based on the cross validation scores, and the validity of the ROMs will be determined via some

local validation metrics (**Note:** currently only one metric is available, i.e. CrowdingDistance). After these ROMs are trained, the **HybridModel** can decide which of the Model (i.e the ROMs or high-fidelity model) to be executed based on the accuracy and validity of the ROMs.

Currently this model is only able to share information (i.e. data) using **PointSet**.

The specifications of a HybridModel must be defined within the XML block **<HybridModel>**. This XML node needs to contain the attributes:

- **name**, ***required string attribute***, user-defined name of this HybridModel. **Note:** As with the other objects, this is the name that can be used to refer to this specific entity from other input blocks in the XML.
- **subType**, ***required string attribute***, must be kept empty.

Within the **<HybridModel>** XML node, the multiple entities that constitute this HybridModel needs to be inputted.

- **<Model>**, ***XML node, required parameter***. The text portion of this node needs to contain the name of the Model This XML node needs to contain the attributes:
  - **class**, ***required string attribute***, the main "class" of the Model.
  - **type**, ***required string attribute***, the sub-type of the Model.
- **<ROM>**, ***XML node, required parameter***. The text portion of this node needs to contain the name of the ROM The user can specify as many **<ROM>** as required by the **<Model>**. **Note:** The outputs of each ROM should be different, and the total set of ROMs' outputs should be the same as the set of *Model's* outputs. This XML node needs to contain the attributes:
  - **class**, ***required string attribute***, the main "class" of the Model.
  - **type**, ***required string attribute***, the sub-type of the Model.
- **<CV>**, ***XML node, required parameter***. The text portion of this node needs to contain the name of the **<PostProcessor>** with **subType** "CrossValidation". This XML node needs to contain the attributes:
  - **class**, ***required string attribute***, the main "class" of the Model.
  - **type**, ***required string attribute***, the sub-type of the Model.
- **<TargetEvaluation>**, ***XML node, required parameter***. The text portion of this node needs to contain the name of a data object defined in the **<DataObjects>** block. **Note:** currently only accept data object with type "PointSet". The **<TargetEvaluation>** is primary used for training ROMs. **Note:** The linked DataObject should be placed as additional output of the Step where the **HybridModel** is used. This XML node needs to contain the attributes:

- **class**, ***required string attribute***, the main "class" of the DataObjects.
- **type**, ***required string attribute***, the sub-type of the DataObjects.

An additional XML sub-node within the main **<HybridModel>** XML node needs to be specified:

- **<settings>**, ***XML node, optional parameter***. The body of this sub-node contains the following XML sub-nodes:
  - **<minInitialTrainSize>**, ***integer, optional field***, the minimum initial number of high-fidelity model runs before starting train the ROMs.
    Default: 10;
  - **<tolerance>**, ***float, optional field***, ROMs convergence criterion indicates the displacement from the optimum results of cross validation. In other words, small tolerance indicates tight convergence criterion of the ROMs, while large tolerance indicates loose convergence criterion of the ROMs. **Note:** Currently, this tolerance can be only used for cross validations with SKL Metrics: *explained_variance_score*, *r2_score*, *median_absolute_error*, *mean_squared_error* and *mean_absolute_error*.
    Default: 0.01;
  - **<maxTrainSize>**, ***XML node, optional field***, the maximum size of training set of ROMs.
    Default: 1.0E6
- **<validationMethod>**, ***XML node, optional parameter***. The validity methods that are used to determine which model to run (i.e. ROMs or high-fidelity Model). This XML node needs to contain the attributes:
  - **name**, ***required string attribute***, user-defined name of this **<validationMethod>**. **Note:** Currently, only one method is available, ie. "CrowdingDistance".

  The body of this sub-node contains the following XML sub-nodes:
  - **<threshold>**, ***XML node, required field***, the threshold that is used for "CrowdingDistance" method.

Example (ExternalModel):

```
<Simulation>
  ...
  <Metrics>
    <SKL name="m1">
      <metricType>mean_absolute_error</metricType>
    </SKL>
```

```xml
</Metrics>

<Models>
  <ExternalModel ModuleToLoad="EM2linear"
     name="thermalConductivityComputation" subType="">
    <variables>leftTemperature,rightTemperature,k,averageTemperature</var
  </ExternalModel>
  <ROM name="knr" subType="SciKitLearn">
    <SKLtype>neighbors|KNeighborsRegressor</SKLtype>
    <Features>leftTemperature, rightTemperature</Features>
    <Target>k</Target>
    <n_neighbors>5</n_neighbors>
    <weights>uniform</weights>
    <algorithm>auto</algorithm>
    <leaf_size>30</leaf_size>
    <metric>minkowski</metric>
    <p>2</p>
  </ROM>
  <PostProcessor name="pp1" subType="CrossValidation">
      <SciKitLearn>
          <SKLtype>KFold</SKLtype>
          <n_splits>10</n_splits>
          <shuffle>False</shuffle>
          <random_state>None</random_state>
      </SciKitLearn>
      <Metric class="Metrics" type="SKL">m1</Metric>
  </PostProcessor>
  <HybridModel name="hybrid" subType="">
      <Model class="Models"
         type="ExternalModel">thermalConductivityComputation</Model>
      <ROM class="Models" type="ROM">knr</ROM>
      <TargetEvaluation class="DataObjects"
         type="PointSet">thermalConductivityComputationContainer</TargetEv
      <CV class="Models" type="PostProcessor">pp1</CV>
      <settings>
          <tolerance>0.01</tolerance>
          <trainStep>1</trainStep>
          <maxTrainSize>1000</maxTrainSize>
          <initialTrainSize>10</initialTrainSize>
      </settings>
      <validationMethod name="CrowdingDistance">
          <threshold>0.2</threshold>
```

```
      </validationMethod>
    </HybridModel>
  </Models>
  ...
</Simulation>
```

Example (Code):

```
<Simulation>
  ...
  <Metrics>
    <SKL name="m1">
      <metricType>mean_absolute_error</metricType>
    </SKL>
  </Metrics>

  <Models>
    <Code name="poly" subType="GenericCode">
      <executable>runCode/poly_inp_io.py</executable>
      <clargs arg="python" type="prepend"/>
      <clargs arg="-i" extension=".one" type="input"/>
      <fileargs arg="aux" extension=".two" type="input"/>
      <fileargs arg="output" type="output"/>
      <prepend>python</prepend>
    </Code>
    <ROM name="knr" subType="SciKitLearn">
      <SKLtype>neighbors|KNeighborsRegressor</SKLtype>
      <Features>x, y</Features>
      <Target>poly</Target>
      <n_neighbors>5</n_neighbors>
      <weights>uniform</weights>
      <algorithm>auto</algorithm>
      <leaf_size>30</leaf_size>
      <metric>minkowski</metric>
      <p>2</p>
    </ROM>
    <PostProcessor name="pp1" subType="CrossValidation">
        <SciKitLearn>
            <SKLtype>KFold</SKLtype>
            <n_splits>10</n_splits>
            <shuffle>False</shuffle>
            <random_state>None</random_state>
        </SciKitLearn>
```

```xml
        <Metric class="Metrics" type="SKL">m1</Metric>
    </PostProcessor>
    <HybridModel name="hybrid" subType="">
        <Model class="Models" type="Code">poly</Model>
        <ROM class="Models" type="ROM">knr</ROM>
        <TargetEvaluation class="DataObjects"
           type="PointSet">samples</TargetEvaluation>
        <CV class="Models" type="PostProcessor">pp1</CV>
        <settings>
            <tolerance>0.1</tolerance>
            <trainStep>1</trainStep>
            <maxTrainSize>1000</maxTrainSize>
            <initialTrainSize>10</initialTrainSize>
        </settings>
        <validationMethod name="CrowdingDistance">
            <threshold>0.2</threshold>
        </validationMethod>
    </HybridModel>
  </Models>
  ...
  <Steps>
    <MultiRun name="hybridModelCode">
      <Input class="Files" type="">gen.one</Input>
      <Input class="Files" type="">gen.two</Input>
      <Input class="DataObjects"
         type="PointSet">inputHolder</Input>
      <Model class="Models" type="HybridModel">hybrid</Model>
      <Sampler class="Samplers" type="Stratified">LHS</Sampler>
      <Output class="DataObjects"
         type="PointSet">samples</Output>
      <Output class="OutStreams" type="Print">samples</Output>
    </MultiRun>
  </Steps>
  ...
</Simulation>
```

Note: For this example, the user needs to provide all the inputs for the **HybridModel**, i.e. Files for the **Code** and DataObject for the **ROM** defined in the **HybridModel**.

18 Functions

The RAVEN code provides support for the usage of user-defined external functions. These functions are python modules, with a format that is automatically interpretable by the RAVEN framework. For example, users can define their own method to perform a particular post-processing activity and the code will be embedded and use the function as though it were an active part of the code itself. In this section, the XML input syntax and the format of the accepted functions are fully specified.

The specifications of an external function must be defined within the XML block **<External>**. This XML node requires the following attributes:

- **name**, ***required string attribute***, user-defined name of this function. **Note:** As with other objects, this name can be used to refer to this specific entity from other input blocks in the XML.
- **file**, ***required string attribute***, absolute or relative path specifying the code associated to this function. **Note:** If a relative path is specified, it must be relative with respect to where the user is running the instance of RAVEN.

In order to make the RAVEN code aware of the variables the user is going to manipulate/use in her/his own python function, the variables need to be specified in the **<External>** input block. The user needs to input, within this block, only the variables directly used by the external code and not the local variables that the user does not want, for example, those stored in a RAVEN internal object. These variables are named within consecutive **<variable>** XML nodes:

- **<variable>**, ***string, required parameter***, in the body of this XML node, the user needs to specify the name of the variable. This variable needs to match a variable used/defined in the external python function.

When the external function variables are defined, at runtime, RAVEN initializes them and keeps track of their values during the simulation. Each variable defined in the **<External>** block is available in the function as a python `self.` member. In the following, an example of a user-defined external function is reported (a python module and its related XML input specifications).

Example Python Function:

```python
import numpy as np
def residuumSign(self):
  if self.var1 < self.var2 :
    return  1
  else:
    return -1
```

Example XML Input:

```xml
  ...
  <Functions>
    ...
    <External name='whatever' file='path_to_python_file'>
     ...
     <variable>var1</variable>
     <variable>var2</variable>
     ...
    </External>
    ...
  </Functions>
   ...
</Simulation>
```

19 Metrics

The Metrics block allows the user to specify the similarity/dissimilarity metrics to be used in specific clustering algorithms available in RAVEN. These metrics are used to calculate the distance values among points and histories. The data mining algorithms in RAVEN which accept the definition of a metric are the following:

- DBSCAN (see Section 17.5.9.3.6)
- Affinity Propagation (see Section 17.5.9.3.3)
- Metric post processor (See Secton 17.5.13)

Both of these algorithms can take as input an $N \times N$ square matrix $D = [d_{ij}]$ where each element d_{ij} of D is the distance between element i $(i = 1, \ldots, N)$ and j $(j = 1, \ldots, N)$.

Available metrics are as follows:

- Minkowski (see Section 19.1)

In the RAVEN input file these metrics are defined as follows:

```
<Simulation>
  ...
  <Metrics>
    ...
    <MetricID name='metricName'>
      ...
     <param1>value</param1>
      ...
    </MetricID>
    ...
  </Metrics>
  ...
</Simulation>
```

19.1 Minkowski

Minkowski distance is the most basic and used metric in any data mining algorithm. Given two multi-dimensional vectors $X = (x_1, x_2, \ldots, x_n)$ and $Y = (y_1, y_2, \ldots, y_n)$, the Minkowski distance

d_p of order p between these two vectors is:

$$d_p = \left(\sum_{i=1}^{n} \|x_i - y_i\|^p \right)^{\frac{1}{p}} \quad (35)$$

Minkowski distance is typically used with p being 1 or 2. The latter one is the Euclidean distance, while the former is sometimes known as the Manhattan distance. Note that this metric can be employed for both PointSets and HistorySets.

Note that this metric can be employed directly on HistorySets. For HistorySets, all histories must contain an equal number of samples. Note that, in this respect, the interfaced Post-Processor (see Section 17.5.10) HistorySetSampling can be employed to perform such action. In addition, several more interfaced Post-Processors can be used to manipulate HistorySets.

The specifications of a Minkowski distance must be defined within the XML block **<Minkowski>**. This XML node needs to contain the attributes:

- **<p>**, ***float, required field***, value for the parameter p
- **<pivotParameter>**, ***string, optional field***, the ID of the temporal variable; this is required in case the metric is applied to historysets instead of pointsets

An example of Minkowski distance defined in RAVEN is provided below:

```
<Simulation>
  ...
  <Metrics>
    ...
    <Minkowski name="example" subType="">
      <p>2</p>
      <pivotParameter>time</pivotParameter>
    </Minkowski>
    ...
  </Metrics>
  ...
</Simulation>
```

19.2 Dynamic Time Warping

The Dynamic Time Warping (DTW) is a distance metrice that can be employed only for HistorySets (i.e., time dependent data).

The specifications of a DTW distance must be defined within the XML block. **<DTW>**.

This XML node needs to contain the attributes:

- **<order>**, ***int, required field***, order of the DTW calculation: 0 specifices a classical DTW caluclation and 1 specifies a derivative DTW calculation
- **<pivotParameter>**, ***string, optional field***, the ID of the temporal variable
- **<localDistance>**, ***string, optional field***, the ID of the distance function to be employed to determine the local distance evaluation of two time series. Available options are provided by the sklearn pairwise_distances (cityblock, cosine, euclidean, $l1$, $l2$, manhattan, braycurtis, canberra, chebyshev, correlation, dice, hamming, jaccard, kulsinski, mahalanobis, matching, minkowski, rogerstanimoto, russellrao, seuclidean, sokalmichener, sokalsneath, sqeuclidean, yule)

An example of Minkowski distance defined in RAVEN is provided below:

```
<Simulation>
  ...
  <Metrics>
    ...
    <DTW name="example" subType="">
      <order>0</order>
      <pivotParameter>time</pivotParameter>
      <localDistance>euclidean</localDistance>
    </DTW>
    ...
  </Metrics>
  ...
</Simulation>
```

19.3 SKL Metrics

This Metric class interfaces directly with the metric distances available within scikit-learn. Note that these distance metrics apply only to PointSets. However, note that this distance metrics can be applied to HistorySets after the HistorySets are converted into PointSets. In this respect, the HSPS interfaced Post-Processor (see Section 17.5.10) can be employed to perform such conversion.

The specifications of a SKL metric must be defined within the XML block **<SKL>**. This XML node needs to contain the subnode:

- **<metricType>**, ***string, required field***, the type of **SKL** metrics from the SciKit-Learn.

Available metrics from SKL are:

- From scikit-learn pairwise kernels:
  - rbf: $k(x, y) = \exp(-\gamma\|x - y\|^2)$, parameters required: gamma
  - sigmoid: $k(x, y) = \tanh(\gamma x^\top y + c_0)$, parameters required: gamma and coef0
  - polynomial: $k(x, y) = (\gamma x^\top y + c_0)^d$, parameters required: gamma, degree, coef0
  - laplacian: $k(x, y) = \exp(-\gamma\|x - y\|_1)$, parameters required: gamma
  - linear: $k(x, y) = x^\top y$
  - cosine: $k(x, y) = \frac{xy^\top}{\|x\|\|y\|}$
  - $chi2$: $k(x, y) = \exp\left(-\gamma \sum_i \frac{(x[i]-y[i])^2}{x[i]+y[i]}\right)$
  - $additive_chi2$: $k(x, y) = -\sum[(x - y)^2/(x + y)]$
- From scikit-learn pairwise distances:
  - euclidean: $||u - v||_2$
  - cityblock: $\sum_i |u_i - v_i|$
  - l1: see cityblock
  - l2: see euclidean
  - manhattan: see cityblock
- From scikit-learn regression metrics:
  - explained_variance_score: $1.0 - \frac{Var[x-y]}{Var[x]}$
  - mean_absolute_error: $\frac{1}{n_{samples}} \sum_{i=0}^{n_{samples}-1} |x_i - y_i|$
  - mean_squared_error: $\frac{1}{n_{samples}} \sum_{i=0}^{n_{samples}-1} (x_i - y_i)^2$
  - median_absolute_error: $median(|x_i - y_i|, \cdots, |x_n - y_n|)$
  - r2_score: $1.0 - \frac{\sum_{i=0}^{n_{samples}-1} (x_i-y_i)^2}{\sum_{i=0}^{n_{samples}-1} (x_i-mean[x])^2}$

In the RAVEN input file these metrics are defined as follows:

```xml
<Simulation>
  ...
  <Metrics>
    <SKL name="euclidean">
        <metricType>euclidean</metricType>
    </SKL>
    <SKL name="laplacian">
        <metricType>laplacian</metricType>
        <gamma>0.5</gamma>
    </SKL>
    <SKL name="rbf">
        <metricType>rbf</metricType>
        <gamma>0.5</gamma>
    </SKL>
    <SKL name="poly">
        <metricType>poly</metricType>
        <gamma>1.0</gamma>
        <degree>2.0</degree>
        <coef0>1.0</coef0>
    </SKL>
    <SKL name="sigmoid">
        <metricType>sigmoid</metricType>
        <gamma>1.0</gamma>
        <coef0>1.0</coef0>
    </SKL>
    <SKL name="polynomial">
        <metricType>polynomial</metricType>
        <gamma>1.0</gamma>
        <degree>2.0</degree>
        <coef0>1.0</coef0>
    </SKL>
    <SKL name="linear">
        <metricType>linear</metricType>
    </SKL>
    <SKL name="cosine">
        <metricType>cosine</metricType>
        <dense_output>True</dense_output>
    </SKL>
    <SKL name="cityblock">
        <metricType>cityblock</metricType>
    </SKL>
    <SKL name="l1">
```

```
        <metricType>l1</metricType>
    </SKL>
    <SKL name="l2">
        <metricType>l2</metricType>
    </SKL>
    <SKL name="manhattan">
        <metricType>manhattan</metricType>
        <sum_over_features>True</sum_over_features>
        <size_threshold>5e8</size_threshold>
    </SKL>
    <SKL name="additive_chi2">
        <metricType>additive_chi2</metricType>
    </SKL>
    <SKL name="chi2">
        <metricType>chi2</metricType>
        <gamma>1.0</gamma>
    </SKL>
    <SKL name="explained_variance_score">
        <metricType>explained_variance_score</metricType>
        <sample_weight>[0.1,0.1,0.1,0.05,0.05,0.2,0.1,0.1,0.1,0.1]</sample_w
    </SKL>
    <SKL name="mean_absolute_error">
        <metricType>mean_absolute_error</metricType>
        <sample_weight>[0.1,0.1,0.1,0.05,0.05,0.2,0.1,0.1,0.1,0.1]</sample_w
    </SKL>
    <SKL name="r2_score">
        <metricType>r2_score</metricType>
        <sample_weight>[0.1,0.1,0.1,0.05,0.05,0.2,0.1,0.1,0.1,0.1]</sample_w
    </SKL>
    <SKL name="mean_squared_error">
        <metricType>mean_squared_error</metricType>
        <sample_weight>[0.1,0.1,0.1,0.05,0.05,0.2,0.1,0.1,0.1,0.1]</sample_w
    </SKL>
    <SKL name="median_absolute_error">
        <metricType>median_absolute_error</metricType>
    </SKL>
  </Metrics>
  ...
</Simulation>
```

19.4 CDFAreaDifference

This calculates the difference in area between the two CDFs. This metric supports using distributions as input. Other inputs are converted to a CDF.

$$\text{CDF area difference} = \int_{-\infty}^{\infty} \|CDF_a(x) - CDF_b(x)\| dx \tag{36}$$

This metric has the same units as x. The closer the number is to zero, the closer the match. A perfect match would be 0.0.

19.5 PDFCommonArea

This calculates the common area between the two PDFs. The higher the value the closer the PDFs are. This metric supports distributions as inputs. Other inputs are converted to a PDF.

$$\text{PDF common area} = \int_{-\infty}^{\infty} \min(PDF_a(x), PDF_b(x)) dx \tag{37}$$

A perfect match would be 1.0.

20 Steps

The core of the RAVEN calculation flow is the **Step** system. The **Step** is in charge of assembling different entities in RAVEN (e.g. Samplers, Models, Databases, etc.) in order to perform a task defined by the kind of step being used. A sequence of different **Steps** represents the calculation flow.

Before analyzing each **Step** type, it is worth to 1) explain how a general **Step** entity is organized, and 2) introduce the concept of step "role" . In the following example, a general example of a **Step** is shown below:

```
<Simulation>
  ...
  <Steps>
    ...
    <WhatEverStepType name='aName'>
        <Role1 class='aMainClassType'
           type='aSubType'>userDefinedName1</Role1>
        <Role2 class='aMainClassType'
           type='aSubType'>userDefinedName2</Role2>
        <Role3 class='aMainClassType'
           type='aSubType'>userDefinedName3</Role3>
        <Role4 class='aMainClassType'
           type='aSubType'>userDefinedName4</Role4>
    </WhatEverStepType>
    ...
  </Steps>
  ...
</Simulation>
```

As shown above each **Step** consists of a list of entities organized into "Roles." Each role represents a behavior the entity (object) will assume during the evaluation of the **Step**. In RAVEN, several different roles are available:

- **Input** represents the input of the **Step**. The allowable input objects depend on the type of **Model** in the **Step**.
- **Output** defines where to collect the results of an action performed by the **Model**. It is generally one of the following types: **DataObjects**, **Databases**, or **OutStreams**.
- **Model** represents a physical or mathematical system or behavior. The object used in this role defines the allowable types of **Inputs** and **Outputs** usable in this step.

- **Sampler** defines the sampling strategy to be used to probe the model.
  It is worth to mention that, when a sampling strategy is employed, the "variables" defined in the `<variable>` blocks are going to be directly placed in the **Output** objects of type **DataObjects** and **Databases**).

- **Function** is an extremely important role. It introduces the capability to perform pre or post processing of Model **Inputs** and **Outputs**. Its specific behavior depends on the **Step** is using it.

- **ROM** defines an acceleration Reduced Order Model to use for a **Step**.

- **SolutionExport** represents the container of the eventual output of a step. For the moment, there are two uses: 1) A **Step** is employing the search of the Limit Surface (LS), through the class of Adaptive **Samplers**); in this case, it contains the coordinates of the LS in the input space; and 2) Some of the post-processors employ clustering algorithms and the cluster centers will be stored in this file with the input being the cluster labels.

Depending on the **Step** type, different combinations of these roles can be used. For this reason, it is important to analyze each **Step** type in details.

The available steps are the following

- SingleRun (see Section 20.1)
- MultiRun(see Section 20.2)
- IOStep(see Section 20.3)
- RomTrainer(see Section 20.4)
- PostProcess(see Section 20.5)

20.1 SingleRun

The **SingleRun** is the simplest step the user can use to assemble a calculation flow: perform a single action of a **Model**. For example, it can be used to run a single job (Code Model) and collect the outcome(s) in a "**DataObjects**" object of type **Point** or **History** (see Section 14 for more details on available data representations).

The specifications of this Step must be defined within a `<SingleRun>` XML block. This XML node has the following definable attributes:

- `name`, ***required string attribute***, user-defined name of this **Step**. **Note:** This name is used to reference this specific entity in the `<RunInfo>` block, under the `<Sequence>` node. If

the name of this **Step** is not listed in the `<Sequence>` block, its action is not going to be performed.

