Clustering Scenarios on Manifolds

Diego Mandelli*[a], Alper Yilmaz[b], Tunc Aldemir[a]

[a]*The Ohio State University, Nuclear Engineering Program, Columbus, OH 43210 U.S.A.*
[b]*The Ohio State University, Photogrammetric Computer Vision Laboratory, Columbus, OH 43210 U.S.A.*
**Corresponding author: mandelli.1@osu.edu*

INTRODUCTION

The recent trend to use a best estimate plus uncertainty (BEPU) approach to nuclear reactor safety analysis [1] instead of the traditional conservative approach can produce very large amounts of data. In that respect, one of the challenges that is currently emerging is the ability to effectively analyze the data generated by these methods. Clustering methodologies, such as the mean-shift methodology [2, 3], offer powerful tools that can help the user to identify scenario groups that are representative of the data [3, 4] and, hence, can reduce the complexity of data analysis efforts. By scenario clustering we mean two actions:

1. Identify the scenarios that have a similar behavior (i.e. identify the most evident classes)
2. Decide for each event sequence to which class it belongs (i.e., classification)

The approach presented in [3, 4] represents each scenario as a multi dimensional vector where each dimension corresponds to the value of one of the n state variables sampled at a specific time instant t. Since the dimensionality of such vector can be very large especially in the case when all the code outputs are considered in the clustering process (e.g. the MELCOR code [5] has 50,000 data channels), the computational time required for clustering may be excessive. By reducing the dimensionality of the vectors it is possible to decrease the computational time.

This paper shows how the dimensionality can be reduced by employing the ISOMAP [6] algorithm. The case study presented in this paper is on the analysis of the data generated by a dynamic event tree (DET) methodology applied to station blackout scenario for a pressurized water reactor (PWR) [4]. For illustration purposes, each scenario is described by nine state variables including time. The objective is to reduce the nine variables and compare the clusters (and particularly the cluster centers) obtained from the reduced dataset and the original dataset.

DIMENSIONALITY REDUCTION

In the mean-shift methodology, we represent each scenario, $\underline{s}_i$, by n state variables (e.g. node pressure, temperature) plus time:

$$\underline{s}_i=[s_i(0), s_i(1),\ldots, s_i(t),\ldots, s_i(T)] \qquad (1)$$

where $s_i(t)$ is an n-tuple which contains values of n variables $(x_{1,\ldots,}x_n)$ sampled at time $t=1,\ldots,T$. Note that the dimensionality of each scenario is $n\cdot T$ and can be extremely high for complex systems (i.e., high number of state variables and high sample instants). In this paper, we focused our attention on the reducing the dimensionality of the state variables.

Dimensionality reduction is the process of finding a bijective mapping function:

$$\mathcal{F}: \mathbb{R}^D \mapsto \mathbb{R}^d \quad (\text{with } d \le D) \qquad (2)$$

which maps the data points from the D-dimensional space into a reduced d-dimensional space, i.e. a manifold, in such a way that the distance between each point and its neighbors is preserved. In our applications $D = n+1$: n state variables plus time.

A classical example of manifolds analysis is the Swiss-roll which can be identified only by using non linear algorithms (Fig.1). In this case, points are distributed in a 3-dimensional space (i.e., $D = 3$) but they are actually lying in a 2-dimensional space (i.e., $d = 2$). The manifold in this case is represented by a 2-dimensinal plane.

For dimensionality reduction, we will implement the ISOMAP algorithm and apply it to the dataset generated by a methodology that is able to assess the impacts of both epistemic and aleatory uncertainties on the system response in a phenomenological consistent manner using dynamic event trees (DETs) [7].

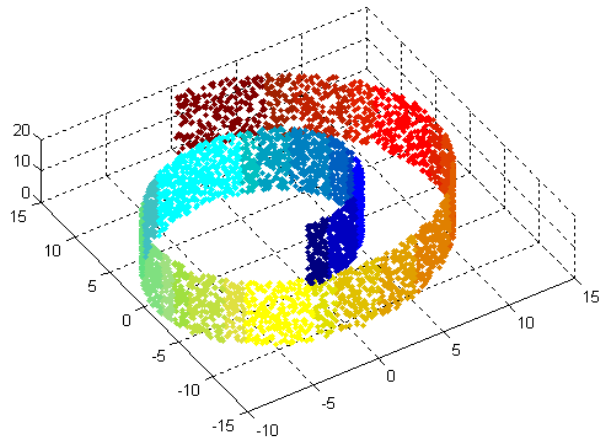

Fig. 1 Swiss-roll: example of a 2-dimenional manifold (d=2) in a 3-dimenisonal space (D=3).

THE ISOMAP ALGORITHM

The ISOMAP algorithm [6] provides a simple method for estimating the intrinsic geometry of a manifold based on a rough estimate of each data point's neighbors on the manifold. ISOMAP extends multidimensional scaling (MDS) [8] and is able to reconstruct the geometry of the manifold by computing geodesic distances[1] (distance along the manifold) using a weighted graph.

This is achieved by:

1. Estimating the geodesic distance between the points using shortest-path within its k nearest neighbors[2]. The connectivity of each data point in the neighborhood graph is defined as its nearest k Euclidean neighbors in the high-dimensional space.
2. Using MDS to find points in low-dimensional Euclidean space whose interpoint distances match the distances found in Step 1.