- `repeatFailureRuns`, ***optional integer attribute***, this optional attribute could be used to set a certain number of repetitions that need to be performed when a realization (i.e. run) fails (e.g. `repeatFailureRuns` = "3", 3 tries).
- `pauseAtEnd`, ***optional boolean/string attribute (case insensitive)***, if True (True values = True, yes, y, t), the code will pause at the end of the step, waiting for a user signal to continue. This is used in case one or more of the **Outputs** are of type **OutStreams**. For example, it can be used when an **OutStreams** of type **Plot** is output to the screen. Thus, allowing the user to interact with the **Plot** (e.g. rotate the figure, change the scale, etc.).
  Default: False.

In the `<SingleRun>` input block, the user needs to specify the objects needed for the different allowable roles. This step accepts the following roles:

- `<Input>`, ***string, required parameter***, names an entity (defined elsewhere in the RAVEN input) that will be used as input for the model specified in this step. This XML node accepts the following attributes:
  - `class`, ***required string attribute***, main object class type. This string corresponds to the tag of the main object's type used in the input. For example, `'Files'`, `'DataObjects'`, `'Databases'`, etc.
  - `type`, ***required string attribute***, the actual entity type. This attribute needs to specify the object type within the main object class. For example, if the `class` attribute is `'DataObjects'`, the `type` attribute might be `'PointSet'`. **Note:** The class `'Files'` has no type (i.e. `type=''`).

  Note: The `class` and, consequently, the `type` usable for this role depends on the particular `<Model>` being used. In addition, the user can specify as many `<Input>` nodes as needed.
- `<Model>`, ***string, required parameter***, names an entity defined elsewhere in the input file to be used as a model for this step. This XML node accepts the following attributes:
  - `class`, ***required string attribute***, main object class type. For this role, only `'Models'` can be used.
  - `type`, ***required string attribute***, the actual entity type. This attribute needs to specify the object type within the `Models` object class. For example, the `type` attribute might be `'Code'`, `'ROM'`, etc.
- `<Output>`, ***string, required parameter*** names an entity defined elsewhere in the input to use as the output for the **Model**. This XML node recognizes the following attributes:

- `class`, ***required string attribute***, main object class type. For this role, only `'DataObjects'`, `'Databases'`, and `'OutStreams'` can be used.
- `type`, ***required string attribute***, the actual entity type. This attribute needs to specify the object type within the main object class. For example, if the `class` attribute is `'DataObjects'`, the `type` attribute might be `'PointSet'`.

Note: The number of <Output> nodes is unlimited.

Example:

```
<Steps>
  ...
  <SingleRun name='StepName' pauseAtEnd='false'>
        <Input    class='Files'
           type=''>anInputFile.i</Input>
        <Input    class='Files'          type=''>aFile</Input>
        <Model   class='Models'        type='Code'>aCode</Model>
        <Output  class='Databases' type='HDF5'>aDatabase</Output>
        <Output  class='DataObjects'
           type='History'>aData</Output>
  </SingleRun>
  ...
</Steps>
```

20.2 MultiRun

The **MultiRun** step allows the user to assemble the calculation flow of an analysis that requires multiple "runs" of the same model. This step is used, for example, when the input (space) of the model needs to be perturbed by a particular sampling strategy.

The specifications of this type of step must be defined within a <MultiRun> XML block. This XML node recognizes the following list of attributes:

- `name`, ***required string attribute***, user-defined name of this Step. **Note:** As with other objects, this name is used to reference this specific entity in the <RunInfo> block, under the <Sequence> node. If the name of this **Step** is not listed in the <Sequence> block, its action is not going to be performed.
- `re-seeding`, ***optional integer/string attribute***, this optional attribute could be used to control the seeding of the random number generator (RNG). If inputted, the RNG can be reseeded. The value of this attribute can be: either 1) an integer value with the seed to be used (e.g. `re-seeding` = "20021986"), or 2) string value named "continue" where the RNG is not re-initialized

- **repeatFailureRuns**, ***optional integer attribute***, this optional attribute could be used to set a certain number of repetitions that need to be performed when a realization (i.e. run) fails (e.g. **repeatFailureRuns** = “3”, 3 tries).
- **pauseAtEnd**, ***optional boolean/string attribute***, if True (True values = True, yes, y, t), the code will pause at the end of the step, waiting for a user signal to continue. This is used in case one or more of the **Outputs** are of type **OutStreams**. For example, it can be used when an **OutStreams** of type **Plot** is output to the screen. Thus, allowing the user to interact with the **Plot** (e.g. rotate the figure, change the scale, etc.).
- **sleepTime**, ***optional float attribute***, in this attribute the user can specify the waiting time (seconds) between two subsequent inquiries of the status of the submitted job (i.e. check if a run has finished).
  Default: 0.05.

In the **<MultiRun>** input block, the user needs to specify the objects that need to be used for the different allowable roles. This step accepts the following roles:

- **<Input>**, ***string, required parameter***, names an entity to be used as input for the model specified in this step. This XML node accepts the following attributes:
  - **class**, ***required string attribute***, main object class type. This string corresponds to the tag of the main object’s type used in the input. For example, **'Files'**, **'DataObjects'**, **'Databases'**, etc.
  - **type**, ***required string attribute***, the actual entity type. This attribute specifies the object type within the main object class. For example, if the **class** attribute is **'DataObjects'**, the **type** attribute might be **'PointSet'**. **Note:** The class **'Files'** has no type (i.e. **type=''**).

  Note: The **class** and, consequently, the **type** usable for this role depend on the particular **<Model>** being used. The user can specify as many **<Input>** nodes as needed.
- **<Model>**, ***string, required parameter*** names an entity defined elsewhere in the input that will be used as the model for this step. This XML node recognizes the following attributes:
  - **class**, ***required string attribute***, main object class type. For this role, only **'Models'** can be used.
  - **type**, ***required string attribute***, the actual entity type. This attribute needs to specify the object type within the `Models` object class. For example, the **type** attribute might be **'Code'**, **'ROM'**, etc.
- **<Sampler>**, ***string, optional parameter*** names an entity defined elsewhere in the input file to be used as a sampler. As mentioned in Section 12, the **Sampler** is in charge of defining the strategy to characterize the input space. This XML node recognizes the following attributes:
  - **class**, ***required string attribute***, main object class type. This string corresponds to the tag of the main object’s type used. Only **'Samplers'** can be used for this role.

  - `type`, ***required string attribute***, the actual entity type. This attribute needs to specify the object type within the `Samplers` object class. For example, the `type` attribute might be **'MonteCarlo'**, **'Adaptive'**, **'AdaptiveDET'**, etc. See Section 12 for all the different types currently supported.

- **<Optimizer>**, ***string, optional parameter*** names an entity defined elsewhere in the input file to be used as an optimizer. As mentioned in Section 13, the **Optimizer** is in charge of defining the strategy to optimize an user-specified variable. This XML node recognizes the following attributes:

  - `class`, ***required string attribute***, main object class type. This string corresponds to the tag of the main object's type used. Only **'Optimizers'** can be used for this role.
  - `type`, ***required string attribute***, the actual entity type. This attribute needs to specify the object type within the `Optimizers` object class. For example, the `type` attribute might be **'SPSA'**, etc. See Section 13 for all the different types currently supported.

  Note: For Multi-Run, either one **<Sampler>** or one **<Optimizer>** is required.

- **<SolutionExport>**, ***string, optional parameter*** identifies an entity to be used for exporting key information coming from the **Sampler** or **Optimizer** object during the simulation. This XML node accepts the following attributes:

  - `class`, ***required string attribute***, main object class type. This string corresponds to the tag of the main object's type used in the input. For this role, only **'DataObjects'** can be used.
  - `type`, ***required string attribute***, the actual entity type. This attribute needs to specify the object type within the `DataObjects` object class. For example, the `type` attribute might be **'PointSet'**, **'HistorySet'**, etc.
    Note: Whether or not it is possible to export the **Sampler** solution depends on the `type`. Currently, only the Samplers in the **'Adaptive'** category and all Optimizers can export their solution into a **<SolutionExport>** entity. For Samplers, the **<Outputs>** node in the `DataObjects` needs to contain the goal **<Function>** name. For example, if **<Sampler>** is of type **'Adaptive'**, the **<SolutionExport>** needs to be of type **'PointSet'** and it will contain the coordinates, in the input space, that belong to the "Limit Surface". For Optimizers, the **<SolutionExport>** needs to be of type **'HistorySet'** and it will contains all the optimization trajectories, each as a history, that record how the variables are updated along each optimization trajectory.

- **<Output>**, ***string, required parameter*** identifies an entity to be used as output for this step. This XML node recognizes the following attributes:

  - `class`, ***required string attribute***, main object class type. This string corresponds to the tag of the main object's type used in the input. For this role, only **'DataObjects'**, **'Databases'**, and **'OutStreams'** may be used.

– `type`, ***required string attribute***, the actual entity type. This attribute specifies the object type within the main object class. For example, if the `class` attribute is `'DataObjects'`, the `type` attribute might be `'PointSet'`.

Note: The number of `<Output>` nodes is unlimited.

Example:

```
<Steps>
  ...
  <MultiRun name='StepName1' pauseAtEnd='False' sleepTime='0.01'>
    <Input class='Files' type=''>anInputFile.i</Input>
    <Input class='Files' type=''>aFile</Input>
    <Sampler class='Samplers' type = 'Grid'>aGridName</Sampler>
    <Model class='Models' type='Code'>aCode</Model>
    <Output class='Databases' type='HDF5'>aDatabase</Output>
    <Output class='DataObjects' type='History'>aData</Output>
  </MultiRun >
  <MultiRun name='StepName2' pauseAtEnd='True' sleepTime='0.02'>
    <Input   class='Files' type=''>anInputFile.i</Input>
    <Input   class='Files' type=''>aFile</Input>
    <Sampler class='Samplers' type='Adaptive'>anAS</Sampler>
    <Model   class='Models' type='Code'>aCode</Model>
    <Output  class='Databases' type='HDF5'>aDatabase</Output>
    <Output  class='DataObjects' type='History'>aData</Output>
    <SolutionExport class='DataObjects' type='PointSet'>
      aTPS
    </SolutionExport>
  </MultiRun>
  ...
</Steps>
```

20.3 IOStep

As the name suggests, the **IOStep** is the step where the user can perform input/output operations among the different I/O entities available in RAVEN. This step type is used to:

- construct/update a *Database* from a *DataObjects* object, and vice versa;
- construct/update a *DataObject* from a *CSV* file contained in a directory;
- construct/update a *Database* or a *DataObjects* object from *CSV* files contained in a directory;

- stream the content of a *Database* or a *DataObjects* out through an **OutStream** object (see section 16);
- store/retrieve a *ROM* to/from an external *File* using Pickle module of Python.

This last function can be used to create and store mathematical model of fast solution trained to predict a response of interest of a physical system. This model can be recovered in other simulations or used to evaluate the response of a physical system in a Python program by the implementing of the Pickle module. The specifications of this type of step must be defined within an **<IOStep>** XML block. This XML node can accept the following attributes:

- **name**, ***required string attribute***, user-defined name of this Step. **Note:** As for the other objects, this is the name that can be used to refer to this specific entity in the **<RunInfo>** block, under the **<Sequence>** node.
- **pauseAtEnd**, ***optional boolean/string attribute (case insensitive)***, if True (True values = True, yes, y, t), the code will pause at the end of the step, waiting for a user signal to continue. This is used in case one or more of the **Outputs** are of type **OutStreams**. For example, it can be used when an **OutStreams** of type **Plot** is output to the screen. Thus, allowing the user to interact with the **Plot** (e.g. rotate the figure, change the scale, etc.).
Default: False.
- **fromDirectory**, ***optional string attribute***, The directory where the input files can be found when loading data from a file or series of files directly into a DataObject.

In the **<IOStep>** input block, the user specifies the objects that need to be used for the different allowable roles. This step accepts the following roles:

- **<Input>**, ***string, required parameter***, names an entity that is going to be used as a source (input) from which the information needs to be extracted. This XML node recognizes the following attributes:
  - **class**, ***required string attribute***, main object class type. This string corresponds to the tag of the main object's type used in the input. As already mentioned, the allowable main classes are **'DataObjects'**, **'Databases'**, **'Models'** and **'Files'**.
  - **type**, ***required string attribute***, the actual entity type. This attribute needs to specify the object type within the main object class. For example, if the **class** attribute is **'DataObjects'**, the **type** attribute might be **'PointSet'**. If the **class** attribute is **'Models'**, the **type** attribute must be **'ROM'** and if the **class** attribute is **'Files'**, the **type** attribute must be **' '**
- **<Output>**, ***string, required parameter*** names an entity to be used as the target (output) where the information extracted in the input will be stored. This XML node needs to contain the following attributes:

- class, ***required string attribute***, main object class type. This string corresponds to the tag of the main object's type used in the input. The allowable main classes are **'DataObjects'**, **'Databases'**, **'OutStreams'**, **'Models'** and **'Files'**.
- type, ***required string attribute***, the actual entity type. This attribute specifies the object type within the main object class. For example, if the class attribute is **'OutStreams'**, the type attribute might be **'Plot'**.

This step acts as a "transfer network" among the different RAVEN storing (or streaming) objects. The number of **<Input>** and **<Output>** nodes is unlimited, but should match. This step assumes a 1-to-1 mapping (e.g. first **<Input>** is going to be used for the first **<Output>**, etc.).
Note: This 1-to-1 mapping is not present when **<Output>** nodes are of class **'OutStreams'**, since **OutStreams** objects are already linked to a Data object in the relative RAVEN input block.

In this case, the user needs to provide all of the "DataObjects" objects linked to the OutStreams objects (see the example below) in the **<Input>** nodes:

```
<Steps>
  ...
  <IOStep name='OutStreamStep'>
    <Input class='DataObjects'
       type='HistorySet'>aHistorySet</Input>
    <Input class='DataObjects' type='PointSet'>aTPS</Input>
    <Output class='OutStreams' type='Plot'>plot_hist
    </Output>
    <Output class='OutStreams' type='Print'>print_hist
    </Output>
    <Output class='OutStreams' type='Print'>print_tps
    </Output>
    <Output class='OutStreams' type='Print'>print_tp
    </Output>
  </IOStep>
  <IOStep name='PushDataObjectsIntoDatabase'>
    <Input  class='DataObjects'
       type='HistorySet'>aHistorySet</Input>
    <Input  class='DataObjects'     type='PointSet'>aTPS</Input>
    <Output class='Databases' type='HDF5'>aDatabase</Output>
    <Output class='Databases' type='HDF5'>aDatabase</Output>
  </IOStep>
  <IOStep name='ConstructDataObjectsFromCSV'>
    <Input class='Files' type=''>aCSVFile</Input>
    <Output class='DataObjects' type='PointSet'>aPS</Output>
  </IOStep>
```

```xml
  <IOStep name='ConstructDataObjectsFromDatabase'>
    <Input class='Databases' type='HDF5'>aDatabase</Input>
    <Input class='Databases' type='HDF5'>aDatabase</Input>
    <Output class='DataObjects'
       type='HistorySet'>aHistorySet</Output>
    <Output class='DataObjects'    type='PointSet'>aTPS</Output>
  </IOStep>
  <IOStep name='PushROMIntoFile'>
    <Input class='Models'  type='ROM'>aROM</Input>
    <Output class='Files'  type=''>aFile</Output>
  </IOStep>
  <IOStep name='ImportROMFromFile'>
    <Input class='Files'  type=''>zFile</Input>
    <Output class='Models'  type='ROM'>zROM</Output>
  </IOStep>
  ...
</Steps>
```

As already mentioned, the <IOStep> can be used to export (serialize) a ROM in a binary file. To use the exported ROM in an external Python (or Python-compatible) code, the RAVEN framework must be present in end-user machine. The main reason for this is that the *Pickle* module uses the class definitions to template the reconstruction of the serialized object in memory.
In order to faciliate the usage of the serialized ROM in an external Python code, the RAVEN team provided a utility class contained in :

```
./raven/scripts/externalROMloader.py
```

An example of how to use this utility class to load and use a serialized ROM (already trained) is reported below: Example Python Function:

```python
from externalROMloader import ravenROMexternal
import numpy as np
rom = ravenROMexternal("path_to_pickled_rom/ROM.pk",
                                     "path_to_RAVEN_framework")
request = {"x1":np.atleast_1d(Value1),"x2":np.atleast_1d(Value2)}
eval  = rom.evaluate(request)
print str(eval)
```

The module above can also be used to evaluate a ROM from input file:

```
python ./raven/scripts/externalROMloader.py input_file.xml
```

The input file has the following format:

```xml
<?xml version="1.0" ?>
<external_rom>
  <RAVENdir>path_to_RAVEN_framework</RAVENdir>
  <ROMfile>ath_to_pickled_rom/ROM.pk</ROMfile>
  <evaluate>
    <x1>0. 1. 0.5</x1>
    <x2>0. 0.4 2.1</x2>
  </evaluate>
  <inspect>true</inspect>
  <outputFile>output_file_name</outputFile>
</external_rom>
```

The output of the above command would look like as follows:

```xml
<?xml version="1.0" ?>
<UROM>
  <settings>
    <Target>ans</Target>
    <name>UROM</name>
    <IndexSet>TensorProduct</IndexSet>
    <Features>[u'x1' u'x2']</Features>
    <PolynomialOrder>2</PolynomialOrder>
  </settings>
  <evaluations>
    <evaluation realization="1">
      <x2>0.0</x2>
      <x1>0.0</x1>
      <ans>-3.1696867353e-14</ans>
    </evaluation>
    <evaluation realization="2">
      <x2>0.4</x2>
      <x1>1.0</x1>
      <ans>1.4</ans>
    </evaluation>
    <evaluation realization="3">
      <x2>2.1</x2>
      <x1>0.5</x1>
      <ans>2.6</ans>
    </evaluation>
  </evaluations>
</UROM>
```

20.4 RomTrainer

The **RomTrainer** step type performs the training of a Reduced Order Model. The specifications of this step must be defined within a **<RomTrainer>** block. This XML node accepts the attributes:

- **name**, ***required string attribute***, user-defined name of this step. **Note:** As for the other objects, this is the name that can be used to refer to this specific entity in the **<RunInfo>** block under **<Sequence>**.

In the **<RomTrainer>** input block, the user will specify the objects needed for the different allowable roles. This step accepts the following roles:

- **<Input>**, ***string, required parameter*** names an entity to be used as a source (input) from which the information needs to be extracted. This XML node accepts the following attributes:
  - **class**, ***required string attribute***, main object class type. This string corresponds to the tag of the main object's type used in the input. The only allowable main class is **'DataObjects'**.
  - **type**, ***required string attribute***, the actual entity type. This attribute specifies the object type within the main object class. For example, the **type** attribute might be **'PointSet'**. **Note:** Since the ROMs currently present in RAVEN are not time-dependent, the only allowable types are **'Point'** and **'PointSet'**.
- **<Output>**, ***string, required parameter***, names a ROM entity that is going to be trained. This XML node recognizes the following attributes:
  - **class**, ***required string attribute***, main object class type. This string corresponds to the tag of the main objects type used in the input. The only allowable main class is **'Models'**.
  - **type**, ***required string attribute***, the actual entity type. This attribute needs to specify the object type within the main object class. The only type accepted here is, currently, **'ROM'**.

Example:

```
<Steps>
  ...
  <RomTrainer name='aStepName'>
    <Input  class='DataObjects'  type='PointSet'>aTPS</Input>
    <Output class='Models' type='ROM'          >aROM</Output>
  </RomTrainer>
  ...
</Steps>
```

20.5 PostProcess

The **PostProcess** step is used to post-process data or manipulate RAVEN entities. It is aimed at performing a single action that is employed by a **Model** of type **PostProcessor**.

The specifications of this type of step is defined within a **<PostProcess>** XML block. This XML node specifies the following attributes:

- **name**, ***required string attribute***, user-defined name of this Step. **Note:** As for the other objects, this is the name that is used to refer to this specific entity in the **<RunInfo>** block under the **<Sequence>** node.
- **pauseAtEnd**, ***optional boolean/string attribute (case insensitive)***, if True (True values = True, yes, y, t), the code will pause at the end of the step, waiting for a user signal to continue. This is used in case one or more of the **Outputs** are of type **OutStreams**. For example, it can be used when an **OutStreams** of type **Plot** is output to the screen. Thus, allowing the user to interact with the **Plot** (e.g. rotate the figure, change the scale, etc.).
  Default: False.

In the **<PostProcess>** input block, the user needs to specify the objects needed for the different allowable roles. This step accepts the following roles:

- **<Input>**, ***string, required parameter***, names an entity to be used as input for the model specified in this step. This XML node accepts the following attributes:
  - **class**, ***required string attribute***, main object class type. This string corresponds to the tag of the main object's type used in the input. For example, **'Files'**, **'DataObjects'**, **'Databases'**, etc.
  - **type**, ***required string attribute***, the actual entity type. This attribute specifies the object type within the main object class. For example, if the **class** attribute is **'DataObjects'**, the **type** attribute might be **'PointSet'**. **Note:** The class **'Files'** has no type (i.e. **type=''**).

  Note: The **class** and, consequently, the **type** usable for this role depends on the particular type of **PostProcessor** being used. In addition, the user can specify as many **<Input>** nodes as needed by the model.

- **<Model>**, ***string, required parameter***, names an entity to be used as a model for this step. This XML node recognizes the following attributes:
  - **class**, ***required string attribute***, main object class type. This string corresponds to the tag of the main object's type used in the input. For this role, only **'Models'** can be used.
  - **type**, ***required string attribute***, the actual entity type. This attribute needs to specify the object type within the **'Models'** object class. The only type accepted here is **'PostProcessor'**.

- **<Output>**, ***string, required/optional parameter***, names an entity to be used as output for the PostProcessor. The necessity of this XML block and the types of entities that can be used as output depend on the type of **PostProcessor** that has been used as a **Model** (see section 17.5). This XML node specifies the following attributes:
  - **class**, ***required string attribute***, main object class type. This string corresponds to the tag of the main object's type used in the input.
  - **type**, ***required string attribute***, the actual entity type. This attribute specifies the object type within the main object class. For example, if the **class** attribute is **'DataObjects'**, the **type** attribute might be **'PointSet'**.

  Note: The number of **<Output>** nodes is unlimited.

Example:

```
<Steps>
  ...
  <PostProcess name='PP1'>
      <Input  class='DataObjects'  type='PointSet' >aData</Input>
      <Model  class='Models' type='PostProcessor'>aPP</Model>
      <Output class='Files'  type=''>anOutputFile</Output>
  </PostProcess>
  ...
</Steps>
```

21 Existing Interfaces

21.1 Generic Interface

The GenericCode interface is meant to handle a wide variety of generic codes that take take straightforward input files and produce output CSV files. There are some limitations for this interface. If a code:

- accepts a keyword-based input file with no cross-dependent inputs,
- has no more than one filetype extension per command line flag,
- and returns a CSV with the input parameters and output parameters,

the GenericCode interface should cover the code for RAVEN.

If a code contains cross-dependent data, the generic interface is not able to edit the correct values. For example, if a geometry-building script specifies inner_radius, outer_radius, and thickness, the generic interface cannot calculate the thickness given the outer and inner radius, or vice versa.

An example of the code interface is shown here. The input parameters are read from the input files gen.one and gen.two respectively. The code is run using python, so that is part of the **<prepend>** node. The command line entry to normally run the code is

```
python poly_inp.py -i gen.one -a gen.two -o myOut
```

and produces the output myOut.csv.

Example:

```xml
<Code name="poly" subType="GenericCode">
  <executable>GenericInterface/poly_inp.py</executable>
  <inputExtentions>.one,.two</inputExtentions>
  <clargs type='prepend' arg='python'/>
  <clargs type='input'   arg='-i' extension='.one'/>
  <clargs type='input'   arg='-a' extension='.two'/>
  <clargs type='output'  arg='-o'/>
  <prepend>python</prepend>
</Code>
```

If a code doesn't accept necessary Raven-editable auxiliary input files or output filenames through the command line, the GenericCode interface can also edit the input files and insert the filenames there. For example, in the previous example, say instead of -a gen.two and -o myOut in the command line, gen.one has the following lines:

```
...
auxfile = gen.two
case = myOut
...
```

Then, our example XML for the code would be

Example:

```
  <Code name="poly" subType="GenericCode">
    <executable>GenericInterface/poly_inp.py</executable>
    <inputExtentions>.one,.two</inputExtentions>
    <clargs   type='prepend' arg='python'/>
    <clargs   type='input'   arg='-i'  extension='.one'/>
    <fileargs type='input'   arg='two' extension='.two'/>
    <fileargs type='output'  arg='out'/>
    <prepend>python</prepend>
  </Code>
```

and the corresponding template input file lines would be changed to read

```
...
auxfile = $RAVEN-two$
case = $RAVEN-out$
...
```

If a code has hard-coded output file names that are not changeable, the GenericCode interface can be invoked using the **<outputFile>** node in which the output file name (CSV only) must be specified. For example, in the previous example, say instead of `-a gen.two` and `-o myOut` in the command line, the code always produce a CSV file named “fixed@output.csv”;

Then, our example XML for the code would be

Example:

```
  <Code name="poly" subType="GenericCode">
    <executable>GenericInterface/poly_inp.py</executable>
    <inputExtentions>.one,.two</inputExtentions>
    <clargs   type='prepend' arg='python'/>
    <clargs   type='input'   arg='-i'  extension='.one'/>
    <fileargs type='input'   arg='two' extension='.two'/>
    <outputFile>fixed_output.csv</outputFile>
    <prepend>python</prepend>
```

```
</Code>
```

In addition, the “wild-cards” above can contain two special and optional symbols:

- **:**, that defines an eventual default value;
- **|**, that defines the format of the value. The Generic Interface currently supports the following formatting options (* in the examples means blank space):
  - **plain integer**, in this case the value that is going to be replaced by the Generic Interface, will be left-justified with a string length equal to the integer value specified here (e.g. “|6”, the value is left-justified with a string length of 6);
  - **d**, signed integer decimal, the value is going to be formatted as an integer (e.g. if the value is 9 and the format “|10d”, the replaced value will be formatted as follows: “*********9”);
  - **e**, floating point exponential format (lowercase), the value is going to be formatted as a float in scientific notation (e.g. if the value is 9.1234 and the format “|10.3e”, the replaced value will be formatted as follows: “*9.123e+00”);
  - **E**, floating point exponential format (uppercase), the value is going to be formatted as a float in scientific notation (e.g. if the value is 9.1234 and the format “|10.3E”, the replaced value will be formatted as follows: “*9.123E+00”);
  - **f or F**, floating point decimal format, the value is going to be formatted as a float in decimal notation (e.g. if the value is 9.1234 and the format “|10.3f”, the replaced value will be formatted as follows: “*****9.123”);
  - **g**, floating point format. Uses lowercase exponential format if exponent is less than -4 or not less than precision, decimal format otherwise (e.g. if the value is 9.1234 and the format “|10.3g”, the replaced value will be formatted as follows: “******9.12”);
  - **G**, floating point format. Uses uppercase exponential format if exponent is less than -4 or not less than precision, decimal format otherwise (e.g. if the value is 0.000009 and the format “|10.3G”, the replaced value will be formatted as follows: “*****9E-06”).

—

For example:

```
...
auxfile = $RAVEN-two:3$
case = $RAVEN-out:5|10$
...
```

Where,

- `:`, in case the variable "two" is not defined in the RAVEN XML input file, the Parser, will replace it with the value "3".;
- `|`, the value that is going to be replaced by the Generic Interface, will be left- justified with a string length of "10";

21.2 RAVEN Interface

The RAVEN interface is meant to provide the possibility to execute a RAVEN input file driving a set of SLAVE RAVEN calculations. For example, if the user wants to optimize the parameters of a surrogate model (e.g. minimizing the distance between the surrogate predictions and the real data), he can achieve this task by setting up a RAVEN input file (master) that performs an optimization on the feature space characterized by the surrogate model parameters, whose training and validation assessment is performed in the SLAVE RAVEN runs.
There are some limitations for this interface:

- only one sub-level of RAVEN can be executed (i.e. if the SLAVE RAVEN input file contains the run of another RAVEN SLAVE, the MASTER RAVEN will error out)
- only data from Outstreams of type Print can be collected by the MASTER RAVEN
- only a maximum of two Outstreams can be collected (1 PointSet and 1 HistorySet)

Like for every other interface, most of the RAVEN workflow stays the same independently of which type of Model (i.e. Code) is used.
Similarly to any other code interface, the user provides paths to executables and aliases for sampled variables within the **<Models>** block. The **<Code>** block will contain attributes **name** and **subType**. **name** identifies that particular **<Code>** model within RAVEN, and **subType** specifies which code interface the model will use (In this case **subType**="RAVEN"). The **<executable>** block should contain the absolute or relative (with respect to the current working directory) path to the RAVEN framework script (**raven_framework**).
In addition to the attributes and xml nodes reported above, the RAVEN accepts the following XML nodes (required and optional):

- **<outputExportOutStreams>**, ***comma separated list, required parameter*** will specify the **<OutStreams>** that will be loaded as outputs of the SLAVE RAVEN. Maximum two **<OutStreams>** can be listed here (1 for PointSet and/or 1 for HistorySet).
- **<conversionModule>**, ***string, optional parameter*** will specify the path to a *Python* module that can contain two methods:

– ***manipulateScalarSampledVariables***, a method that is aimed to manipulate sampled variables and to create more in case needed. Example:

```
def manipulateScalarSampledVariables(sampledVariables):
  """
  This method is aimed to manipulate scalar variables.
  The user can create new variables based on the
  variables sampled by RAVEN
   @ In, sampledVariables, dict, dictionary of
       sampled variables ({"var1":value1,"var2":value2})
   @ Out, None, the new variables should be
                 added in the "sampledVariables" dictionary
  """
  newVariableValue =
    sampledVariables['Distributions|Uniform@name:a_dist|lowerBound']
    + 1.0
  sampledVariables['Distributions|Uniform@name:a_dist|upperBound'] =
    newVariableValue
  return
```

– ***convertNotScalarSampledVariables***, a method that is aimed to convert not scalar variables (e.g. 1D arrays) into multiple scalar variables (e.g. **<constant>**(s) in a sampling strategy). This method is going to be required in case not scalar variables are detected by the interface. Example:

```
 def convertNotScalarSampledVariables(noScalarVariables):
  """
  This method is aimed to convert not scalar
  variables into multiple scalar variables. The user MUST
   create new variables based on the not Scalar Variables
    sampled (and passed in) by RAVEN
  @ In, noScalarVariables, dict, dictionary of sampled
       variables that are not scalar ({"var1":1Darray1,"var2":1Darra
  @ Out, newVars, dict,  the new variables that have
       been created based on the not scalar variables
       contained in "noScalarVariables" dictionary
  """
  oneDimensionalArray =
      noScalarVariables['temperatureHistory']
  newVars = {}
  for cnt, value in enumerate(oneDimensionalArray):
    newVars['Samplers|MonteCarlo@name:myMC|constant'+
               '@name=temperatureHistory'+str(cnt)] =
               oneDimensionalArray[cnt]
```

```
return newVars
```

Code input example:

```
<Code name="RAVENrunningRAVEN" subType="RAVEN">
  <executable>../../../raven_framework</executable>
  <outputExportOutStreams>
     HistorySetOutStream,PointSetOutStream
  </outputExportOutStreams>
  <conversionModule>
    ~/Users/username/whateverConversionModule.py
  </conversionModule>
</Code>
```

Like for every other interface, the syntax of the variable names is important to make the parser understand how to perturb an input file.
For the RAVEN interface, a syntax inspired by the XPath nomenclature is used.

```
<Samplers>
  <MonteCarlo name="MC_external">
     ...
    <variable
       name="Models|ROM@subType:SciKitLearn@name:ROM1|C">
      <distribution>C_distrib</distribution>
    </variable>
    <variable
       name="Models|ROM@subType:SciKitLearn@name:ROM1|tol">
      <distribution>toll_distrib</distribution>
    </variable>
    <variable name="Samplers|Grid@name:'+
␣␣␣␣␣␣␣␣␣␣␣␣
   'GridName|variable@name:var1|grid@construction:equal@type:value@steps">
      <distribution>categorical_step_distrib</distribution>
    </variable>
    ...
  </MonteCarlo>
</Samplers>
```

In the above example, it can be inferred that each XML node (subnode) needs to be separated by a "—" separator. In addition, every time an XML node has attributes, the user can specify them using the "@" separator to specify a value for them. The first variable above will be pointing to the following XML sub-node (<C>):

```xml
      <Models>
        <ROM name="ROM1" subType="SciKitLearn">
           ...
           <C>10.0</C>
          ...
        </ROM>
      </Models>
```

The second variable above will be pointing to the following XML sub-node (<tol>):

```xml
      <Models>
        <ROM name="ROM1" subType="SciKitLearn">
           ...
           <tol>0.0001</tol>
          ...
        </ROM>
      </Models>
```

The third variable above will be pointing to the following XML attribute (steps):

```xml
  <Samplers>
    <Grid name="GridName">
       ...
      <variable name="var1">
         ...
        <grid construction="equal" type="value" steps="1">0
           1</grid>
         ...
      </variable>

      ...
    </MonteCarlo>
  </Samplers>
```

The above nomenclature must be used for all the variables to be sampled and for the variables generated by the two methods contained, in case, in the module that gets specified by the <conversionModule> in the <Code> section.
Finally the SLAVE RAVEN input file (s) must be "tagged" with the attribute type="raven" in the Files section. For example,

```xml
<Files>
    <Input name="slaveRavenInputFile" type="raven" >
      test_rom_trainer.xml
```

```
    </Input>
</Files>
```

21.2.1 ExternalXML and RAVEN interface

Care must be taken if the SLAVE RAVEN uses **<ExternalXML>** nodes. In this case, each file containing external XML nodes must be added in the **<Step>** as an **<Input>** class **Files** to make sure it gets copied to the individual run directory. The type for these files can be anything, with the exception of type **'raven'**.

21.3 RELAP5 Interface

21.3.1 Sequence

In the **<Sequence>** section, the names of the steps declared in the **<Steps>** block should be specified. As an example, if we called the first multirun "Grid_Sampler" and the second multirun "MC_Sampler" in the sequence section we should see this:

```
<Sequence>Grid_Sampler,MC_Sampler</Sequence>
```

21.3.2 batchSize and mode

For the **<batchSize>** and **<mode>** sections please refer to the **<RunInfo>** block in the previous chapters.

21.3.3 RunInfo

After all of these blocks are filled out, a standard example RunInfo block may look like the example below:

```
<RunInfo>
  <WorkingDir>~/workingDir</WorkingDir>
  <Sequence>Grid_Sampler,MC_Sampler</Sequence>
  <batchSize>1</batchSize>
  <mode>mpi</mode>
  <expectedTime>1:00:00</expectedTime>
  <ParallelProcNumb>1</ParallelProcNumb>
```

```
</RunInfo>
```

21.3.4 Files

In the <Files> section, as specified before, all of the files needed for the code to run should be specified. In the case of RELAP5, the files typically needed are:

- RELAP5 Input file
- Table file or files that RELAP needs to run

Example:

```xml
<Files>
  <Input name='tpfh2o' type=''>tpfh2o</Input>
  <Input name='inputrelap.i' type=''>X10.i</Input>
</Files>
```

It is a good practice to put inside the working directory all of these files and also:

- the RAVEN input file
- the license for the executable of RELAP5

It is important to notice that the interface output collection relies on the MINOR EDITS. The user must specify the MINOR EDITS block and those variables are the only one the INTERFACE will read and make available to RAVEN. In addition, it is important to notice that:

- **the simulation time is stored in a variable called *"time"***;
- **all the variables specified in the MINOR EDIT block are going to be converted using underscores (e.g. an edit such as** $301 \quad p \quad 345010000$ **will be named in the converted CSVs as** $p_345010000$**).In addition, if a variable contains spaces, the trailing spaces are going to be removed and internal spaces are replaced with underscores (e.g.** $HTTEMP113100812$ **will become** $HTTEMP_1131008_12$.

Remeber also that a RELAP5 simulation run is considered successful (i.e., the simulation did not crash) if it terminates with the following message: Transient terminated by end of time step cards or Transient terminated by trip

If the a RELAP5 simulation run stops with messages other than this one (e.g., " Transient terminated by failure.") than the simulation is considered as crashed, i.e., it will not be saved. Hence, it is strongly recommended to set up the RELAP5 input file so that the simulation exiting conditions are set through control logic trip variables (e.g., simulation mission time and clad temperature equal to clad failure temperature).

21.3.5 Models

For the **<Models>** block here is a standard example of how it would look when using RELAP5 as the external model:

```
<Models>
  <Code name='MyRELAP' subType='Relap5'>
    <executable>~/path_to_the_executable</executable>
  </Code>
</Models>
```

In case the **multi-deck** approach is used in RELAP5, the interface is going to load all the outputs in one CSV RAVEN is going to read. This means that all the decks' outputs are going to be loaded in one of the Output of RAVEN. In case the user wants to select the outputs coming from only one deck, the following XML node needs to be specified:

- **<outputDeckNumber>**, ***integer, optional parameter***, the deck number from which the results needs to be retrieved.
  Default: all.

In addition, if some command line parameters need to be passed to RELAP5 (e.g. "-r $restartFileWithCustomName.r$"), the user might use (optionally) the **<clargs>** XML nodes.

```
<Models>
  <Code name='MyRELAP' subType='Relap5'>
    <executable>~/path_to_the_executable</executable>
    <outputDeckNumber>1</outputDeckNumber>
    <clargs type="text" arg="-r restartFileWithCustomName.r"/>
  </Code>
</Models>
```

21.3.6 Distributions

The <Distribution> block defines the distributions that are going to be used for the sampling of the variables defined in the <Samplers> block. For all the possibile distributions and all their possible inputs please see the chapter about Distributions (see 11). Here we give a general example of three different distributions:

```xml
<Distributions verbosity='debug'>
  <Triangular name='BPfailtime'>
    <apex>5.0</apex>
    <min>4.0</min>
    <max>6.0</max>
  </Triangular>
  <LogNormal name='BPrepairtime'>
    <mean>0.75</mean>
    <sigma>0.25</sigma>
  </LogNormal>
  <Uniform name='ScalFactPower'>
    <lowerBound>1.0</lowerBound>
    <upperBound>1.2</upperBound>
  </Uniform>
 </Distributions>
```

It is good practice to name the distribution something similar to what kind of variable is going to be sampled, since there might be many variables with the same kind of distributions but different input parameters.

21.3.7 Samplers

In the <Samplers> block we want to define the variables that are going to be sampled. **Example**: We want to do the sampling of 3 variables:

- Battery Fail Time
- Battery Repair Time
- Scaling Factor Power Rate

We are going to sample these 3 variables using two different sampling methods: grid and MonteCarlo.

In RELAP5, the sampler reads the variable as, given the name, the first number is the card number and the second number is the word number. In this example we are sampling:

- For card 0000588 (trip) the word 6 (battery failure time)
- For card 0000575 (trip) the word 6 (battery repair time)
- For card 20210000 (reactor power) the word 4 (reactor scaling factor)

We proceed to do so for both the Grid sampling and the MonteCarlo sampling.

```xml
<Samplers verbosity='debug'>
  <Grid name='Grid_Sampler' >
    <variable name='0000588:6'>
      <distribution>BPfailtime</distribution>
      <grid type='value' construction='equal'  steps='10'>0.0
         28800</grid>
    </variable>
    <variable name='0000575:6'>
      <distribution>BPrepairtime</distribution>
      <grid type='value' construction='equal' steps='10'>0.0
         28800</grid>
    </variable>
    <variable name='20210000:4'>
      <distribution>ScalFactPower</distribution>
      <grid type='value' construction='equal' steps='10'>1.0
         1.2</grid>
    </variable>
  </Grid>
  <MonteCarlo name='MC_Sampler'>
     <samplerInit>
       <limit>1000</limit>
     </samplerInit>
    <variable name='0000588:6'>
      <distribution>BPfailtime</distribution>
    </variable>
    <variable name='0000575:6'>
      <distribution>BPrepairtime</distribution>
    </variable>
    <variable name='20210000:4'>
      <distribution>ScalFactPower</distribution>
    </variable>
  </MonteCarlo>
```

```
</Samplers>
```

In case the RELAP5 input file is a multi-deck, the user can specify the deck to which each sampled variable corresponds to. As an example, the following sampling strategy:

```
<MonteCarlo name='MC_Sampler'>
   <samplerInit>
     <limit>1000</limit>
   </samplerInit>
  <variable name='1|0000588:6'>
    <distribution>BPfailtime</distribution>
  </variable>
  <variable name='2|0000575:6'>
    <distribution>BPrepairtime</distribution>
  </variable>
</MonteCarlo>
</Samplers>
```

performs:

- the sampling of the distribution <BPfailtime> and it provides the sampled value to the 6th word of card 0000588 for the first deck
- the sampling of the distribution <BPrepairtime> and it provides the sampled value to the 6th word of card 0000575 for the second deck

It can be seen that each variable is connected with a proper distribution defined in the <Distributions> block (from the previous example). The following demonstrates how the input for the first variable is read.

We are sampling a a variable situated in word 6 of the card 0000588 using a Grid sampling method. The distribution that this variable is following is a Triangular distribution (see section above). We are sampling this variable beginning from 0.0 in 10 *equal* steps of 2880.

21.3.8 Steps

For a RELAP interface, the <MultiRun> step type will most likely be used. First, the step needs to be named: this name will be one of the names used in the <Sequence> block. In our example, `Grid_Sampler` and `MC_Sampler`.

```
    <MultiRun name='Grid_Sampler' verbosity='debug'>
```

With this step, we need to import all the files needed for the simulation:

- RELAP input file
- element tables – tpfh2o

```
    <Input    class='Files' type=''>inputrelap.i</Input>
    <Input    class='Files' type=''>tpfh2o</Input>
```

We then need to define which model will be used:

```
    <Model   class='Models' type='Code'>MyRELAP</Model>
```

We then need to specify which Sampler is used, and this can be done as follows:

```
    <Sampler class='Samplers' type='Grid'>Grid_Sampler</Sampler>
```

And lastly, we need to specify what kind of output the user wants. For example the user might want to make a database (in RAVEN the database created is an HDF5 file). Here is a classical example:

```
    <Output   class='Databases' type='HDF5'>Grid_out</Output>
```

Following is the example of two MultiRun steps which use different sampling methods (grid and Monte Carlo), and creating two different databases for each one:

```
<Steps verbosity='debug'>
  <MultiRun name='Grid_Sampler' verbosity='debug'>
    <Input    class='Files' type=''>inputrelap.i</Input>
    <Input    class='Files'     type=''    >tpfh2o</Input>
    <Model    class='Models'    type='Code'>MyRELAP</Model>
    <Sampler class='Samplers'  type='Grid'>Grid_Sampler</Sampler>
    <Output   class='Databases' type='HDF5'>Grid_out</Output>
  </MultiRun>
  <MultiRun name='MC_Sampler' verbosity='debug'
     re-seeding='210491'>
    <Input    class='Files' type=''>inputrelap.i</Input>
    <Input    class='Files'     type=''          >tpfh2o</Input>
    <Model    class='Models'    type='Code'      >MyRELAP</Model>
    <Sampler class='Samplers'
       type='MonteCarlo'>MC_Sampler</Sampler>
```

```
    <Output  class='Databases' type='HDF5'      >MC_out</Output>
  </MultiRun>
</Steps>
```

21.3.9 Databases

As shown in the <Steps> block, the code is creating two database objects called `Grid_out` and `MC_out`. So the user needs to input the following:

```
<Databases>
  <HDF5 name="Grid_out" readMode="overwrite"/>
  <HDF5 name="MC_out" readMode="overwrite"/>
</Databases>
```

As listed before, this will create two databases. The files will have names corresponding to their **name** appended with the .h5 extension (i.e. `Grid_out.h5` and `MC_out.h5`).

21.3.10 Modified Version of the Institute of Nuclear Safety System Incorporated (Japan)

The Institute of Nuclear Safety System Incorporated (Japan) has modified the **RELAP5** source code in order to be able to control some additional parameters from an auxiliary input file (**modelPar.inp**). In order to use this interface, the user needs to input the $subType$ attribute**Relap5inssJp**:

```
<Models>
  <Code name='MyRELAP' subType='Relap5'>
    <executable>~/path_to_the_executable</executable>
    <!-- here is taking the output from the first deck only -->
    <outputDeckNumber>1</outputDeckNumber>
  </Code>
</Models>
```

For perturbing such input file, the approach presented in section 21.1 (Generic Interface) has been employed. For the standard **RELAP5** input, the same approach previously in this section is used. For example, in the following Sampler block, the card 9100101 is perturbed with the same approach used in standard **RELAP5**; in addition, the variable $modelParTest$ is going to be perturbed in the **modelPar.inp** input file.

```
    <MonteCarlo name="mc_loca">
      <samplerInit>
        <limit>1</limit>
      </samplerInit>
```

```
      <variable name="9100101:3">
        <distribution>break_size</distribution>
      </variable>
      <variable name="modelParTest">
          <distribution>break_size</distribution>
      </variable>
    </MonteCarlo>
```

21.4 RELAP7 Interface

This section covers the input specifications for running RELAP7 through RAVEN. It is important to notice that this short explanation assumes that the reader already knows how to use the control logic system in RELAP7. Since the presence of the control logic system in RELAP7, this code interface is different with respect to the others and uses some special keyword available in RAVEN (see the following).

21.4.1 Files

In the **<Files>** section, as specified before, all of the files needed for the code to run should be specified. In the case of RELAP7, the files typically needed are the following:

- RELAP7 Input file
- Control Logic file

Example:

```
<Files>
  <Input name='nat_circ.i' type=''>nat_circ.i</Input>
  <Input name='control_logic.py' type=''>control_logic.py</Input>
</Files>
```

The RAVEN/RELAP7 interface recognizes as RELAP7 inputs the files with the extensions “*.i”, “*.inp” and “*.in”.

21.4.2 Models

For the **<Models>** block RELAP7 uses the RAVEN executable, since through this executable the stochastic environment gets activated (possibility to sample parameters directly in the control logic

system) Here is a standard example of what can be used to use RELAP7 as the model:

```
<Models>
    <Code name='MyRAVEN'
       subType='RAVEN'><executable>~path/to/RAVEN-opt</executable></Code>
</Models>
```

21.4.3 Distributions

As for all the other codes interfaces the **<Distributions>** block needs to be specified in order to employ as sampling strategy (e.g. MonteCarlo, Stratified, etc.). In this block, the user specifies the distributions that need to be used. Once the user defines the distributions in this block, RAVEN activates the Distribution environment in the RAVEN/RELAP7 control logic system. The sampling of the parameters is then performed directly in the control logic input file.

For example, let's consider the sampling of a normal distribution for the primary pressure in RELAP7:

```
<Distributions>
 <Normal name="Prim_Pres">
 <mean>1000000</mean>
 <sigma>100<sigma/>
 </Normal>
</Distributions>
```

In order to change a parameter (independently on the sampling strategy), the control logic input file should be modified as follows:

```
def initial_function(monitored, controlled, auxiliary)
    print("monitored",monitored,"controlled",
    controlled,"auxiliary",auxiliary)

    controlled.pressureInPressurizer =
     distributions.Prim_Pres.getDistributionRandom()
    return
```

21.4.4 Samplers

In the **<Samplers>** block, all the variables that needs to be sampled must be specified. In case some of these variables are directly sampled in the Control Logic system, the **<variable>** needs

to be replaced with <Distribution>. In this way, RAVEN is able to understand which variables needs to be directly modified through input file (i.e. modifying the original input file *.i) and which variables are going to be "sampled" through the control logic system. For the example, we are performing Grid Sampling. The global initial pressure wasn't specified in the control logic so it is going to be specified using the node <variable>. The "pressureInPressurizer" variable is instead sampled in the control logic system; for this reason, it is going to be specified using the node <Distribution>. For example,

```
<Samplers>
 <Grid name="MC_samp">
   <samplerInit> <limit>500</limit> </samplerInit>
   <variable name="GlobalParams|global_init_P">
      <distribution>Prim_Pres</distribution>
      <grid construction="equal" steps="10" type="CDF">0.0
         1.0</grid>
   </variable>
   <Distribution name="pressureInPressurizer">
      <distribution>Prim_Pres</distribution>
      <grid construction="equal" steps="10" type="CDF">0.0
         1.0</grid>
   </Distribution>
 </Grid>
</Samplers>
```

21.5 MooseBasedApp Interface

21.5.1 Files

In the <Files> section, as specified before, all of the files needed for the code to run should be specified. In the case of any MooseBasedApp, the files typically needed are the following:

- MooseBasedApp YAML input file
- Restart Files (if the calculation is instantiated from a restart point)

Example:

```
<Files>
  <Input name='mooseBasedApp.i' type=''>mooseBasedApp.i</Input>
  <Input name='0020_mesh.cpr' type=''>0020_mesh.cpr</Input>
  <Input name='0020.xdr.0000'>0020.xdr.0000</Input>
```

```
  <Input name='0020.rd-0'>0020.rd-0</Input>
</Files>
```

21.5.2 Models

In the <Models> block particular MooseBasedApp executable needs to be specified. Here is a standard example of what can be used to use with a typical MooseBasedApp (Bison) as the model:

```
<Models>
    <Code name='MyMooseBasedApp'
       subType='MooseBasedApp'><executable>~path/to/Bison-opt</executable></
</Models>
```

21.5.3 Distributions

The <Distributions> block defines the distributions that are going to be used for the sampling of the variables defined in the <Samplers> block. For all the possible distributions and all their possible inputs please see the chapter about Distributions (see 11). Here we give a general example of three different distributions:

```
<Distributions>
    <Normal name='ThermalConductivity1'>
        <mean>1</mean>
        <sigma>0.001</sigma>
        <lowerBound>0.5</lowerBound>
        <upperBound>1.5</upperBound>
    </Normal>
    <Normal name='SpecificHeat'>
        <mean>1</mean>
        <sigma>0.4</sigma>
        <lowerBound>0.5</lowerBound>
        <upperBound>1.5</upperBound>
    </Normal>
    <Triangular name='ThermalConductivity2'>
        <apex>1</apex>
        <min>0.1</min>
        <max>4</max>
    </Triangular>
</Distributions>
```

It is good practice to name the distribution something similar to what kind of variable is going to be sampled, since there might be many variables with the same kind of distributions but different input parameters.

21.5.4 Samplers

In the `<Samplers>` block we want to define the variables that are going to be sampled. **Example**: We want to do the sampling of 3 variables:

- Thermal Conductivity of the Fuel;
- Specific Heat Transfer Ratio of the Cladding;
- Thermal Conductivity of the Cladding.

We are going to sample these 3 variables using two different sampling methods: Grid and Monte-Carlo.

In order to perturb any MooseBasedApp, the user needs to specify the variables to be sampled indicating the path to the value separated with the symbol "|". For example, if the variable that we want to perturb is specified in the input as follows:

```
[Materials]
  ...
  [./heatStructure]
     ...
     thermal_conductivity = 1.0
     ...
  [../]
  ...
[]
```

the variable name in the Sampler input block needs to be named as follows:

```
...
<Samplers>
  <aSampler name='aUserDefinedName' >
    <variable
       name='Materials|heatStructure|thermal_conductivity'>
      ...
    </variable>
  </aSampler>
```

```
</Samplers>
...
```

In this example, we proceed to do so for both the Grid sampling and the Monte-Carlo sampling.

```
<Samplers verbosity='debug'>
    <Grid name='myGrid'>
      <variable
         name='Materials|heatStructure1|thermal_conductivity' >
        <distribution>ThermalConductivity1</distribution>
        <grid          type='value' construction='custom' >0.6
           0.7 0.8</grid>
      </variable>
      <variable name='Materials|heatStructure1|specific_heat' >
        <distribution >SpecificHeat</distribution>
        <grid          type='CDF'    construction='custom'>0.5
           1.0 0.0</grid>
      </variable>
      <variable
         name='Materials|heatStructure2|thermal_conductivity'>
        <distribution  >ThermalConductivity2</distribution>
        <grid type='value' upperBound='4' construction='equal'
           steps='1'>0.5</grid>
      </variable>
    </Grid>
  <MonteCarlo name='MC_Sampler' limit='1000'>
      <variable
         name='Materials|heatStructure1|thermal_conductivity' >
        <distribution>ThermalConductivity1</distribution>
      </variable>
      <variable name='Materials|heatStructure1|specific_heat' >
        <distribution >SpecificHeat</distribution>
      </variable>
      <variable
         name='Materials|heatStructure2|thermal_conductivity'>
        <distribution  >ThermalConductivity2</distribution>
      </variable>
  </MonteCarlo>
</Samplers>
```

21.5.5 Steps

For a MooseBasedApp, the <MultiRun> step type will most likely be used, as first step. First, the step needs to be named: this name will be one of the names used in the <Sequence> block. In our example, Grid_Sampler and MC_Sampler.

```
    <MultiRun name='Grid_Sampler' >
```

With this step, we need to import all the files needed for the simulation:

- MooseBasedApp YAML input file;
- eventual restart files (optional);
- other auxiliary files (e.g., powerHistory tables, etc.).

```
    <Input    class='Files' type=''>mooseBasedApp.i</Input>
    <Input    class='Files' type=''>0020_mesh.cpr</Input>
    <Input    class='Files' type=''>0020.xdr.0000</Input>
    <Input    class='Files' type=''>0020.rd-0</Input>
```

We then need to define which model will be used:

```
    <Model   class='Models' type='Code'>MyMooseBasedApp</Model>
```

We then need to specify which Sampler is used, and this can be done as follows:

```
    <Sampler class='Samplers' type='Grid'>Grid_Sampler</Sampler>
```

And lastly, we need to specify what kind of output the user wants. For example the user might want to make a database (in RAVEN the database created is an HDF5 file) and a DataObject of type PointSet, to use in sub-sequential post-processing. Here is a classical example:

```
    <Output   class='Databases' type='HDF5'>MC_out</Output>
    <Output   class='DataObjects'
      type='PointSet'>MCOutData</Output>
```

Following is the example of two MultiRun steps which use different sampling methods (grid and Monte Carlo), and creating two different databases for each one:

```
<Steps verbosity='debug'>
  <MultiRun name='Grid_Sampler' verbosity='debug'>
    <Input   class='Files' type=''>mooseBasedApp.i</Input>
```

```
    <Input  class='Files' type=''>0020_mesh.cpr</Input>
    <Input  class='Files' type='' >0020.xdr.0000</Input>
    <Input  class='Files' type=''>0020.rd-0</Input>
    <Model  class='Models'    type='Code'>MyMooseBasedApp</Model>
    <Sampler class='Samplers'  type='Grid'>Grid_Sampler</Sampler>
    <Output  class='Databases' type='HDF5'>Grid_out</Output>
    <Output  class='DataObjects'
      type='PointSet'>gridOutData</Output>
  </MultiRun>
  <MultiRun name='MC_Sampler' verbosity='debug'
     re-seeding='210491'>
    <Input  class='Files' type=''>mooseBasedApp.i</Input>
    <Input  class='Files' type=''>0020_mesh.cpr</Input>
    <Input  class='Files' type='' >0020.xdr.0000</Input>
    <Input  class='Files' type=''>0020.rd-0</Input>
    <Model  class='Models'    type='Code'>MyMooseBasedApp</Model>
    <Sampler class='Samplers'  type='MonteCarlo'
      >MC_Sampler</Sampler>
    <Output  class='Databases' type='HDF5' >MC_out</Output>
    <Output  class='DataObjects'
      type='PointSet'>MCOutData</Output>
  </MultiRun>
</Steps>
```

21.5.6 Databases

As shown in the **<Steps>** block, the code is creating two database objects called Grid_out and MC_out. So the user needs to input the following:

```
<Databases>
  <HDF5 name="Grid_out" readMode="overwrite"/>
  <HDF5 name="MC_out" readMode="overwrite"/>
</Databases>
```

As listed before, this will create two databases. The files will have names corresponding to their **name** appended with the .h5 extension (i.e. Grid_out.h5 and MC_out.h5).

21.5.7 DataObjects

As shown in the **<Steps>** block, the code is creating two DataObjects of type PointSet called `gridOutData` and `MCOutData`. So the user needs to input the following:

```
<DataObjects>
    <PointSet name='gridOutData'>
      <Input>
          Materials|heatStructure2|thermal_conductivity,
          Materials|heatStructure1|specific_heat,
          Materials|heatStructure2|thermal_conductivity
      </Input>
      <Output>aveTempLeft</Output>
    </PointSet>
    <PointSet name='MCOutData'>
      <Input>
          Materials|heatStructure2|thermal_conductivity,
          Materials|heatStructure1|specific_heat,
          Materials|heatStructure2|thermal_conductivity
      </Input>
      <Output>aveTempLeft</Output>
    </PointSet>
</DataObjects>
```

As listed before, this will create two DataObjects that can be used in sub-sequential post-processing.

21.5.8 OutStreams

As fully explained in section 16, if the user want to print out or plot the content of a **DataObjects**, he needs to create an **OutStream** in the **<OutStreams>** XML block.
As it shown in the example below, for MooseBasedApp (and any other Code interface that might use the symbol | for the Sampler's variable syntax), in the Plot **<x>** and **<y>** specification, the user needs to utilize curly brackets.

```
<OutStreams>
  <Print name='gridOutDataDumpCSV'>
    <type>csv</type>
    <source>gridOutData</source>
  </Print>
   <Plot verbosity='debug' name='test'   overwrite='False'>
    <plotSettings>
       <plot>
```

```
        <type>line</type>
        <x>MCOutData|Input|{Materials|heatStructure2|thermal_conductivity}</
        <y>MCOutData|Output|aveTempLeft</y>
        <kwargs><color>blue</color></kwargs>
      </plot>
    </plotSettings>
    <actions><how>screen,png</how></actions>
  </Plot>
</OutStreams>
```

21.6 MooseVPP Interface

The Moose Vector Post Processor is used mainly in the solid mechanics analysis. This interface loads the values of the vector ouput processor to a **<DataObjects>** object.

To use this interface the [DomainIntegral] needs to be present in the MooseBasedApp's input file and the subnode **<fileargs>** should be defined in the subnode **<Code>** in the **<Models>** block of the RAVEN input file. The **<fileargs>** is required to have attributes with the below specified values:

- type, ***string, required field***, must be "MooseVPP"
- arg, ***string, required field***, the string value attached to the vector post processor action while creating the output files.

This interface is actually identical to the MooseBasedApp interface, however there is few constraints on defining the output values of the post processor. The definition of these outputs in the **<DataObjects>** depends on the definition of the [DomainIntegral].

The location of the value outputted is defined as *ID#* and the value is as *value#*. The "#" defines the number of the location. The example below contains 3 locations in the [DomainIntegral] where the values are outputted.

Example:

```
...
  <Models>
    <Code name="MOOSETestApp" subType="MooseBasedApp">
      <executable>%FRAMEWORK_DIR%/../../moose/
        modules/combined/modules-%METHOD%</executable>
      <fileargs type = "MooseVPP" arg = "_J_1_" />
```

```
      <alias variable = "poissonsRatio" >
        Materials|stiffStuff|poissons_ratio</alias>
      <alias variable = "youngModulus"   >
        Materials|stiffStuff|youngs_modulus</alias>
    </Code>
  </Models>
 ...
  <DataObjects>
    <PointSet name="collset">
      <Input>youngModulus,poissonsRatio</Input>
      <Output>ID1,ID2,ID3,value1,value2,value3</Output>
    </PointSet>
  </DataObjects>
 ...
```

21.7 OpenModelica Interface

OpenModelica (http://www.openmodelica.org) is an open souce implementation of the Modelica simulation language. Modelica is "a non-proprietary, object-oriented, equation based language to conveniently model complex physical systems containing, e.g., mechanical, electrical, electronic, hydraulic, thermal, control, electric power or process-oriented subcomponents."[1]. Modelica models are specified in text files with a file extension of .mo. A standard Modelica example called BouncingBall which simulates the trajectory of an object falling in one dimension from a height is shown as an example:

```
model BouncingBall
  parameter Real e=0.7 "coefficient␣of␣restitution";
  parameter Real g=9.81 "gravity␣acceleration";
  Real h(start=1) "height␣of␣ball";
  Real v "velocity␣of␣ball";
  Boolean flying(start=true) "true,␣if␣ball␣is␣flying";
  Boolean impact;
  Real v_new;
  Integer foo;

equation
  impact = h <= 0.0;
  foo = if impact then 1 else 2;
  der(v) = if flying then -g else 0;
  der(h) = v;
```

[1]http://www.modelica.org

```
  when {h <= 0.0 and v <= 0.0,impact} then
    v_new = if edge(impact) then -e*pre(v) else 0;
    flying = v_new > 0;
    reinit(v, v_new);
  end when;

end BouncingBall;
```

21.7.1 Files

An OpenModelica installation specific to the operating system is used to create a stand-alone executable program that performs the model calculations. A separate XML file containing model parameters and initial conditions is also generated as part of the build process. The RAVEN OpenModelica interface modifies input parameters by changing copies of this file. Both the executable and XML parameter file names must be provided to RAVEN. In the case of the BouncingBall model previously mentioned on the Windows operating system, the <Files>specification would look like:

```
<Files>
  <Input name='BouncingBall_init.xml'
     type=''>BouncingBall_init.xml</Input>
  <Input name='BouncingBall.exe' type=''>BouncingBall.exe</Input>
</Files>
```

21.7.2 Models

OpenModelica models may provide simulation output in a number of formats. The particular format used is specified during the model generation process. RAVEN works best with Comma-Separated Value (CSV) files, which is one of the possible output format options. Models are generated using the OpenModelica Shell (OMS) command-line interface, which is part of the OpenModelica installation. To generate an executable that provides CSV-formatted output, use OMSl commands as follows:

1. Change to the directory containing the .mo file to generate an executable for:

```
>> cd("C:/MinGW/msys/1.0/home/bobk/projects/raven/framework/
   ↪ CodeInterfaces/OpenModelica")
"C:/MinGW/msys/1.0/home/bobk/projects/raven/framework/
   ↪ CodeInterfaces/OpenModelica"
```

2. Load the model file into memory:

```
>> loadFile("BouncingBall.mo")
true
```

3. Create the model executable, specifying CSV output format:

```
>> buildModel(BouncingBall, outputFormat="csv")
{"C:/MinGW/msys/1.0/home/bobk/projects/raven/framework/
   ↪ CodeInterfaces/OpenModelica/BouncingBall","
   ↪ BouncingBall_init.xml"}
Warning: The initial conditions are not fully specified. Use
   ↪ +d=initialization for more information.
```

At this point the model executable and XML initialization file should have been created in the same directory as the original model file.

The model executable is specified to RAVEN using the <Models>section of the input file as follows:

```
<Simulation>
    ...
  <Models>
    <Code name="BouncingBall" subType = "OpenModelica">
      <executable>BouncingBall.exe</executable>
    </Code>
  </Models>
    ...
</Simulation>
```

21.7.3 CSV Output

The CSV files produced by OpenModelica model executables require adjustment before it may be read by RAVEN. The first few lines of original CSV output from the BouncingBall example is shown below:

```
"time","h","v","der(h)","der(v)","v_new","foo","flying","impact",
0,1,0,0,-9.810000000000001,0,2,1,0,
  ...
```

RAVEN will not properly read this file as-generated for two reasons:

- The variable names in the first line are each enclosed in double-quotes.
- Each line has a trailing comma.

The OpenModelica inteface will automatically remove the double-quotes and trailing commas through its implementation of the finalizeCodeOutput function.

21.8 Dymola Interface

Modelica is ”a non-proprietary, object-oriented, equation-based language to conveniently model complex physical systems containing, e.g., mechanical, electrical, electronic, hydraulic, thermal, control, electric power or process-oriented subcomponents.”[2]. Modelica models (with a file extension of .mo) are built, translated (compiled), and simulated in Dymola (http://www.modelon.com/products/dymola/), which is a commercial modeling and simulation environment based on the Modelica modeling language. A standard Modelica example called BouncingBall, which simulates the trajectory of an object falling in one dimension from a height, is shown as an example:

```
model BouncingBall
  parameter Real e=0.7 "coefficient of restitution";
  parameter Real g=9.81 "gravity acceleration";
  parameter Real hstart = 10 "height of ball at time zero";
  parameter Real vstart = 0 "velocity of ball at time zero";
  Real h(start=hstart,fixed=true) "height of ball";
  Real v(start=vstart,fixed=true) "velocity of ball";
  Boolean flying(start=true) "true, if ball is flying";
  Boolean impact;
  Real v_new;
  Integer foo;

equation
  impact = h <= 0.0;
  foo = if impact then 1 else 2;
  der(v) = if flying then -g else 0;
  der(h) = v;

  when {h <= 0.0 and v <= 0.0,impact} then
    v_new = if edge(impact) then -e*pre(v) else 0;
    flying = v_new > 0;
    reinit(v, v_new);
  end when;
```

[2]http://www.modelica.org

```
  annotation (uses(Modelica(version="3.2.1")),
    experiment(StopTime=10, Interval=0.1),
    __Dymola_experimentSetupOutput);

end BouncingBall;
```

21.8.1 Files

When a modelica model, e.g., BouncingBall model, is implemented in Dymola, the platform dependent C-code from a Modelica model and the corresponding executable code (i.e., by default dymosim.exe on the Windows operating system) are generated for simulation. After the executable is generated, it may be run multiple times (with Dymola license). A separate TEXT file (by default dsin.txt) containing model parameters and initial conditions are also generated as part of the build process. The RAVEN Dymola interface modifies input parameters by changing copies of this file. Both the executable and TEXT parameter file (or simulation initialization file) names must be provided to RAVEN. The TEXT parameter file must be of type 'DymolaInitialisation'. In the case of the BouncingBall model previously mentioned on the Windows operating system, the <Files>specification would look like:

```
<Files>
  <Input name='dsin.txt'
     ↪ type='DymolaInitialisation'>dsin.txt</Input>
</Files>
```

The Dymola interface can only pass scalar values into the TEXT parameter file. If the user wants to pass vector information to Dymola, he can do so by providing an optional TEXT vector file to Dymola. This file must have the type 'DymolaVectors'. This additional file can then be read by the Dymola model. If vecor data is passed from RAVEN to the Dymola interface and the TEXT vector file is not specified, the interface will display an error and stop the Dymola execution. If the TEXT vector file is specified (and vector data is passed to the interface), the interface will write the datd into the specified file, but also display a warning, saying that the Dymola interface found vector data to be passed and if this data is supposed to go into the simulation initialisation file of type 'DymolaInitialisation' the array must be split into scalars. The <Files>specification for the vector data look as follows:

```
<Files>
  <Input name='timeSeriesData.txt'
     ↪ type='DymolaVectors'>timeSeriesData.txt</Input>
</Files>
```

21.8.2 Models

An executable (dymosim.exe) and a simulation initialization file (dsin.txt) can be generated after either translating or simulating the Modelica model (BouncingBall.mo) using the Dymola Graphical User Interface (GUI) or Dymola Application Programming Interface (API)-routines. To generate an executable and a simulation initialization file, use the Dymola API-routines (or Dymola GUI) to translate the model as follows:

1. Change to the directory containing the .mo file to generate an executable. In Dymola GUI, this corresponds to File/Change Directory in menus:

```
>> cd("C:/msys64/home/KIMJ/projects/raven/framework/
    ↪ CodeInterfaces/Dymola");
C:/msys64/home/KIMJ/projects/raven/framework/CodeInterfaces/
    ↪ Dymola
 = true
```

2. Reads the specified file and displays its window. In Dymola GUI, this corresponds to File/Open in the menus:

```
>> openModel("BouncingBall.mo")
 = true
```

3. Compile the model (with current settings), and create the model executable and the corresponding simulation initialization file. In Dymola GUI, this corresponds to Translate Model in the menus:

```
>> translateModel("BouncingBall");
 = true
```

   At this point the model executable and the simulation initialization file should have been created in the same directory as the original model file. Additionally, they could be created by simulating the model. The following command corresponds to Simulate in the menus in Dymola GUI:

```
>> simulateModel("BouncingBall", stopTime=10,
    ↪ numberOfIntervals=0, outputInterval=0.1, method="dassl"
    ↪ , resultFile="BouncingBall");
 = true
```

   The file extension (.mat) is automatically added to a output file (resultFile), e.g., BouncingBall.mat. If the generated executable code is triggered directly from a command prompt, the output file is always named as ”dsres.mat”.

The model executable is specified to RAVEN using the <Models>section of the input file as follows:

```
<Simulation>
    ...
  <Models>
    <Code name="BouncingBall" subType = "Dymola">
      <executable>dymosim.exe</executable>
    </Code>
  </Models>
    ...
</Simulation>
```

RAVEN works best with Comma-Separated Value (CSV) files. Therefore, the default .mat output type needs to be converted to .csv output. The Dymola interface will automatically convert the .mat output to human-readable forms, i.e., .csv output, through its implementation of the finalizeCodeOutput function.
In order to speed up the reading and conversion of the .mat file, the user can specify the list of variables (in addition to the Time variable) that need to be imported and converted into a csv file minimizing the IO memory usage as much as possible. Within the **<Code>** the following XML node (in addition ot the **<executable>** one) can be inputted:

- **<outputVariablesToLoad>**, ***space separated list, optional parameter***, a space separated list of variables that need be exported from the .mat file (in addition to the Time variable).
  Default: all the variables in the .mat file.

For example:

```
<Simulation>
    ...
  <Models>
    <Code name="BouncingBall" subType = "Dymola">
      <executable>dymosim.exe</executable>
      <outputVariablesToLoad>var1 var2
         ↪ var3</outputVariablesToLoad>
    </Code>
  </Models>
    ...
</Simulation>
```

21.9 Mesh Generation Coupled Interfaces

Some software requires a provided mesh that requires a separate code run to generate. In these cases, we use sampled geometric variables to generate a new mesh for each perturbation of the original problem, then run the input with the remainder of the perturbed parameters and the perturbed mesh. RAVEN currently provides two interfaces for this type of calculation, listed below.

21.9.1 MooseBasedApp and Cubit Interface

Many MOOSE-based applications use Cubit (`https://cubit.sandia.gov`) to generate Exodus II files as geometry and meshing for calculations. To use the developed interface, Cubit's bin directory must be added to the user's PYTHONPATH. Input parameters for Cubit can be listed in a journal (`.jou`) file. Parameter values are typically hardcoded into the Cubit command syntax, but variables may be predefined in a journal file through Aprepro syntax. This is an example of a journal file that generates a rectangle of given height and width, meshes it, defines its volume and sidesets, lists its element type, and writes it as an Exodus file:

```
#{x = 3}
#{y = 3}
#{out_name = "'out_mesh.e'"}
create surface rectangle width {x} height {y} zplane
mesh surface 1
set duplicate block elements off
block 1 surface 1
Sideset 1 curve 3
Sideset 2 curve 4
Sideset 3 curve 1
Sideset 4 curve 2
Block all element type QUAD4
export genesis {out_name} overwrite
```

The first three lines are the Aprepro variable definitions that RAVEN requires to insert sampled variables. All variables that RAVEN samples need to be defined as Aprepro variables in the journal file. One essential caveat to running this interface is that an Aprepro variable MUST be defined with the name ”out_name”. In order to run this script without RAVEN inserting the correct syntax for the output file name and properly generate the Exodus file for a mesh, the output file name is REQUIRED to be in both single and double quotation marks with the file extension appended to the end of the file base name (e.g. ’”output_file.e”’).

21.9.1.1 Files

<Files> works the same as in other interfaces with name and type attributes for each node entry. The name attribute is a user-chosen internal name for the file contained in the node, and type identifies which base-level interface the file is used within. <type> should only be specified for inputs that RAVEN will perturb. For Moose input files, <type> should be 'MooseInput' and for Cubit journal files, the <type> should be 'CubitInput'. The node should contain the path to the file from the working directory. The following is an example of a typical <Files> block.

```
<Files>
  <Input name='moose_test'
     ↪ type='MooseInput'>simple_diffusion.i</Input>
  <Input name='mesh_in'
     ↪ type='CubitInput'>rectangle.jou</Input>
  <Input name='other_file' type=''
     ↪ >some_file_moose_input_needs.ext</Input>
</Files>
```

21.9.1.2 Models

A user provides paths to executables and aliases for sampled variables within the <Models> block. The <Code> block will contain attributes name and subType. Name identifies that particular <Code> model within RAVEN, and subType specifies which code interface the model will use. The <executable> block should contain the absolute or relative (with respect to the current working directory) path to the MooseBasedApp that RAVEN will use to run generated input files. The absolute or relative path to the Cubit executable is specified within <preexec>. If the <preexec> block is not needed, the MooseBasedApp interface is probably preferable to the Cubit-Moose interface.

Aliases are defined by specifying the variable attribute in an <alias> node with the internal RAVEN variable name chosen with the node containing the model variable name. The Cubit-Moose interface uses the same syntax as the MooseBasedApp to refer to model variables, with pipes separating terms starting with the highest YAML block going down to the individual parameter that RAVEN will change. To specify variables that are going to be used in the Cubit journal file, the syntax is ”Cubit—aprepro_var”. The Cubit-Moose interface will look for the Cubit tag in all variables passed to it and upon finding it, send it to the Cubit interface. If the model variable does not begin with 'Cubit', the variable MUST be specified in the MooseBasedApp input file. While the model variable names are not required to have aliases defined (the <alias> blocks are optional), it is highly suggested to do so not only to ensure brevity throughout the RAVEN input, but to easily identify where variables are being sent in the interface.

An example <Models> block follows.

```
<Models>
  <Code name="moose-modules" subType="CubitMoose">
    <executable>%FRAMEWORK_DIR%/../../moose/modules/combined/...
      modules-%METHOD%</executable>
    <preexec>/hpc-common/apps/local/cubit/13.2/bin/cubit</preexec>
    <alias variable="length">Cubit@y</alias>
    <alias variable="bot_BC">BCs|bottom|value</alias>
  </Code>
</Models>
```

21.9.1.3 Distributions

The <Distributions> block defines all distributions used to sample variables in the current RAVEN run.

For all the possible distributions and their possible inputs please refer to the Distributions chapter (see 11). It is good practice to name the distribution something similar to what kind of variable is going to be sampled, since there might be many variables with the same kind of distributions but different input parameters.

21.9.1.4 Samplers

The <Samplers> block defines the variables to be sampled.

After defining a sampling scheme, the variables to be sampled and their distributions are identified in the <variable> blocks. The name attribute in the <variable> block must either be the full MooseBasedApp model variable name or the alias name specifed in <Models>. If the sampled variable is a geometric property that will be used to generate a mesh with Cubit, remember the syntax for variables being passed to journal files (Cubit—aprepro_var).

For listings of available samplers refer to the Samplers chapter (see 12).

See the following for an example of a grid based sampler for length and the bottom boundary condition (both of which have aliases defined in <Models>).

```
<Samplers>
  <Grid name="Grid_sampling">
    <variable name="length" >
      <distribution>length_dist</distribution>
      <grid type="value" construction="custom">1.0 2.0</grid>
    </variable>
```

```xml
    <variable name="bot_BC">
      <distribution>bot_BC_dist</distribution>
      <grid type="value" construction="custom">3.0 6.0</grid>
    </variable>
  </Grid>
</Samplers>
```

21.9.1.5 Steps,OutStreams,DataObjects

This interface's **<Steps>**, **<OutStreams>**, and **<DataObjects>** blocks do not deviate significantly from other interfaces' respective nodes. Please refer to previous entries for these blocks if needed.

21.9.1.6 File Cleanup

The Cubit-Moose interface automatically removes files that are commonly unwanted after the RAVEN run reaches completion. Cubit has been described as ”talkative” due to additional journal files with execution information being generated by the program after every completed journal file run. The quantity of these files can quickly become unwieldly if the working directory is not kept clean; thus these files are removed. In addition, some users may wish to remove Exodus files after the RAVEN run is complete as the typical size of each file is quite large and it is assumed that any output quantities of interest will be collected by appropriate postprocessors and the OutStreams. Exodus files are not automatically removed, but by using the **<deleteOutExtension>** node in **<RunInfo>**, one may specify the Exodus extension to save a fair amount of storage space after RAVEN completes a sequence. For example:

```xml
<RunInfo>
  ...
  <deleteOutExtension>e</deleteOutExtension>
  ...
</RunInfo>
```

21.9.2 MooseBasedApp and Bison Mesh Script Interface

For BISON users, a Python mesh generation script is included in the %BISON_DIR%/tools/UO2/ directory. This script generates 3D or 2D (RZ) meshes for nuclear fuel rods using Cubit with templated commands. The BISON Mesh Script (BMS) is capable of generating rods with discrete fuel pellets of various size in assorted configurations. To use this interface, Cubit's bin directory must be added to the user's PYTHONPATH.

21.9.2.1 Files

Similar to the Cubit-Moose interface, the BisonAndMesh interface requires users to specify all files required to run their input so that these file may be copied into the respective sequence's working directory. The user will give each file an internal RAVEN designation with the name attribute, and the MooseBasedApp and BISON Mesh Script inputs must be assigned their respective types in another attribute of the <Input> node. An example follows.

```
<Files>
  <Input name='bison_test'
     ↪ type='MooseInput'>simple_bison_test.i</Input>
  <Input name='mesh_in'
     ↪ type='BisonMeshInput'>coarse_input.py</Input>
  <Input name='other_file'
     ↪ type=''>some_file_moose_input_needs.ext</Input>
</Files>
```

21.9.2.2 Models

A user provides paths to executables and aliases for sampled variables within the <Models> block. The <Code> block will contain attributes name and subType. name identifies that particular <Code> model within RAVEN, and subType specifies which code interface the model will use. The <executable> block should contain the absolute or relative (with respect to the current working directory) path to the MooseBasedApp that RAVEN will use to run generated input files. The absolute or relative path to the mesh script python file is specified within <preexec>. If the <preexec> block is not needed, use the MooseBasedApp interface.

Aliases are defined by specifying the variable attribute in an <alias> node with the internal RAVEN variable name chosen with the node containing the model variable name. The BisonAndMesh interface uses the same syntax as the MooseBasedApp to refer to model variables, with pipes separating terms starting with the highest YAML block going down to the individual parameter that RAVEN will change. To specify variables that are going to be used in the BISON Mesh Script python input, the syntax is "Cubit—dict_name—var_name". The interface will look for the Cubit tag in all variables passed to it and upon finding the tag, send it to the BISON Mesh Script interface. If the model variable does not begin with Cubit, the variable MUST be specified in the MooseBasedApp input file. While the model variable names are not required to have aliases defined (the <alias> blocks are optional), it is highly suggested to do so not only to ensure brevity throughout the RAVEN input, but to easily identify where variables are being sent in the interface.

An example <Models> block follows.

```
<Models>
```

```
  <Code name="Bison-opt" subType="BisonAndMesh">
    <executable>%FRAMEWORK_DIR%/../../bison/bison-%METHOD%</executable>
    <preexec>%FRAMEWORK_DIR%/../../bison/tools/UO2/mesh_script.py</preexec>
    <alias variable="pellet_radius"
       ↪ >Cubit@Pellet1|outer_radius</alias>
    <alias
       ↪ variable="clad_thickness">Cubit@clad|clad_thickness</alias>
    <alias variable="fuel_k"
       ↪ >Materials|fuel_thermal|thermal_conductivity</alias>
    <alias variable="clad_k"
       ↪ >Materials|clad_thermal|thermal_conductivity</alias>
  </Code>
</Models>
```

21.9.2.3 Distributions

The <Distributions> block defines all distributions used to sample variables in the current RAVEN run.

For all the possible distributions and their possible inputs please refer to the Distributions chapter (see 11). It is good practice to name the distribution something similar to what kind of variable is going to be sampled, since there might be many variables with the same kind of distributions but different input parameters.

21.9.2.4 Samplers

The <Samplers> block defines the variables to be sampled.

After defining a sampling scheme, the variables to be sampled and their distributions are identified in the <variable> blocks. The name attribute in the <variable> block must either be the full MooseBasedApp model variable name or the alias name specified in <Models>. If the sampled variable is a geometric property that will be used to generate a mesh with Cubit, remember the syntax for variables being passed to journal files (Cubit—aprepro_var).

For listings of available samplers refer to the Samplers chapter (see 12).

See the following for an example of a grid based sampler for length and the bottom boundary condition (both of which have aliases defined in <Models>).

```
<Samplers>
  <Grid name="Grid_sampling">
```

```
    <variable name="length" >
      <distribution>length_dist</distribution>
      <grid type="value" construction="custom">1.0 2.0</grid>
    </variable>
    <variable name="bot_BC">
      <distribution>bot_BC_dist</distribution>
      <grid type="value" construction="custom">3.0 6.0</grid>
    </variable>
  </Grid>
</Samplers>
```

21.9.2.5 Steps,OutStreams,DataObjects

This interface's **<Steps>**, **<OutStreams>**, and **<DataObjects>** blocks do not deviate significantly from other interfaces' respective nodes. Please refer to previous entries for these blocks if needed.

21.9.2.6 File Cleanup

The BisonAndMesh interface automatically removes files that are commonly unwanted after the RAVEN run reaches completion. Cubit has been described as ”talkative” due to additional journal files with execution information being generated by the program after every completed journal file run. The BISON Mesh Script creates a journal file to run with cubit after reading input parameters; so Cubit will generate its ”redundant” journal files, and .pyc files will litter the working directory as artifacts of the python mesh script reading from the .py input files. The quantity of these files can quickly become unwieldly if the working directory is not kept clean, thus these files are removed. Some users may wish to remove Exodus files after the RAVEN run is complete as the typical size of each file is quite large and it is assumed that any output quantities of interest will be collected by appropriate postprocessors and the OutStreams. Exodus files are not automatically removed, but by using the **<deleteOutExtension>** node in **<RunInfo>**, one may specify the Exodus extension (*.e) to save a fair amount of storage space after RAVEN completes a sequence. For example:

```
<RunInfo>
  ...
  <deleteOutExtension>e</deleteOutExtension>
  ...
</RunInfo>
```

21.10 Rattlesnake Interfaces

This section covers the input specification for running Rattlesnake through RAVEN. It is important to notice that this short explanation assumes that the reader already knows how to use Rattlesnake. The interface can be used to perturb the Rattlesnake MOOSE-based input file as well as the Yak cross section libraries XML input files (e.g. multigroup cross section libraries) and Instant format cross section libraries.

21.10.1 Files

<Files> works the same as in other interfaces with name and type attributes for each node entry. The **name** attribute is a user-chosen internal name for the file contained in the node, and **type** identifies which base-level interface the file is used within. **type** should only be specified for inputs that RAVEN will perturb. Take Rattlesnake input files for example, **type** should be **'RattlesnakeInput'**.

21.10.1.1 Perturb Yak Multigroup Cross Section Libraries

If the user would like to perturb the Yak multigroup cross section libraries, the user need to use the **'YakXSInput'** for the **type** of the libaries. In addition, the **type** of the alias files that are used to perturb the Yak multigroup cross section libraries should be **'YakXSAliasInput'**. The following is an example of a typical **<Files>** block.

```
<Files>
  <Input name='rattlesnakeInput'
     ↪ type='RattlesnakeInput'>simple_diffusion.i</Input>
  <Input name='crossSection'    type='YakXSInput'>xs.xml</Input>
  <Input name='alias' type='YakXSAliasInput'>alias.xml</Input>
</Files>
```

The alias files are employed to define the variables that will be used to perturb Yak multigroup cross section libraries. The following is an example of a typical alias file:

```
<Multigroup_Cross_Section_Libraries Name="twigl" NGroup="2"
   ↪ Type="rel">
    <Multigroup_Cross_Section_Library ID="1">
        <Fission gridIndex="1" mat="pseudo-seed1"
           ↪ gIndex="1">f11</Fission>
        <Capture gridIndex="1" mat="pseudo-seed1"
           ↪ gIndex="1">c11</Capture>
```

```
        <TotalScattering gridIndex="1" mat="pseudo-seed1"
           ↪ gIndex="1">t11</TotalScattering>
        <Nu gridIndex="1" mat="pseudo-seed1" gIndex="1">n11</Nu>
        <Fission gridIndex="1" mat="pseudo-seed2"
           ↪ gIndex="2">f22</Fission>
        <Capture gridIndex="1" mat="pseudo-seed2"
           ↪ gIndex="2">c22</Capture>
        <TotalScattering gridIndex="1" mat="pseudo-seed2"
           ↪ gIndex="2">t22</TotalScattering>
        <Nu gridIndex="1" mat="pseudo-seed2" gIndex="2">n22</Nu>
        <Fission gridIndex="1" mat="pseudo-seed1-dup"
           ↪ gIndex="1">f11</Fission>
        <Capture gridIndex="1" mat="pseudo-seed1-dup"
           ↪ gIndex="1">c11</Capture>
        <TotalScattering gridIndex="1" mat="pseudo-seed1-dup"
           ↪ gIndex="1">t11</TotalScattering>
        <Nu gridIndex="1" mat="pseudo-seed1-dup"
           ↪ gIndex="1">n11</Nu>
        <Transport gridIndex="1" mat="pseudo-seed1-dup"
           ↪ gIndex="1">d11</Transport>
    </Multigroup_Cross_Section_Library>
</Multigroup_Cross_Section_Libraries>
```

In the above alias file, the **Name** of **<Multigroup_Cross_Section_Libraries>** are used to indicate which Yak multigroup cross section library input file will be perturbed. The **NGroup**, **ID**, and **<Multigroup_Cross_Section_Library>** should be consistent with the Yak multigroup cross section library input files. The **<Fission>**, **<Capture>**, **<TotalScattering>**, **<Nu>**, **gridIndex**, **mat**, and **gIndex** are used to find the corresponding cross sections in the Yak multigroup cross section library input files. For example:

```
<Fission gridIndex="1" mat="pseudo-seed1"
   ↪ gIndex="1">f11</Fission>
```

This node defines an alias with name **'f11'** used to represent the fission cross section at energy group **'1'** for material with name 'pseudo-seed1' at grid index **'1'** in the Yak multigroup cross section library input files.

Note: The attribute **Type="rel"** indicates that the cross sections will be perturbed relatively (i.e. perturbed by percents). In this case, the user also needs to specify a relative covariance matrix for **<covaraince type="rel">** in **<MultivariateNormal>** distribution, and the values for **<mu>** should be 'ones'. In the other case, if the user choose **Type="abs"**, the cross sections will be perturbed absolutely (i.e. perturbed by values), and the user needs to provide an absolute covariance matrix and specify 'zeros' for **<mu>** in **<MultivariateNormal>** distribution.

Note: Currently, only the following cross sections can be perturbed by the user: Fission, Capture, Nu, TotalScattering, and Transport.

21.10.1.2 Perturb Instant format Cross Section Libraries

If the user would like to perturb the Instant cross section libraries, the user need to use the `'InstantXSInput'` for the `type` of the libaries. In addition, the `type` of the alias files that are used to perturb the Instant format cross section libraries should be `'InstantXSAliasInput'`. The following is an example of a typical `<Files>` block.

```
<Files>
  <Input name='rattlesnakeInput'
     ↪ type='RattlesnakeInput'>iaea2d_ls_sn.i</Input>
  <Input name='crossSection'
     ↪ type='InstantXSInput'>iaea2d_materials.xml</Input>
  <Input name='alias'
     ↪ type='InstantXSAliasInput'>alias.xml</Input>
</Files>
```

The alias files are employed to define the variables that will be used to perturb Instant format cross section libraries. The following is an example of a typical alias file:

```
<Materials>
  <Macros NG="2" Type="rel">
    <material ID="1">
      <FissionXS gIndex="1">f11</FissionXS>
      <CaptureXS gIndex="1">c11</CaptureXS>
      <TotalScatteringXS gIndex="1">t11<TotalScatteringXS>
      <Nu gIndex="1">n11</Nu>
      <DiffusionCoefficient gIndex="1">d11</DiffusionCoefficient>
    </material>
  </Macros>
</Materials>
```

In the above alias file, the `NG` and `ID` should be consistent with the Instant format cross section library input files. The `<FissionXS>`, `<CaptureXS>`, `<TotalScatteringXS>`, `<Nu>`, `gIndex`, are used to find the corresponding cross sections in the Instant format cross section library input files. For example, the variable `'f11'` used to represent the fission cross section at energy group `'1'` for material with `'ID'` equal `'1'` in the given cross section library.

Note: The attribute `Type="rel"` indicates that the cross sections will be perturbed relatively (i.e. perturbed by percents). In this case, the user also needs to specify a relative covariance matrix

for <covaraince type="rel"> in <MultivariateNormal> distribution, and the values for <mu> should be 'ones'. In the other case, if the user choose Type="abs", the cross sections will be perturbed absolutely (i.e. perturbed by values), and the user needs to provide an absolute covariance matrix and specify 'zeros' for <mu> in <MultivariateNormal> distribution.

Note: Currently, only the following cross sections can be perturbed by the user: FissionXS, CaptureXS, Nu, TotalScatteringXS, and DiffusionCoefficient.

21.10.2 Models

A user provides paths to executables and aliases for sampled variables within the <Models> block. The <Code> block will contain attributes <name> and <subType>. The <name> identifies that particular <Code> model within RAVEN, and <subType> specifies which code interface the model will use. The <executable> block should contain the absolute or relative (with respect to the current working directory) path to Rattlesnake that RAVEN will use to run generated input files.

An example <Models> block follows.

```
<Models>
  <Code name="Rattlesnake" subType="Rattlesnake">
    <executable>%FRAMEWORK_DIR%/../../rattlesnake/
     rattlesnake-%METHOD%</executable>
  </Code>
</Models>
```

21.10.3 Distributions

The <Distributions> block defines all distributions used to sample variables in the current RAVEN run.

For all the possible distributions and their possible inputs please refer to the Distributions chapter (see 11). It is good practice to name the distribution something similar to what kind of variable is going to be sampled, since there might be many variables with the same kind of distributions but different input parameters.

21.10.3.1 Samplers

The <Samplers> block defines the variables to be sampled. After defining a sampling scheme, the variables to be sampled and their distributions are identified in the <variable> blocks. The

name attribute in the <variable> block must either be the full MooseBasedApp (Rattlesnake) model variable name, the alias name specifed in <Models>, or the variable name specified in the provided alias files.

For listings of available samplers, please refer to the Samplers chapter (see 12). See the following for an example of a grid based sampler for the first energy group fission and capture cross sections (both of which have defined in alias files provided in <Files>).

```
<Samplers>
  <Grid name="Grid_sampling">
    <variable name="fission_group_1" >
      <distribution>fission_dist</distribution>
      <grid type="value" construction="custom">1.0 2.0</grid>
    </variable>
    <variable name="capture_group_1">
      <distribution>capture_dist</distribution>
      <grid type="value" construction="custom">3.0 6.0</grid>
    </variable>
  </Grid>
</Samplers>
```

21.10.4 Steps

For a Rattlesnake interface, the <MultiRun> step type will most likely be used. First, the step needs to be named: this name will be one of the names used in the <Sequence> block. In our example, 'Grid_Rattlesnake'.

```
<MultiRun name='Grid_Rattlesnake' verbosity='debug'>
    <Input   class='Files' type=''>RattlesnakeInput.i</Input>
    <Input   class='Files' type=''>xs.xml</Input>
    <Input   class='Files' type=''>alias.xml</Input>
    <Model   class='Models' type='Code'>Rattlesnake</Model>
    <Sampler class='Samplers'
       ↪ type='Grid'>Grid_Samplering</Sampler>
    <Output  class='DataObjects' type='PointSet'>solns</Ouput>
```

With this step, we need to import all the files needed for the simulation:

- Rattlesnake MOOSE-based input file;
- Yak multigroup cross section libraries input files (XML);

- Yak alias files used to define the perturbed variables (XML).

We then need to define <Model>, <Sampler> and <Output>. The <Output> can be <DataObjects> or <OutStreams>.

21.11 MAAP5 Interface

This section presents the main aspects of the interface coupling RAVEN with MAAP5, the consequent RAVEN input adjustments and the modifications of the MAAP5 files required to run the two coupled codes. The interface works both for forward sampling and the DET, however there are some differences depending on the selected sampling strategy.

21.11.1 RAVEN Input file

21.11.1.1 Files

MAAP5 requires more than one file to run a simulation. This means that, since the <Files> section has to contain all the files required by the external model (MAAP5) to be run, all these files need to be included within this node. This involves not only the input file (.inp) but also the include file, the parameter file, all the files defining the different "PLOTFILs", if any, and the other files which could result useful for the MAAP5 simulation run.

Example:

```
<Files>
  <Input name="test.inp" type="">test.inp</Input>
  <Input name="include" type="">include</Input>
  <Input name="plot.txt" type="">plot.txt</Input>
  <Input name="plant.par" type="">plant.par</Input>
</Files>
```

All the files here mentioned in this section need, then, to be put into the working directory specified by the <workingDir> node into the <RunInfo> block.

21.11.1.2 Models

The <Models> block contains the name of the executable file of MAAP5 (with the path, if necessary), and the name of the interface (e.g. MAAP5_GenericV7). The block has also some required nodes:

- **<boolMaapOutputVariables>**: containing the number of the MAAP5 IEVNT corresponding to the boolean events of interest;
- **<contMaapOutputVariables>**: containing the list of all the continuous variables we are interested at, and that we want to monitor;
- **<stopSimulation>**: this node is required only in case of DET sampling strategy. The user needs to specify if the MAAP5 simulation run stops due to the reached END TIME, specifying "mission_time", or due to the occurrence of a specific event by inserting the number of the corresponding MAAP5 IEVNT (e.g IEVNT(691) for core uncovery)
- **<includeForTimer>**: also this node is required only in case of DET sampling strategy and it contains the name of the MAAP5 include file where the TIMERS for the different variables are defined (see paragraph "MAAP5 include file below" for more information about timers).

A **<Models>** block is shown as an example below:

```
<Models>
  <Code name="MyMAAP" subType="MAAP5\_GenericV7">
    <executable>MAAP5.exe</executable>
    <clargs type='input' extension='.inp'/>
    <boolMaapOutputVariables>691</boolMaapOutputVariables>
    <contMaapOutputVariables>PPS,PSGGEN(1),ZWDC2SG(1)
    </contMaapOutputVariables>
    <stopSimulation>mission_time</stopSimulation>
    <includeForTimer>include</includeForTimer>
  </Code>
</Models>
```

21.11.1.3 Other blocks

All the other blocks (e.g. **<Distributions>**, **<Samplers>**, **<Steps>**, **<Databases>**, **<OutStream>**, etc.) do not require any particular arrangements than already provided by a RAVEN input. User can, therefore, refer to the corresponding sections of the User's Manual. This is valid for both forward sampling and DET.

21.11.2 MAAP5 Input files

The coupling of RAVEN and MAAP5 requires modifications to some MAAP5 files in order to work. This is particularly true when a DET analysis is performed. The MAAP5 input files that need to be modified are:

- MAAP5 include file
- MAAP5 input file (.inp)
- PLOTFIL blocks

21.11.2.1 MAAP5 include file

Usually MAAP5 simulation provides the presence of some include files, for example, containing the user-defined variables, timers, definition of the plotfil, etc. The adjustments explained in this section are required only in case of a DET analysis. The user needs to modify the include file containing the set of the timers used into the run, by adding the definition of the different timers, one for each variable that causes a branching. The include file to be modified should correspond to that one defined in the **<includeForTimer>** block of the RAVEN xml input.

User is supposed to check that the numbers used for the different timers definition are not already used in any of the other MAAP5 files. These timers should be preceeded by a line reporting "C Branching + name of the variable sampled by RAVEN causing the branching".

For example, we assume that DIESEL is the name of the variable corresponding to the failure time of the Diesel generators (user defined). User has to firstly ensure that, for example, "TIMER 100" is not already used into the model, then the following lines need to be added into the selected include file for the set of the timer corresponding to the Diesel generators failure:

```
C Branching DIESEL
WHEN (TIM>DIESEL)
   SET TIMER 100
END
```

It is worth mentioning that at this step a TIMER should be defined also for the event IEVNT specified into the **<stopSimulation>**, if this is the stop condition for the MAAP5 run:

```
WHEN IEVNT(691) == 1.0
  SET TIMER 10
END
```

The interface will check that one timer is defined for each variable of the DET. If not, an error arises suggesting to user the name of the variable having no timer defined.

21.11.2.2 MAAP5 input file

In the ”parameter change” section of the MAAP5 input file, the user should declare the name of the variables sampled by RAVEN according to the following statement:

```
 variable = $ RAVEN-variable:default$
```

where the dafault value is optional.

For example:

```
DIESEL = $RAVEN-DIESEL:-1$
```

This is valid for both forward and DET sampled variables. In particular, in case of DET analysis, the variables causing the occurrence of the branch should be assigned within a block identified by the comment ”C DET Sampled variables”:

```
C DET Sampled Variables
DIESEL = $RAVEN-DIESEL:-1$
C End DET Sampled Variables
```

If the sampled variables are user-defined, then the user shall ensure that they are initialized (to the default value) and set within the user-defined variables section of one of the include file. As usual, a distribution and a sampling strategy should then correspond to each of these variables into the RAVEN xml input file.

Only for the DET analysis, then, the occurence of a branch will be identified by a comment before. This comment is ”C BRANCHING + name of the variable determining the branch” and acts as a sort of branching marker. Looking for these markers, indeed, the interface (in case of DET sampler) verifies that at least one branching exists, and furthermore, that one branching is defined for each of the variables contained into ”DET sampled variables”.

Within the block, the occurrence of the branching leads the value of a variable (user-defined) called ”TIM+number of the corresponding timer set into the include file” to switch to 1.0. The code, in fact, detects if a branch has occurred by monitoring the value of these kind of variables. SInce these variables are user-defined, they need to be initialized to a value (different from 1.0), into the ”user-defined variables” section of one of the include files.

Therefore following the previous example, if we want that, when the diesels failure occurs it leads to the event ”Loss of AC Power” (IEVNT(205) of MAAP5), we will have:

```
C Branching TIMELOCA
WHEN TIM > DIESEL
 TIM100=1.0
 IEVNT(205)=1.0
```

```
END
```

It is worth noticing that no comments should be contained within the line of assignment (i.e. IEVNT(205)=1.0 //LOSS OF AC POWER is not allowed).

Finally, only in case of DET analysis, a stop simulation condition (provided by the comment ”C Stop Simulation condition”) needs to be put into the input. The original input should have all the timers (linked with the branching) separated by an OR condition, even including that one of the event that stops the simulation (e.g. IEVNT(691)), if any.

```
C Stop Simulation condition
IF (TIMER 10 > 0) OR (TIMER 100 > 0) OR ... (TIMER N > 0)
 TILAST=TIM
END
```

This allows the simulation run to stop when a branch condition occurs, creating the restart file that will be used by the two following branches.

For each branch, then, the interface will automatically update the name of the RESTART FILE to be used and of the RESTART TIME that will be equal to the difference between the END TIME of the ”parent” simulation and the PRINT INTERVAL (which specifies the interval at which the restart output is written).

21.11.2.3 MAAP5 PLOTFIL blocks

This section refers to the ”PLOTFIL blocks” used to modify the plot file (.csv) defined into the parameter file. These blocks need to be modified in order to include some variables. It is important, indeed, that the MAAP5 csv PLOTFIL files contain the evolution of:

- RAVEN sampled variables (e.g. DIESEL) (both for Forward and DET sampling)
- the variables whose value is modified by the occurrence of one of the branches, either continuous or boolean (e.g. IEVNT(225))
- the variables of interest defined within **<boolMaapOutputVariables>** and **<contMaapOutputVariables>** blocks (both for Forward and DET sampling)

If one of these variables is not contained into one of csv files, RAVEN will give an error.

21.12 MAMMOTH Interface

This section covers the input specification for running MAMMOTH through RAVEN. It is important to notice that this short explanation assumes that the reader already knows how to use MAMMOTH. The interface can be used to perturb Bison, Rattlesnake, RELAP-7, and general MOOSE input files that utilize MOOSE's standard YAML input structure as well as Yak multigroup cross section library XML input files.

21.12.1 Files

<Files> works the same as in other interfaces with name and type attributes for each node entry. The **name** attribute is a user-chosen internal name for the file contained in the node, and **type** identifies which base-level interface the file is used within. **type** should be specified for all inputs used in RAVEN's MultiRun for MAMMOTH (including files not perturbed by RAVEN). The MAMMOTH input file's **type** should have **'MAMMOTHInput'** prepended to the driver app's input specification (e.g. **'MAMMOTHInput|appNameInput'**). Any other app's input file needs a **type** with the app's name prepended to **'Input'** (e.g. **'BisonInput'**, **'Relap7Input'**, etc.). In addition, the **type** for any mesh input is the app in which that mesh is utilized prepended to **'|Mesh'**; so a Bison mesh would have a **type** of **'Bison|Mesh'** and similarly a mesh for Rattlesnake would have **'Rattlesnake|Mesh'** as its **type**. In cases where a file needs to be copied to each perturbed run directory (to be used as function input, control logic, etc.), one can use the **type** **'AncillaryInput'** to make it clear in the RAVEN input file that this is file is required for the simulation to run but contains no perturbed parameters. For Yak multigroup cross section libraries, the **type** should be **'YakXSInput'**, and for the Yak alias files that are used to perturb the Yak multigroup cross section libraries, the **type** should be **'YakXSAliasInput'**.

The node should contain the path to the file from the working directory. The following is an example of a typical **<Files>** block.

```
<Files>
  <Input name='mammothInput'
    ↪ type='MAMMOTHInput|RattlesnakeInput'>test_mammoth.i</Input>
  <Input name='crossSection'    type='YakXSInput'>xs.xml</Input>
  <Input name='alias' type='YakXSAliasInput'>alias.xml</Input>
  <Input name='bisonInput'
    ↪ type='BisonInput'>test_bison.xml</Input>
  <Input name='bisonMesh'
    ↪ type='Bison|Mesh'>bisonMesh.e</Input>
  <Input name='fuelCTEfunct'
    ↪ type='AncillaryInput'>uo2_CTE.csv</Input>
  <Input name='rattlesnakeMesh'
    ↪ type='Rattlesnake|Mesh'>rattlesnakeMesh.e</Input>
```

```
</Files>
```

The alias files are employed to define the variables that will be used to perturb Yak multigroup cross section libraries. Please see the section 21.10 for the example.

21.12.2 Models

A user provides paths to executables and aliases for sampled variables within the <Models> block. The <Code> block will contain name and subType. The attribute name identifies that particular <Code> model within RAVEN, and subType specifies which code interface the model will use. The <executable> block should contain the absolute or relative (with respect to the current working directory) path to MAMMOTH that RAVEN will use to run generated input files.

An example <Models> block follows.

```
<Models>
  <Code name="Mammoth" subType="MAMMOTH">
    <executable>\%FRAMEWORK_DIR\%/../../mammoth/
     mammoth-%METHOD%</executable>
  </Code>
</Models>
```

21.12.3 Distributions

The <Distributions> block defines all distributions used to sample variables in the current RAVEN run.

For all the possible distributions and their possible inputs please refer to the Distributions chapter (see 11). It is good practice to name the distribution something similar to what kind of variable is going to be sampled, since there might be many variables with the same kind of distributions but different input parameters.

21.12.3.1 Samplers

The <Samplers> block defines the variables to be sampled. After defining a sampling scheme, the variables to be sampled and their distributions are identified in the <variable> blocks. The name attribute in the <variable> block must either be the app's name prepended to the full MooseBasedApp model variable name, the alias name specifed in <Models>, or the variable name specified in the provided alias files.

For listings of available samplers, please refer to the Samplers chapter (see 12). See the following for an example of a grid based sampler used to generate the samples for the first energy group fission and capture cross sections (both of which have defined in alias files provided in <Files>), the initial condition temperature defined in Rattlesnake input file and the poissons ratio, clad thickness, and gap width defined in Bison input files with clad and gap parameters calculated using an external function with sampled clad inner and outer diameters as inputs.

```
<Samplers>
  <Grid name="Grid_sampling">
    <variable name="Rattlesnake@fission_group_1" >
      <distribution>fission_dist</distribution>
      <grid type="value" construction="custom">1.0 2.0</grid>
    </variable>
    <variable name="Rattlesnake@capture_group_1">
      <distribution>capture_dist</distribution>
      <grid type="value" construction="custom">3.0 6.0</grid>
    </variable>
    <variable
      ↪ name="Rattlesnake@AuxVariables|Temp|initial_condition">
      <distribution>uniform</distribution>
      <grid type="value" construction="custom">3.0 6.0</grid>
    </variable>
    <variable
      ↪ name="Bison@Materials|fuel_solid_mechanics_elastic|poissons_ratio"
      <distribution>normal</distribution>
      <grid type="value" construction="custom">3.0 6.0</grid>
    </variable>
    <variable name='clad_outer_diam'>
      <distribution>clad_outer_diam_dist</distribution>
      <grid construction='equal' steps='144' type='CDF'>0.02275
        ↪ 0.97725</grid>
    </variable>
    <variable name='clad_inner_diam'>
      <distribution>clad_inner_diam_dist</distribution>
      <grid construction='equal' steps='144' type='CDF'>0.02275
        ↪ 0.97725</grid>
    </variable>
    <variable name='Bison@Mesh|clad_thickness'>
      <function>clad_thickness_calc</function>
    </variable>
    <variable name='Bison@Mesh|clad_gap_width'>
      <function>clad_gap_width_calc</function>
    </variable>
```

```
  </Grid>
</Samplers>
```

In order to make the input variables of one application distinct from input variables of another, an app's name followed by the '@' symbol is prepended to the variable name (e.g. **'appName@varName'**). Each variable to be used in an app's input file and sampled in the MAMMOTH interface is required to have a destination app specified. All variables utilizing Rattlesnake's executable (whether they are in the Rattlesnake input file or not) are listed as Rattlesnake variables as that application's interface will sort input file and cross section variables itself. Notice that the clad inner and outer diameter sampled parameters have no app name specified. These parameters are utilized to sample values used as inputs for the clad thickness and gap width variables in BISON, so by not specifying a destination app, these are passed through the interface having only been used in an external function to calculate parameters usable in an app's input.

21.12.4 Steps

For a MAMMOTH interface run, the **<MultiRun>** step type will most likely be used. First, the step needs to be named: this name will be one of the names used in the **<Sequence>** block. In our example, **'Grid_Mammoth'**.

```
<MultiRun name='Grid_Mammoth' verbosity='debug'>
    <Input   class='Files' type=''>mammothInput</Input>
    <Input   class='Files' type=''>crossSection</Input>
    <Input   class='Files' type=''>alias</Input>
    <Input   class='Files' type=''>bisonInput</Input>
    <Input   class='Files' type=''>bisonMesh</Input>
    <Input   class='Files' type=''>fuelCTEfunct</Input>
    <Input   class='Files' type=''>rattlesnakeMesh</Input>
    <Model   class='Models' type='Code'>Mammoth</Model>
    <Sampler class='Samplers'
       ↪ type='Grid'>Grid_Samplering</Sampler>
    <Output  class='DataObjects' type='PointSet'>solns</Output>
</MultiRun>
```

With this step, we need to import all the files needed for the simulation:

- MAMMOTH—Rattlesnake YAML input file;
- Yak multigroup cross section libraries input files (XML);
- Yak alias files used to define the perturbed variables (XML);

- Bison YAML input file;
- Bison mesh file;
- Bison function file for the fuel's coefficient of thermal expansion as a function of temperature;
- Rattlesnake mesh file.

As well as **<Model>**, **<Sampler>** and outputs, such as **<OutStreams>** and **<DataObjects>**.

21.13 MELCOR Interface

The current implementation of MELCOR interfaface is valid for MELCOR 2.1/2.2; its validity for MELCOR 1.8 is **not been tested**.

21.13.1 Sequence

In the **<Sequence>** section, the names of the steps declared in the **<Steps>** block should be specified. As an example, if we called the first multirun "Grid_Sampler" and the second multirun "MC_Sampler" in the sequence section we should see this:

```
<Sequence>Grid_Sampler,MC_Sampler</Sequence>
```

21.13.2 batchSize and mode

For the **<batchSize>** and **<mode>** sections please refer to the **<RunInfo>** block in the previous chapters.

21.13.3 RunInfo

After all of these blocks are filled out, a standard example RunInfo block may look like the example below:

```
<RunInfo>
  <WorkingDir>~/workingDir</WorkingDir>
  <Sequence>Grid_Sampler,MC_Sampler</Sequence>
  <batchSize>8</batchSize>
</RunInfo>
```

In this example, the <batchSize> is set to 8; this means that 8 simulatenous (parallel) instances of MELCOR are going to be executed when a sampling strategy is employed.

21.13.4 Files

In the <Files> section, as specified before, all of the files needed for the code to run should be specified. In the case of MELCOR, the files typically needed are:

- MELCOR Input file (file extension ".i" or ".inp")
- Restart file (if present)

Example:

```
<Files>
  <Input name='melcorInputFile' type=''>inputFileMelcor.i</Input>
  <Input name='aRestart' type=''>restartFile</Input>
</Files>
```

It is a good practice to put inside the working directory (<WorkingDir>) all of these files.

It is important to notice that the interface output collection (i.e. the parser of the MELCOR output) currently is able to extract *CONTROL VOLUME HYDRODYNAMICS EDIT* data only. Only those variables are going to be exported and make available to RAVEN. In addition, it is important to notice that:

- **the simulation time is stored in a variable called *"time"*;**
- **all the variables specified in the *CONTROL VOLUME HYDRODYNAMICS EDIT* block are going to be converted using underscores. For example, the following EDITs:**

VOLUME	PRESSURE	TLIQ	TVAP	MASS
	PA	K	K	KG
1	1.00E+07	584.23	584.23	1.66E+03

will be converted in the following way (CSV):

$time$	$volume_1_PRESSURE$	$volume_1_TLIQ$	$volume_1_TVAP$	$volume_1_MASS$
1.0	1.00E+07	584.23	584.23	1.66E+03

Remeber also that a MELCOR simulation run is considered successfull (i.e., the simualtion did not crash) if it terminates with the following message:

Normal termination

If the a MELCOR simulation run stops with messages other than this one than the simulation is considered as crashed, i.e., it will not be saved. Hence, it is strongly recommended to set up the MELCOR input file so that the simulation exiting conditions are set through control logic trip variables.

21.13.5 Models

For the **<Models>** block here is a standard example of how it would look when using MELCOR 2.1/2.2 as the external code:

```
<Models>
  <Code name='MyMELCOR' subType='Melcor'>
    <executable>~/path_to_the_executable_of_melcor</executable>
    <preexec>~/path_to_the_executable_of_melgen</preexec>
  </Code>
</Models>
```

As it can be seen above, the **<preexec>** node must be specified, since MELCOR 2.1/2.2 must run the MELGEN utility code before executing. Once the **<preexec>** node is inputted, the execution of MELGEN is performed automatically by the Interface.
In addition, if some command line parameters need to be passed to MELCOR, the user might use (optionally) the **<clargs>** XML nodes.

```
<Models>
  <Code name='MyMELCOR' subType='Melcor'>
    <executable>~/path_to_the_executable_of_melcor</executable>
    <preexec>~/path_to_the_executable_of_melgen</preexec>
    <clargs type="text" arg="-r whatever command line
       instruction"/>
  </Code>
</Models>
```

21.13.6 Distributions

The **<Distribution>** block defines the distributions that are going to be used for the sampling of the variables defined in the **<Samplers>** block. For all the possibile distributions and all their

possible inputs please see the chapter about Distributions (see 11). Here we report an example of a Normal distribution:

```
<Distributions verbosity='debug'>
    <Normal name="temper">
      <mean>1.E+7</mean>
      <sigma>1.5</sigma>
      <upperBound>9.E+6</upperBound>
      <lowerBound>1.1E+7</lowerBound>
    </Normal>
 </Distributions>
```

It is good practice to name the distribution something similar to what kind of variable is going to be sampled, since there might be many variables with the same kind of distributions but different input parameters.

21.13.7 Samplers

In the <Samplers> block we want to define the variables that are going to be sampled. **Example**: We want to do the sampling of 1 single variable:

- The in pressure (P_in) of a control volume regolated by a Tabular Function TF_TAB

We are going to sample this variable using two different sampling methods: Grid and Monte-Carlo.

The interface of MELCOR uses the ***GenericCode*** (see section 21.1) interface for the input perturbation; this means that the original input file (listed in the <Files> XML block) needs to implement wild-cards. In this example we are sampling the variable:

- *PRE*, which acts on the Tabular Function TF_TAB whose TF_ID is P_in.

We proceed to do so for both the Grid sampling and the MonteCarlo sampling.

```
<Samplers verbosity='debug'>
  <Grid name='Grid_Sampler' >
    <variable name='PRE'>
      <distribution>temper</distribution>
      <grid type='CDF' construction='equal'  steps='10'>0.001
         ↪ 0.999</grid>
    </variable>
```

```
  </Grid>
  <MonteCarlo name='MC_Sampler'>
     <samplerInit>
       <limit>1000</limit>
     </samplerInit>
    <variable name='PRE'>
      <distribution>temper</distribution>
  </MonteCarlo>
</Samplers>
```

It can be seen that each variable is connected with a proper distribution defined in the **<Distributions>** block (from the previous example). The following demonstrates how the input for the variable is read.

We are sampling a variable whose wild-card in the original input file is named $\$RAVEN - PRE\$$ using a Grid sampling method. The distribution that this variable is following is a Normal distribution (see section above). We are sampling this variable beginning from 0.001 (CDF) in 10 *equal* steps of 0.0998 (CDF).

21.13.8 Steps

For a MELCOR interface, the **<MultiRun>** step type will most likely be used. First, the step needs to be named: this name will be one of the names used in the **<sequence>** block. In our example, `Grid_Sampler` and `MC_Sampler`.

```
     <MultiRun name='Grid_Sampler' verbosity='debug'>
```

With this step, we need to import all the files needed for the simulation:

- MELCOR input file
- any other file needed by the calculation (e.g. restart file)

```
    <Input    class='Files' type=''>inputFileMelcor.i</Input>
    <Input    class='Files' type=''>restartFile</Input>
```

We then need to define which model will be used:

```
    <Model   class='Models' type='Code'>MyMELCOR</Model>
```

We then need to specify which Sampler is used, and this can be done as follows:

```xml
    <Sampler class='Samplers' type='Grid'>Grid_Sampler</Sampler>
```

And lastly, we need to specify what kind of output the user wants. For example the user might want to make a database (in RAVEN the database created is an HDF5 file). Here is a classical example:

```xml
    <Output  class='Databases' type='HDF5'>Grid_out</Output>
```

Following is the example of two MultiRun steps which use different sampling methods (Grid and Monte Carlo), and creating two different databases for each one:

```xml
<Steps verbosity='debug'>
  <MultiRun name='Grid_Sampler' verbosity='debug'>
    <Input   class='Files' type=''>inputFileMelcor.i</Input>
    <Input   class='Files' type=''>restartFile</Input>
    <Model   class='Models'    type='Code'>MyMELCOR</Model>
    <Sampler class='Samplers'  type='Grid'>Grid_Sampler</Sampler>
    <Output  class='Databases' type='HDF5'>Grid_out</Output>
    <Output  class='DataObjects' type='PointSet'
       ↪ >GridMelcorPointSet</Output>
    <Output  class='DataObjects'
       ↪ type='HistorySet'>GridMelcorHistorySet</Output>
  </MultiRun>
  <MultiRun name='MC_Sampler' verbosity='debug'
     ↪ re-seeding='210491'>
    <Input   class='Files' type=''>inputFileMelcor.i</Input>
    <Input   class='Files' type=''>restartFile</Input>
    <Model   class='Models'    type='Code'>MyMELCOR</Model>
    <Sampler class='Samplers'
       ↪ type='MonteCarlo'>MC_Sampler</Sampler>
    <Output  class='Databases' type='HDF5'      >MC_out</Output>
    <Output  class='DataObjects' type='PointSet'
       ↪ >MonteCarloMelcorPointSet</Output>
    <Output  class='DataObjects'
       ↪ type='HistorySet'>MonteCarloMelcorHistorySet</Output>
  </MultiRun>
</Steps>
```

21.13.9 Databases

As shown in the **<Steps>** block, the code is creating two database objects called `Grid_out` and `MC_out`. So the user needs to input the following:

```xml
<Databases>
  <HDF5 name="Grid_out" readMode="overwrite"/>
  <HDF5 name="MC_out" readMode="overwrite"/>
</Databases>
```

As listed before, this will create two databases. The files will have names corresponding to their **name** appended with the .h5 extension (i.e. `Grid_out.h5` and `MC_out.h5`).

21.13.10 DataObjects

As shown in the **<Steps>** block, the code is creating 4 data objects (2 HistorySet and 2 PointSet) called `GridMelcorPointSet` `GridMelcorHistorySet` `MonteCarloMelcorPointSet` and `MonteCarloMelcorHistorySet`. So the user needs to input the following block as well, where the Input and Output variables are listed:

```xml
  <DataObjects>
    <PointSet name="GridMelcorPointSet">
      <Input>PRE</Input>
      <Output>
        time,volume_1_PRESSURE,volume_1_TLIQ,
        volume_1_TVAP,volume_1_MASS
      </Output>
    </PointSet>
    <HistorySet name="GridMelcorHistorySet">
      <Input>PRE</Input>
      <Output>
        time,volume_1_PRESSURE,volume_1_TLIQ,
        volume_1_TVAP,volume_1_MASS
      </Output>
    </HistorySet>
    <PointSet name="MonteCarloMelcorPointSet">
      <Input>PRE</Input>
      <Output>
        time,volume_1_PRESSURE,volume_1_TLIQ,
        volume_1_TVAP,volume_1_MASS
      </Output>
    </PointSet>
    <HistorySet name="MonteCarloMelcorHistorySet">
      <Input>PRE</Input>
      <Output>
        time,volume_1_PRESSURE,volume_1_TLIQ,
```

```
      volume_1_TVAP,volume_1_MASS
    </Output>
  </HistorySet>
</DataObjects>
```

As mentioned before, this will create 4 DataObjects.

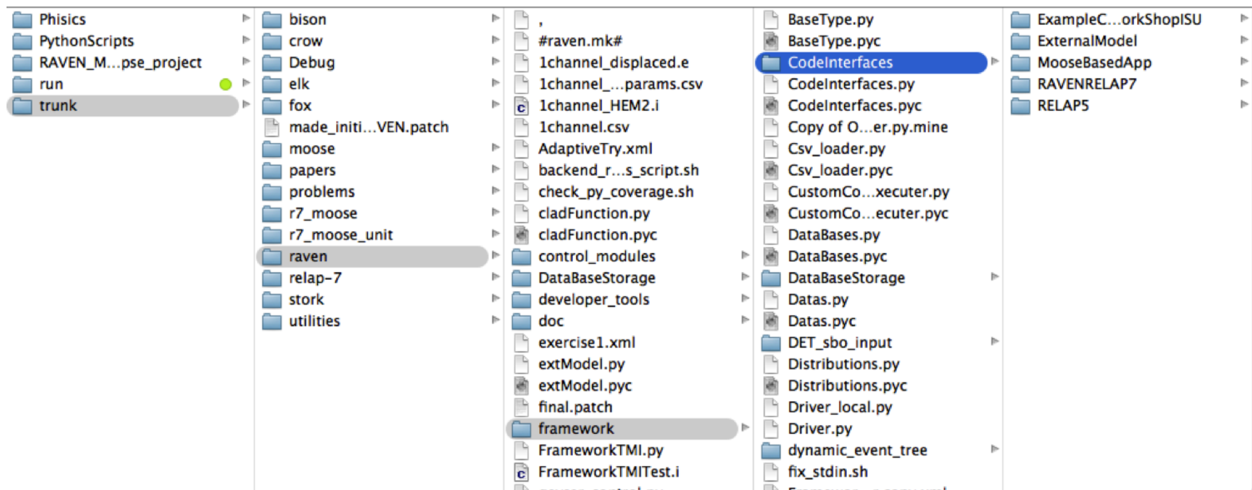


Figure 1. Code Interface Location.

22 Advanced Users: How to couple a new code

The procedure of coupling a new code/application with RAVEN is a straightforward process. For all the codes currently supported by RAVEN (e.g. RELAP-7, RELAP5-3D, BISON, MOOSE,etc.), the coupling is performed through a Python interface that interprets the information coming from RAVEN and translates them into the input of the driven code. The coupling procedure does not require modifying RAVEN itself. Instead, the developer creates a new Python interface that is going to be embedded in RAVEN at run-time (no need to introduce hard-coded coupling statements). This interface needs to be placed in a folder (whatever name) located in (see figure 1):

```
path/to/raven/distribution/raven/framework/CodeInterfaces/
```

At the initialization stage, RAVEN imports all the Interfaces that are contained in this directory and performs some preliminary cross-checks.
It is important to notice that the name of class in the Interface module is the one the user needs to specify when the new interface needs to be used. For example, if the Interface module contains the class “NewCode”, the *subType* in the **<Code>** block will be “NewCode”:

```
class NewCode(CodeInterfaceBase):
  ...
```

```
  ...
  <Code name='whatever' subType='NewCode'>
   ...
  </Code>
```

```
  ...
</Models>
```

In the following sub-sections, a step-by-step procedure for coupling a code to RAVEN is outlined.

22.1 Pre-requisites.

In order to couple a newer application to the RAVEN code, some pre-requisites need to be satisfied.

Input

The first pre-requisite is the knowledge of the input syntax of the application the developer wants to couple. Indeed, RAVEN task "ends" at the Code Interface stage. RAVEN transfers the information needed to perturb the input space into the Code interface and expects that the newly developed Interface is able to perturb the input files based on the information passed through.
This means that the developer needs to code a Python-compatible parser of the system code input (a module that is able to read and modify the input of the code that needs to be coupled).
For example, let's suppose the input syntax of the code the developer needs to couple is as follows:

```
kewword1 =  aValue1
kewword2 =  aValue2
kewword3 =  aValue3
kewword4 =  aValue4
```

The Python input parser would be:

```
class simpleInputParser():
  def __init__(self,filename):
    #
    # @ In, string, filename, input file name (with path)
    #
    self.keywordDictionary = {}
    # open the file
    fileobject = open(filename)
    # store all the lines into a list
    lines = fileobject.readlines()
    # parse the list to construct
    # self.keywordDictionary dictionary
    for line in lines:
      # split the line with respect
      # to the symbol "=" and store the
      # outcomes into the dictionary
```

```
      # listSplitted[0] is the keword
      # listSplitted[1] is the value
      listSplitted = line.split("=")
      keyword = listSplitted[0]
      value   = listSplitted[1]
      self.keywordDictionary[keyword] = value
    # close the file
    fileobject.close()

  def modifyInternalDictionary(self,inDictionary):
      #
      # @ In, dictionary {keyword:value},
      # inDictionary, dictionary containing
      # the keywords to perturb
      #

    # we just parse the dictionary and replace the
    # matching keywords
    for keyword,newvalue in inDictionary.items():
      self.keywordDictionary[keyword] = newvalue

  def writeNewInput(self,filename):
    #
    # @ In, string, filename, newer input file name (with path)
    #

    # open the file
    fileobject = open(filename)
    # write line by line
    for keyword,newvalue in self.keywordDictionary.items():
      fileobject.write(keyword + ``='' + str(newvalue) + ``\n'')
    # close the file
    fileobject.close()
```

It is important to notice that for most of the codes, a wild-card approach can be used. In case this approach fits the user's needs, the RAVEN developer team suggests to inherit from the $GenericCode$ Interface (see section **??**).

Output

RAVEN is able to handle Comma Separated Value (CSV) files (as outputs of the system code). In order make RAVEN able to retrieve the information from the newly coupled code, these files

need to be either generated by the system code itself or the developer needs to code a Python-compatible module to convert the whatever code output format to a CSV one. This module can be directly called in the new code interface (see following section).
Let's suppose that the output format of the code (the same of the previous input parser example) is as follows:

```
result1 = aValue1
result2 = aValue2
result3 = aValue3
```

The Python output converter would be as simple as:

```
def convertOutputFileToCSV(outputfile):
    keywordDictionary = {}
    # open the original file
    fileobject = open(outputfile)
    outputCSVfile = open (outputfile + '.csv')
    # store all the lines into a list
    lines = fileobject.readlines()
    # parse the list to construct
    # self.keywordDictionary dictionary
    for line in lines:
      # split the line with respect
      # to the symbol "=" and store the
      # outcomes into the dictionary
      # listSplitted[0] is the keword
      # listSplitted[1] is the value
      listSplitted = line.split("=")
      keyword = listSplitted[0]
      value   = listSplitted[1]
      keywordDictionary[keyword] = value
    outputCSVfile.write(','.join(keywordDictionary.keys()))
    outputCSVfile.write(','.join(keywordDictionary.values()))
    outputCSVfile.close()
```

And the output CSV becomes:

```
result1, result2, result3
aValue1, aValue2, aValue3
```

22.2 Code Interface Creation

As already mentioned, RAVEN imports all the "Code Interfaces" at run-time, without actually knowing the syntax of the driven codes. In order to make RAVEN able to drive a newer software, the developer needs to code a Python module that will contain few methods (with strict syntax) that are called by RAVEN during the simulation.
When loading a "Code Interface", RAVEN expects to find, in the class representing the code, the following required methods:

```
from CodeInterfaceBaseClass import CodeInterfaceBase
class NewCode(CodeInterfaceBase):
  def generateCommand(self, inputFiles, executable, clargs=None,
     ↪ fargs=None)
  def createNewInput(self, currentInputFiles, oriInputFiles,
                                samplerType, **Kwargs)
```

In addition, the following optional methods can be specified:

```
from CodeInterfaceBaseClass import CodeInterfaceBase
class NewCode(CodeInterfaceBase):
  ...
  def finalizeCodeOutput(self, command, output, workingDir)
  def getInputExtension(self)
  def checkForOutputFailure(self, output, workingDir)
```

In the following sub-sections all the methods are fully explained, providing examples (referring to the simple code used as example for the previous sections)

22.2.1 Method: `generateCommand`

```
def generateCommand(self, inputFiles, executable, clargs=None,
   ↪ fargs=None)
```

The **generateCommand** method is used to generate the commands (in `string` format) needed to launch the driven Code, as well as the root name of the output of the perturbed inputs (in `string` format). The return for this command is a two-part `Python` tuple. The first entry is a list of two-part tuples, each which specifies whether the corresponding command should be run exclusively in serial, or whether it can be run in parallel, as well as the command itself. For example, for a command where two successive commands are called, the first in serial and the second in parallel,

```
  def generateCommand(self,inputFiles,executable,clargs=None,
     ↪ fargs=None):
```

```
  . . .
  commands = [('serial',first_command), ('parallel',
    ↪ second_command)]
  return (commmands,outFileRoot)
```

For each command, the second entry in the tuple is a string containing the full command that the internal JobHandler is going to use to run the Code this interface refers to. The return data type must be a Python `tuple` with a list of tuples and a string: (`commands, outfileRoot`). Note that in most cases, only a single command needs to be run, so only a single command tuple is necessary. At run time, RAVEN will string together commands attached by double ampersands (&&), and each command labeled as parallel-compatible will be prepended with appropriate mpi arguments. For the example above, the command executed will be (with **<NumMPI>** equal to 4)

```
$  first_command && mpiexec -n 4 second_command
```

RAVEN is going to call the `generateCommand` function passing in the following arguments:

- **`inputFiles`**, data type = list: List of input files (length of the list depends on the number of inputs listed in the Step which is running this code);
- **`executable`**, data type = string, executable name with absolute path (e.g. /home/path_to_executable/code.exe);
- **`clargs`**, *optional*, data type = dictionary, a dictionary containing the command-line flags the user can specify in the input (e.g. under the node $< Code >< clargstype =' input' arg =' -i' extension =' .inp' / >< /Code >$).
- **`fargs`**, *optional*, data type = dictionary, a dictionary containing the axuiliary input file variables the user can specify in the input (e.g. under the node $< Code >< clargstype =' input' arg =' aux' extension =' .aux' / >< /Code >$).

For the example referred to in the previous section, this method would be implemented as follows:

```
def generateCommand(self,inputFiles,executable,clargs=None,
   ↪ fargs=None):
  found = False
  for index, inputFile in enumerate(inputFiles):
    if inputFile.endswith(self.getInputExtension()):
      found = True
      break
  if not found: raise IOError(
        `None of the input files has one of the following
          ↪ extensions: ` +
```

```
        ` `.join(self.getInputExtension()))
    outputfile = 'out~'+os.path.split(inputFiles[index])[1].split
       ↪ ('.')[0]
    executeCommand = [('parallel',executable+ ` -i ` +os.path.
       ↪ split(inputFiles[index])[1])]
    return executeCommand,outputfile
```

22.2.2 Method: `createNewInput`

```
def createNewInput(self,currentInputFiles,oriInputFiles,
   ↪ samplerType,**Kwargs)
```

The **createNewInput** method is used to generate an input based on the information RAVEN passes in. In this function the developer needs to call the driven code input parser in order to modify the input file, accordingly with respect to the variables RAVEN is providing. This method needs to return a list containing the path and filenames of the modified input files. **Note:** RAVEN expects that at least one input file of the original list gets modified.

RAVEN is going to call this function passing in the following arguments:

- **`currentInputFiles`**, data type = list: List of current input files. This list of files is the one the code interface needs to use to print the new perturbed list of files. Indeed, RAVEN already changes the file location in sub-directories and the Code Interface does not need to change the filename or location of the files. For example, the files are going to have a absolute path as following: $.\\path_to_working_directory\\stepName\\anUniqueIdentifier\\filename.extension$. In case of sampling, the "*anUniqueIdentifier*" is going to be an integer (e.g. 1).
- **`oriInputFiles`** , data type = list, List of the original input files;
- **`samplerType`** , data type = string, Sampler type (e.g. MonteCarlo, Adaptive, etc.). **Note:** None if no Sampler has been used;
- **`Kwargs`** , data type = kwarded dictionary, dictionary of parameters. In this dictionary there is another dictionary called "SampledVars" where RAVEN stores the variables that got sampled (Kwargs['SampledVars'] = {'var1':10,'var2':40});

For the example referred in the previous section, this method would implemented as follows:

```
  def createNewInput(self,currentInputFiles,
        oriInputFiles,samplerType,**Kwargs):
    for index, inputFile in enumerate(oriInputFiles):
```

```
      if inputFile.endswith(self.getInputExtension()):
        break
    parser = simpleInputParser(currentInputFiles[index])
    parser.modifyInternalDictionary(**Kwargs['SampledVars'])
    parser.writeNewInput(newInputFiles[index])
    return newInputFiles
```

22.2.3 Method: `getInputExtension`

```
def getInputExtension(self)
```

The **getInputExtension** function is an optional method. If present, it is called by RAVEN code at run time. This function can be considered an utility method, since its main goal is to return a tuple of strings, where the developer can place all the input extensions the code interface needs to support (i.e. the extensions of the input(s) the code interface is going to "perturb"). If this method is not implemented, the default extensions are **(".i", ".inp", ".in")**. This function does not accept any input argument. For the example referred in the previous section, this method would implemented as follows:

```
def getInputExtension(self):
    return (".i",".input")
```

22.2.4 Method: `finalizeCodeOutput`

```
def finalizeCodeOutput(self, command, output, workingDir)
```

The **finalizeCodeOutput** function is an optional method. If present, it is called by RAVEN code at the end of each run. It can be used for those codes, that do not create CSV files as output to convert the whatever output format into a CSV. RAVEN checks if a string is returned; if so, RAVEN interprets that string as the new output file name (CSV).
RAVEN is going to call this function passing in the following arguments:

- **`command`**, data type = string: the command used to run the just ended job;
- **`output`**, data type = string, the Output name root;
- **`workingDir`**, data type = string, current working directory.

For the example referred in the previous section, this method would implemented as follows:

```
def finalizeCodeOutput(self, command, output, workingDir):
    outfile = os.path.join(workingDir,output+".o")
    convertOutputFileToCSV(outfile)
```

22.2.5 Method: **checkForOutputFailure**

```
def checkForOutputFailure(self, output, workingDir)
```

The **checkForOutputFailure** function is an optional method. If present, it is called by RAVEN code at the end of each run. This method needs to be implemented by the codes that, if a run fails, return a "returncode" = 0. This can happen in those codes that record the failure of a run (e.g. not converged, etc.) as normal termination (returncode == 0) This method can be used, for example, to parse the outputfile looking for a special keyword that testifies that a particular job failed (e.g. in RELAP5 would be the keyword "********"). This method MUST return a boolean (True if failed, False otherwise).
RAVEN is going to call this function passing in the following arguments:

- **`output`**, data type = string,the Output name root;
- **`workingDir`**, data type = string, current working directory.

For the example referred in the previous section, this method would implemented as follows:

```
def checkForOutputFailure(self, command, output, workingDir):
    from  __builtin__ import any
    errorWord = "ERROR"
    return any(errorWord in x for x in
      open(os.path.join(workingDir,output+'.o'),"r").readlines())
```

22.3 Tools for Developing Code Interfaces

To make generating a code interface as simple as possible, there are several tools RAVEN makes available within the Code Interface objects.

22.3.1 File Objects

RAVEN has created a wrapper for files within Python in order to carry along some additional information. This allows the user to tag particular files for reference in the Code Interface, using the **type** XML attribute in **<Files>** nodes. To differentiate, RAVEN file objects will use the capital Files, whereas typical files will use the lowercase files.

When the Files are passed in to `createNewInput`, they are passed in as Files objects. To access the xmlAttrtype of a file, use the method `getType`. For instance, instead of looking for an extension, a Code Interface might identify an input file by looking for a particular type, as shown in the example below. **Note:** RAVEN does not access a File's **type**; it is exclusively an optional tool for Code Interface developers.

```
found = False
for inFile in inputFiles:
  if inFile.getType()=='mainInput':
    found = True
    break
if not found:
  raise IOError('Desired file with type ``mainInput'' not found!'
    ↪ )
```

Using Files **type** attributes can especially help when multiple input files have the same extension. For example, say a Code execution command normally has the following appearance on the command line:

```
/home/path/to/executable/myexec.sh -i mainInp.xml -a auxInp.xml
  ↪ --mesh cube.e
```

The **<Files>** block in the RAVEN XML might appear as follows:

```
<Files>
  <Input name='main' type='base'>mainInp.xml</Input>
  <Input name='two'  type='aux' >auxInp.xml</Input>
  <Input name='cube' type='mesh' perturbable='False'>cube.e</
    ↪ Input>
</Files>
```

The search for these files in the Code Interface might then look like the example below, assuming one file per type:

```
# populate a type dictionary
typesDict={}
for inFile in inputFiles:
```

```
  typesDict[inFile.getType()]=inFile
# check all the necessary files are there
if 'base' not in typesDict.keys():
  raise IOError('File type ``base'' not listed in input file!')
if 'aux' not in typesDict.keys():
  raise IOError('File type ``aux'' not listed in input file!')
if 'mesh' not in typesDict.keys():
  raise IOError('File type ``mesh'' not listed in input file!')
mainFile = typesDict['base']
# do operations on file, etc.
```

Additionally, a Code Interface developer can access the **perturbable** through the `getPerturbable()` method of a Files object. This can be useful, for example, in preventing searching binary files for variable names when creating new input. For example,

```
for inFile in inputFiles:
  if not inFile.getPerturbable(): continue
  # etc
```

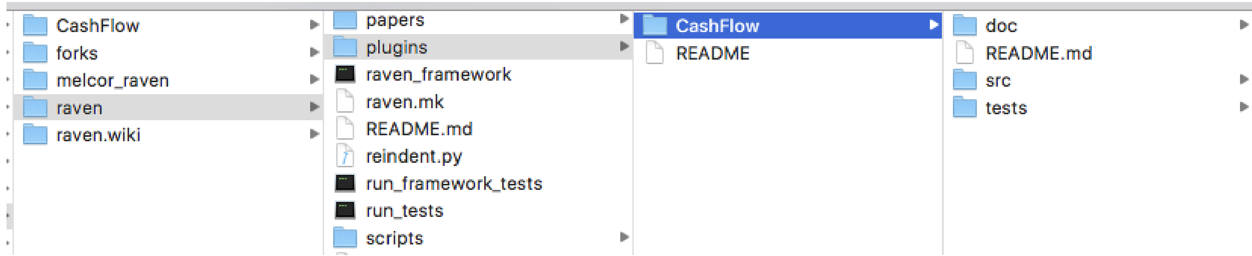


Figure 2. Plugins Location

23 Advanced Users: How to create a RAVEN ExternalModel plugin

The procedure of adding a plugin for the ExternalModel is a straightforward process. The addition of a plugin does not require modifying RAVEN itself. Instead, the developer creates a new Python module that is going to be embedded in RAVEN at run-time (no need to introduce hard-coded statements). This plugin needs to be placed in a folder (whatever name) located in (see figure 2):

```
path/to/raven/plugins/
```

In order to install a new plugin, the user can run the script contained in the RAVEN script folder:

```
python path/to/raven/scripts/install_plugins.py  **directory**
```

where $**directory**$ should be replaced with the absolute path to the plugin directory. (e.g. "path/to/my/plugins/folder"). If the plugin developer wants to make of his plugin an official supported plugin in RAVEN (by the submodule system), he needs to check the raven wiki under the "contribution" section).

At the initialization stage, RAVEN imports all the Plugins that are contained in this directory and performs some preliminary cross-checks.
It is important to notice that the name of class in the Plugin module is the one the user needs to specify when the new plugin needs to be used. For example, if the Plugin module contains the class "NewPlugin", the *subType* in the **<ExternalModel>** block will be "NewPlugin":

```
class NewPlugin(ExternalModelPluginBase):
  ...
```

```
  ...
  <ExternalModel name='whatever' subType='NewPlugin'>
   ...
```

```
  </ExternalModel>
  ...
</Models>
```

In the following sub-sections, a step-by-step procedure for creating a new ExternalModel plugin is outlined.

23.1 ExternalModel Plugin Input

When a new ExternalModel plugin is developed, its RAVEN input is almost identical to the general ExternalModel entity (see 17.4). The specifications of an ExternalModel Plugin must be defined within the XML block **<ExternalModel>**. This XML node needs to contain the attributes:

- **name**, ***required string attribute***, user-defined name of this External Model. **Note:** As with the other objects, this is the name that can be used to refer to this specific entity from other input blocks in the XML.
- **subType**, ***required string attribute***, must be equal to the name of the new plugin the user wants to use (e.g. $NewPlugin$). **Note:** In case a plugin is requested (through the **subType** attribute) the attribute **ModuleToLoad** must not be inputted.

In order to make the RAVEN code aware of the variables the user is going to manipulate/use in her/his ExternalModel Plugin, the variables need to be specified in the **<ExternalModel>** input block. The user needs to input, within this block, only the variables that RAVEN needs to be aware of (i.e. the variables are going to directly be used by the Plugin) and not the local variables that the ExternalModel Plugin developer does not want to, for example, store in a RAVEN internal object. These variables are specified within a **<variables>** block:

- **<variables>**, ***string, required parameter***. Comma-separated list of variable names. Each variable name needs to match a variable used/defined in the external python model.

In addition, if the user wants to use the alias system, the following XML block can be inputted:

- **<alias>** ***string, optional field*** specifies alias for any variable of interest in the input or output space for the ExternalModel. These aliases can be used anywhere in the RAVEN input to refer to the ExternalModel variables. In the body of this node the user specifies the name of the variable that the model is going to use (during its execution). The actual alias, usable throughout the RAVEN input, is instead defined in the **variable** attribute of this tag.
  The user can specify aliases for both the input and the output space. As sanity check, RAVEN requires an additional required attribute **type**. This attribute can be either "input" or "output". **Note:** The user can specify as many aliases as needed.
  Default: None

When the Plugin variables are defined, at run time, RAVEN initializes them and tracks their values during the simulation. Each variable defined in the **<ExternalModel>** block is available in the Plugin class (in each implemented method) as the object "container" that "acts" as a Python "self". For example,

```
def run (self, container, inputs):
  print(container.variableA)
```

23.2 ExternalModel Plugin Creation

As already mentioned, RAVEN imports all the "ExternalModel Plugins" at run-time. In order to make RAVEN able to drive a newer ExternalModel plugin, the developer needs to code a Python class containing few methods (with strict syntax) that are called by RAVEN during the simulation. Every new "ExternalModel Plugin" must inherit from a RAVEN base class named $ExternalModelPluginBase$:

```
class NewPlugin(ExternalModelPluginBase):
  ...
```

This base class is needed by RAVEN to identify in the plugins folder which class must be considered an "ExternalModel Plugin".
In addition, when loading an "ExternalModel Plugin", RAVEN expects to find, in the class representing the plugin, the following required methods:

```
from ExternalModelPluginBase import ExternalModelPluginBase
class NewPlugin(ExternalModelPluginBase):
  def run (self, container, Inputs)
```

In addition, the following optional methods can be specified:

```
from ExternalModelPluginBase import ExternalModelPluginBase
class NewPlugin(ExternalModelPluginBase):
  ...
  def createNewInput(self, container, inputs, samplerType, **
    ↪ Kwargs)
  def _readMoreXML(self, container, xmlNode)
  def initialize(self,container, runInfo, inputs)
```

In the following sub-sections all the methods are fully explained, providing examples.

23.2.1 Method: `run`

```
def run (self, container, Inputs)
```

As stated previously, the only method that *must* be present in an ExternalModel Plugin is the **run** function. In this function, the plugin developer needs to implement the algorithm that RAVEN will execute. The **run** method is generally called after having inquired the "createNewInput" method (either the internal RAVEN one or the one implemented by the plugin developer). The only two attributes this method is going to receive are a Python list of inputs (the inputs coming from the `createNewInput` method) and a "self-like" object named "container". If the user wants RAVEN to collect the results of this method, the outcomes of interest need to be stored in the above mentioned "container" object. **Note:** RAVEN is trying to collect the values of the variables listed only in the **<ExternalModel>** XML block. In the following an example is reported:

```
def run(self, container, Input):
  # in here the actual run of the
  # model is implemented
  input = Input[0]
  container.outcome = container.sigma*container.rho*input[``
    ↪ whatEver'']
```

23.2.2 Method: `createNewInput`

```
def createNewInput(self, container, inputs, samplerType, **Kwargs
  ↪ )
```

The **createNewInput** method can be implemented by the ExternalModel Plugin developer to create a new input with the information coming from the RAVEN framework. In this function, the developer can retrieve the information coming from the RAVEN framework, during the employment of a calculation flow, and use them to construct a new input that is going to be transferred to the "run" method. The new input created needs to be returned to RAVEN (i.e. "return NewInput"). This method expects that the new input is returned in a Python "dictionary". RAVEN communicates, thorough a set of method attributes, all the information that are generally needed to create a new input:

- `inputs`, ***python list***, a list of all the inputs that have been defined in the "Step" using this model.
- `samplerType`, ***string***, the type of Sampler, if a sampling strategy is employed; will be None otherwise.

- Kwargs, ***dictionary***, a dictionary containing several pieces of information (that can change based on the "Step" type). If a sampling strategy is employed, this dictionary contains another dictionary identified by the keyword "SampledVars", in which the variables perturbed by the sampler are reported.

Note: If the "Step" that is using this Model has as input(s) an object of main class type "DataObjects" (see Section 14), the internal "createNewInput" method is going to convert it in a dictionary of values. Here we present an example:

```
def createNewInput(self, container, inputs,samplerType,**Kwargs):
  # in here the actual createNewInput of the
  # model is implemented
  if samplerType == 'MonteCarlo':
    avariable = inputs['something']*inputs['something2']
  else:
    avariable = inputs['something']/inputs['something2']
  return avariable*Kwargs['SampledVars']['aSampledVar']
```

23.2.3 Method: _readMoreXML

```
def _readMoreXML(self, container, xmlNode)
```

As already mentioned, the **_readMoreXML** method can be implemented by the ExternalModel Plugin developer if the XML input that belongs to this ExternalModel plugin needs to be extended to contain other information. The read information needs to be stored in the "self-like" object "container" in order to be available to all the other methods (e.g. if the developer needs to add a couple of newer XML nodes with information needed by the algorithm implemented in the "run" method). If this method is implemented in the **ExternalModel**, RAVEN is going to call it when the node **<ExternalModel>** is found parsing the XML input file. The method receives from RAVEN an attribute of type "xml.etree.ElementTree", containing all the sub-nodes and attribute of the XML block **<ExternalModel>**.

Example XML:

```
<Simulation>
  ...
  <Models>
     ...
    <ExternalModel name='AnExtModule' subType=''NewPlugin">
␣␣␣␣␣␣␣<variables>sigma,rho,outcome</variables>
␣␣␣␣␣␣␣<!--
```

```
          here we define other XML nodes RAVEN does not read 
   ↪ automatically.
          We need to implement, in the external model Plugin 
   ↪ class the _readMoreXML
          method
        -->
        <newNodeWeNeedToRead>
          whatNeedsToBeRead
        </newNodeWeNeedToRead>
    </ExternalModel>
     ...
  </Models>
  ...
</Simulation>
```

Corresponding Python function:

```python
def _readMoreXML(self, container, xmlNode):
  # the xmlNode is passed in by RAVEN framework
  # <newNodeWeNeedToRead> is unknown (in the RAVEN framework)
  # we have to read it on our own
  # get the node
  ourNode = xmlNode.find('newNodeWeNeedToRead')
  # get the information in the node
  container.ourNewVariable = ourNode.text
  # end function
```

23.2.4 Method: initialize

```python
def initialize(self, container, runInfo, inputs)
```

The **initialize** method can be implemented in the **ExternalModel** Plugin in order to initialize some variables needed by it. For example, it can be used to compute a quantity needed by the "run" method before performing the actual calculation. If this method is implemented in the **ExternalModel** Plugin, RAVEN is going to call it at the initialization stage of each "Step" (see section 20). RAVEN will communicate, thorough a set of method attributes, all the information that are generally needed to perform an initialization:

- runInfo, a dictionary containing information regarding how the calculation is set up (e.g. number of processors, etc.). It contains the following attributes:

- `DefaultInputFile` – default input file to use
- `SimulationFiles` – the xml input file
- `ScriptDir` – the location of the pbs script interfaces
- `FrameworkDir` – the directory where the framework is located
- `WorkingDir` – the directory where the framework should be running
- `TempWorkingDir` – the temporary directory where a simulation step is run
- `NumMPI` – the number of mpi process by run
- `NumThreads` – number of threads by run
- `numProcByRun` – total number of core used by one run (number of threads by number of mpi)
- `batchSize` – number of contemporaneous runs
- `ParallelCommand` – the command that should be used to submit jobs in parallel (mpi)
- `numNode` – number of nodes
- `procByNode` – number of processors by node
- `totalNumCoresUsed` – total number of cores used by driver
- `queueingSoftware` – queueing software name
- `stepName` – the name of the step currently running
- `precommand` – added to the front of the command that is run
- `postcommand` – added after the command that is run
- `delSucLogFiles` – if a simulation (code run) has not failed, delete the relative log file (if True)
- `deleteOutExtension` – if a simulation (code run) has not failed, delete the relative output files with the listed extension (comma separated list, for example: 'e,r,txt')
- `mode` – running mode, curently the only mode supported is mpi (but custom modes can be created)
- *expectedTime* – how long the complete input is expected to run
- *logfileBuffer* – logfile buffer size in bytes

- inputs, a list of all the inputs that have been specified in the "Step" using this model.

As all the others method in the ExternalModel Plugin, the information *must* be stored in the "self-like" object "container". In the following an example is reported:

```
def initialize(self, container, runInfo, inputs):
 # Let's suppose we just need to initialize some variables
  container.sigma = 10.0
  container.rho   = 28.0
  # end function
```

A Appendix: Example Primer

In this Appendix, a set of examples are reported. In order to be as general as possible, the *Model* type "ExternalModel" has been used.

A.1 Example 1.

This simple example is about the construction of a "Lorentz attractor", sampling the relative input space. The parameters that are sampled represent the initial coordinate (x0,y0,z0) of the attractor origin.

```
<?xml version="1.0" encoding="UTF-8"?>
<Simulation verbosity="debug">
<!-- RUNINFO -->
<RunInfo>
    <WorkingDir>externalModel</WorkingDir>
    <Sequence>FirstMRun</Sequence>
    <batchSize>3</batchSize>
</RunInfo>
<!-- Files -->
<Files>
    <Input name='lorentzAttractor.py'
       ↪ type=''>lorentzAttractor</Input>
</Files>
<!-- STEPS -->
<Steps>
    <MultiRun name='FirstMRun'  re-seeding='25061978'>
        <Input   class='Files'     type=''
           ↪ >lorentzAttractor.py</Input>
        <Model   class='Models'    type='ExternalModel'
           ↪ >PythonModule</Model>
        <Sampler class='Samplers'  type='MonteCarlo'
           ↪ >MC_external</Sampler>
        <Output  class='DataObjects'     type='HistorySet'
           ↪ >testPrintHistorySet</Output>
        <Output  class='Databases' type='HDF5'
           ↪ >test_external_db</Output>
        <Output  class='OutStreams' type='Print'
           ↪ >testPrintHistorySet_dump</Output>
    </MultiRun >
</Steps>
```

```
<!-- MODELS -->
<Models>
    <ExternalModel name='PythonModule' subType=''
       ↪ ModuleToLoad='externalModel/lorentzAttractor'>
       <variables>sigma,rho,beta,x,y,z,time,x0,y0,z0</variables>
    </ExternalModel>
</Models>
<!-- DISTRIBUTIONS -->
<Distributions>
    <Normal name='x0_distrib'>
        <mean>4</mean>
        <sigma>1</sigma>
    </Normal>
    <Normal name='y0_distrib'>
        <mean>4</mean>
        <sigma>1</sigma>
    </Normal>
    <Normal name='z0_distrib'>
        <mean>4</mean>
        <sigma>1</sigma>
    </Normal>
</Distributions>
<!-- SAMPLERS -->
<Samplers>
    <MonteCarlo name='MC_external'>
      <samplerInit>
        <limit>3</limit>
      </samplerInit>
      <variable name='x0' >
        <distribution  >x0_distrib</distribution>
      </variable>
      <variable name='y0' >
        <distribution  >y0_distrib</distribution>
      </variable>
      <variable name='z0' >
        <distribution  >z0_distrib</distribution>
      </variable>
    </MonteCarlo>
</Samplers>
<!-- DATABASES -->
<Databases>
  <HDF5 name="test_external_db"/>
```

```xml
</Databases>
<!-- OUTSTREAMS -->
<OutStreams>
  <Print name='testPrintHistorySet_dump'>
    <type>csv</type>
    <source>testPrintHistorySet</source>
  </Print>
</OutStreams>
<!-- DATA OBJECTS -->
<DataObjects>
    <HistorySet name='testPrintHistorySet'>
        <Input>x0,y0,z0</Input>
        <Output>time,x,y,z</Output>
   </HistorySet>
</DataObjects>
</Simulation>
```

The Python *ExternalModel* is reported below:

```python
import numpy as np

def run(self,Input):
  max_time = 0.03
  t_step = 0.01

  numberTimeSteps = int(max_time/t_step)

  self.x = np.zeros(numberTimeSteps)
  self.y = np.zeros(numberTimeSteps)
  self.z = np.zeros(numberTimeSteps)
  self.time = np.zeros(numberTimeSteps)

  self.x0 = Input['x0']
  self.y0 = Input['y0']
  self.z0 = Input['z0']

  self.x[0] = Input['x0']
  self.y[0] = Input['y0']
  self.z[0] = Input['z0']
  self.time[0]= 0

  for t in range (numberTimeSteps-1):
    self.time[t+1] = self.time[t] + t_step
```

```
    self.x[t+1]     = self.x[t] +  self.sigma*
                       (self.y[t]-self.x[t]) * t_step
    self.y[t+1]     = self.y[t] + (self.x[t]*
                       (self.rho-self.z[t])-self.y[t]) * t_step
    self.z[t+1]     = self.z[t] + (self.x[t]*
                            self.y[t]-self.beta*self.z[t]) * t_step
```

A.2 Example 2.

This example shows a slightly more complicated example, that employs the usage of:

- *Samplers:* Grid and Adaptive;
- *Models:* External, Reduce Order Models and Post-Processors;
- *OutStreams:* Prints and Plots;
- *Data Objects:* PointSets;
- *Functions:* ExternalFunctions.

The goal of this input is to compute the "SafestPoint". It provides the coordinates of the farthest point from the limit surface that is given as an input. The safest point coordinates are expected values of the coordinates of the farthest points from the limit surface in the space of the "controllable" variables based on the probability distributions of the "non-controllable" variables.

The term "controllable" identifies those variables that are under control during the system operation, while the "non-controllable" variables are stochastic parameters affecting the system behaviour randomly.

The "SafestPoint" post-processor requires the set of points belonging to the limit surface, which must be given as an input.

```
<Simulation verbosity='debug'>

<!-- RUNINFO -->
<RunInfo>
  <WorkingDir>SafestPointPP</WorkingDir>
  <Sequence>pth1,pth2,pth3,pth4</Sequence>
  <batchSize>50</batchSize>
</RunInfo>

```

```
<!-- STEPS -->
<Steps>
  <MultiRun name = 'pth1' pauseAtEnd = 'False'>
    <Sampler  class = 'Samplers'  type = 'Grid'
       ↪ >grd_vl_ql_smp_dpt</Sampler>
    <Input    class = 'DataObjects'     type = 'PointSet'
       ↪ >grd_vl_ql_smp_dpt_dt</Input>
    <Model    class = 'Models'    type = 'ExternalModel'
       ↪ >xtr_mdl</Model>
    <Output   class = 'DataObjects'     type = 'PointSet'
       ↪ >nt_phy_dpt_dt</Output>
  </MultiRun >

  <MultiRun name = 'pth2' pauseAtEnd = 'True'>
    <Sampler          class = 'Samplers'  type = 'Adaptive'
       ↪ >dpt_smp</Sampler>
    <Input            class = 'DataObjects'     type =
       ↪ 'PointSet'  >bln_smp_dt</Input>
    <Model            class = 'Models'    type = 'ExternalModel'
       ↪ >xtr_mdl</Model>
    <Output           class = 'DataObjects'     type =
       ↪ 'PointSet'  >nt_phy_dpt_dt</Output>
    <SolutionExport   class = 'DataObjects'     type =
       ↪ 'PointSet'  >lmt_srf_dt</SolutionExport>
  </MultiRun>

  <PostProcess name='pth3' pauseAtEnd = 'False'>
    <Input    class = 'DataObjects'          type = 'PointSet'
       ↪      >lmt_srf_dt</Input>
    <Model    class = 'Models'         type = 'PostProcessor'
       ↪ >SP</Model>
    <Output   class = 'DataObjects'          type = 'PointSet'
       ↪     >sfs_pnt_dt</Output>
  </PostProcess>

  <OutStreamStep name = 'pth4' pauseAtEnd = 'True'>
        <Input   class = 'DataObjects'           type =
           ↪ 'PointSet'  >lmt_srf_dt</Input>
        <Output class = 'OutStreams' type = 'Print'
           ↪ >lmt_srf_dmp</Output>
    <Input   class = 'DataObjects'           type = 'PointSet'
       ↪ >sfs_pnt_dt</Input>
```

```
        <Output class = 'OutStreams' type = 'Print'
           ↪ >sfs_pnt_dmp</Output>
  </OutStreamStep>
</Steps>

<!-- DATA OBJECTS -->
<DataObjects>
  <PointSet name = 'grd_vl_ql_smp_dpt_dt'>
    <Input>x1,x2,gammay</Input>
    <Output>OutputPlaceHolder</Output>
  </PointSet>

  <PointSet name = 'nt_phy_dpt_dt'>
    <Input>x1,x2,gammay</Input>
    <Output>g</Output>
  </PointSet>

  <PointSet name = 'bln_smp_dt'>
    <Input>x1,x2,gammay</Input>
    <Output>OutputPlaceHolder</Output>
  </PointSet>

  <PointSet name = 'lmt_srf_dt'>
    <Input>x1,x2,gammay</Input>
    <Output>g_zr</Output>
  </PointSet>

  <PointSet name = 'sfs_pnt_dt'>
    <Input>x1,x2,gammay</Input>
    <Output>p</Output>
  </PointSet>
</DataObjects>

<!-- DISTRIBUTIONS -->
<Distributions>
  <Normal name = 'x1_dst'>
    <upperBound>10</upperBound>
    <lowerBound>-10</lowerBound>
        <mean>0.5</mean>
    <sigma>0.1</sigma>
  </Normal>
```

```
  <Normal name = 'x2_dst'>
    <upperBound>10</upperBound>
    <lowerBound>-10</lowerBound>
    <mean>-0.15</mean>
    <sigma>0.05</sigma>
  </Normal>

  <Normal name = 'gammay_dst'>
    <upperBound>20</upperBound>
    <lowerBound>-20</lowerBound>
    <mean>0</mean>
    <sigma>15</sigma>
  </Normal>
</Distributions>

<!-- SAMPLERS -->
<Samplers>
  <Grid name = 'grd_vl_ql_smp_dpt'>
    <variable name = 'x1' >
      <distribution>x1_dst</distribution>
      <grid type = 'value' construction = 'equal' steps = '10'
         ↪ upperBound = '10'>2</grid>
    </variable>
    <variable name='x2' >
      <distribution>x2_dst</distribution>
      <grid type = 'value' construction = 'equal' steps = '10'
         ↪ upperBound = '10'>2</grid>
    </variable>
    <variable name='gammay' >
      <distribution>gammay_dst</distribution>
      <grid type = 'value' construction = 'equal' steps = '10'
         ↪ lowerBound = '-20'>4</grid>
    </variable>
  </Grid>

  <Adaptive name = 'dpt_smp' verbosity='debug'>
    <ROM              class = 'Models'    type = 'ROM'
       ↪ >accelerated_ROM</ROM>
    <Function         class = 'Functions' type = 'External'
       ↪ >g_zr</Function>
    <TargetEvaluation class = 'DataObjects'     type =
       ↪ 'PointSet'  >nt_phy_dpt_dt</TargetEvaluation>
```

```
    <Convergence limit = '3000' forceIteration = 'False' weight
       ↪ = 'none' persistence = '5'>1e-2</Convergence>
      <variable name = 'x1'>
        <distribution>x1_dst</distribution>
      </variable>
      <variable name = 'x2'>
        <distribution>x2_dst</distribution>
      </variable>
      <variable name = 'gammay'>
        <distribution>gammay_dst</distribution>
      </variable>
  </Adaptive>
</Samplers>

<!-- MODELS -->
<Models>
  <ExternalModel name = 'xtr_mdl' subType = '' ModuleToLoad =
     ↪ 'SafestPointPP/safest_point_test_xtr_mdl'>
    <variables>x1,x2,gammay,g</variables>
  </ExternalModel>

  <ROM name = 'accelerated_ROM' subType = 'SciKitLearn'>
    <Features>x1,x2,gammay</Features>
    <Target>g_zr</Target>
    <SKLtype>svm|SVC</SKLtype>
    <kernel>rbf</kernel>
    <gamma>10</gamma>
    <tol>1e-5</tol>
    <C>50</C>
  </ROM>

  <PostProcessor name='SP' subType='SafestPoint'>
    <!-- List of Objects (external with respect to this PP)
       ↪ needed by this post-processor -->
    <Distribution     class = 'Distributions'  type =
       ↪ 'Normal'>x1_dst</Distribution>
    <Distribution     class = 'Distributions'  type =
       ↪ 'Normal'>x2_dst</Distribution>
    <Distribution     class = 'Distributions'  type =
       ↪ 'Normal'>gammay_dst</Distribution>
    <!- end of the list -->
    <controllable>
```

```
        <variable name = 'x1'>
                <distribution>x1_dst</distribution>
                <grid type = 'value' steps = '20'>1</grid>
        </variable>
        <variable name = 'x2'>
                <distribution>x2_dst</distribution>
                <grid type = 'value' steps = '20'>1</grid>
        </variable>
    </controllable>
    <non-controllable>
        <variable name = 'gammay'>
                <distribution>gammay_dst</distribution>
                <grid type = 'value' steps = '20'>2</grid>
        </variable>
    </non-controllable>
  </PostProcessor>
</Models>

<!-- FUNCTIONS -->
<Functions>
  <External name='g_zr'
     ↪ file='SafestPointPP/safest_point_test_g_zr.py'>
    <variable>g</variable>
  </External>
</Functions>

<!-- OUT-STREAMS -->
<OutStreams>
  <Print name = 'lmt_srf_dmp'>
        <type>csv</type>
        <source>lmt_srf_dt</source>
  </Print>

  <Print name = 'sfs_pnt_dmp'>
        <type>csv</type>
        <source>sfs_pnt_dt</source>
  </Print>
</OutStreams>

</Simulation>
```

The Python *ExternalModel* is reported below:

```
def run(self,Input):
  self.g = self.x1+4*self.x2-self.gammay
```

The “Goal Function”,the function that defines the transitions with respect the input space coordinates, is as follows:

```
def __residuumSign(self):
  if self.g<0 : return  1
  else        : return -1
```

Document Version Information

tag_number_23-180-gb3fa3b93a
b3fa3b93a2ce4b34e6c0c319d5dc0b5f1cca8922 Paul Talbot
Thu, 15 Mar 2018 13:21:20 -0600

References

[1] M. P. Forum, "Mpi: A message-passing interface standard," tech. rep., Knoxville, TN, USA, 1994.

[2] "Portable batch system." http://www.pbsworks.com.

[3] J. C. Spall, "Implementation of the simultaneous perturbation algorithm for stochastic optimization," *IEEE Transactions on aerospace and electronic systems*, vol. 34, no. 3, pp. 817–823, 1998.

[4] J. D. Hunter, "Matplotlib: A 2d graphics environment," *Computing In Science & Engineering*, vol. 9, no. 3, pp. 90–95, 2007.

[5] D. Cournapeau, "Scikit-learn library for machine learning." http://scikit-learn.org/stable/user_guide.html.

[6] S. Wilcox and W. Marion, *Users manual for TMY3 data sets*. National Renewable Energy Laboratory Golden, CO, 2008.

[7] J. M. Finkelstein and R. E. Schafer, "Improved goodness-of-fit tests," *Biometrika*, vol. 58, no. 3, pp. 641–645, 1971.

v1.28

INL
Idaho National Laboratory