Since the algorithm uses non-linear combination of the state-variables (as opposed to principal component analysis [9] which considers only linear combinations of the state variables and hence is more restrictive[3]), the basis set for the low-dimensional space does not directly correspond to particular physical variables which reside in the high-dimensional space.

CASE STUDIED

The initiating event investigated was that of a station blackout (SBO) at a U.S. PWR and the MELCOR code [5] was linked to the ADAPT tool [7] to determine the evolution for each DET scenario. The simulations using MELCOR model the transient from the occurrence of the SBO through the core melting phase and up to point of containment failure and release of radionuclides to the environment. All the 104 scenarios (i=1,…,104) generated in this DET led to containment failure at some point in the scenario evolution.

For the purposes of this paper, we choose 8 state variables of interests (i.e. n=8):

1. Seal LOCA flow rate [gpm]
2. Hydrogen mass generated [kg]
3. Core water level [m]
4. System Pressure [Pa]
5. Core vapor temperature [K]
6. Hot leg vapor temperature [K]
7. Intact core fraction [%]
8. Fuel Temperature [K]

We sampled each state variable 100 times (hence, T=100) which gave us an accurate description of all the 104 transients.

RESULTS

The state space of the system described in the previous section is composed by 8 state variables and, hence, D=9. The overall number of data points distributed in this 9-dimensional space is 100·104=10400. We determined a new dataset which has the same number of points but with a reduced number of dimensions from D = 9 to d = 6.

In order to validate the new dataset against the reduced one, we compared the clusters obtained from the two datasets. While the same number of clusters were obtained from both the original and reduced datasets and each cluster contained the same number of scenarios for the high- and low-dimensional cases, some differences were observed for in the cluster centers for Clusters 1, 2 and 3 (highlighted in bold in Table 1).

When the Euclidean distances between the scenarios in each pair (62,60), (13,12) and (20,21) were determined it was found out that the differences were very small. The differences are possibly due to the fact that the dimensionality reduction process described in Eq.(2) may change slightly the geometrical distribution of the original dataset.

Figure 2 shows the cluster centers and cluster envelopes obtained from both the original and the reduced data sets using two sample state variables (core water level and system pressure), Figure 2 shows that not only the cluster centers obtained from the original and the reduced data sets are similar, but their envelopes as well.

Table 1: Comparison of cluster centers obtained from the original and the reduced datasets. Entries denote scenario identifiers.

Cluster #	Original dataset (D=9)	Reduced dataset (d=6)
1	**62**	**60**
2	**13**	**12**
3	**20**	**21**
4	7	7
5	29	29
6	34	34
7	59	59
8	12	12

CONCLUSION

This paper presents an application of manifold analysis to reduce the dimensionality of the datasets prior to clustering. A methodology based on the ISOMAP algorithm was able to identify the dependence between the set of initial variables and determine a smaller set of variables that can still describes the evolution of each scenario correctly.

We applied the methodology to the dataset generated by a DET methodology. Results showed that it is possible

[1] The geodesic distance refers to the shortest paths between two points by traversing the topology of the surface they reside.

[2] ISOMAP defines the geodesic distance to be the sum of edge weights along the shortest path between two nodes.

[3] However, ISOMAP will reduce to principal component analysis for linear data sets.

to reduce the number of variables for clustering from 9 to 6 while still identifying the clusters obtained from the original dataset. Larger reduction in dimensionality is expected if more variables are chosen to represent the scenarios for better discrimination between scenarios due to possibly higher level of correlation among the variables.

REFERENCES

1. A. BUCALOSSI, A. PETRUZZI, M. KRISTOF, F. D'AURIA, "Comparison between Best-Estimate-Plus-Uncertainty Methods and Conservative Tools for Nuclear Power Plant Licensing", *Nuclear Technology*, **172**, 29-47 (2010).
2. K. FUKUNAGA and L. HOSTETLER, "The estimation of the gradient of a density function, with applications in pattern recognition," *IEEE Transactions on Information Theory*, **21**, no. 1, pp. 32-40 (1975).
3. D. MANDELLI, T. ALDEMIR, A. YILMAZ, "Scenario Aggregation in Dynamic PRA Uncertainty Quantification", *Trans. Am. Nucl. Soc.*, **102**, 246-248 (June 2010).
4. D. MANDELLI, K. METZROTH, A. YILMAZ, R. DENNING, T. ALDEMIR, "Probabilistic Clustering for Scenario Analysis", *Trans. Am. Nucl. Soc.*, **103**, 371-374 (November 2010).
5. R. O. GAUNTT, R. K. COLE, S. A. HODGE, S. B. RODRIGUEZ, R. L. SANDERS, R. C. SMITH, D. S. STUART, R. M. SUMMERS, And M. F. YOUNG, "MELCOR Computer Code Manuals", NUREG/CR-6119/SAND 2005-5713, U.S. Nuclear Regulatory Commission, Washington, D.C. (2005).
6. J. B. TENEMBAUM, V. de SILVA, and J. C. LANGFORD, "A global geometric framework for nonlinear dimensionality reduction," *Science*, **290**, pp. 2319-2323, (2000).
7. A. HAKOBYAN, T. ALDEMIR, R. DENNING, S. DUNAGAN, D. KUNSMAN, B. RUTT, and U. CATALYUREK, "Dynamic generation of accident progression event trees," *Nuclear Engineering and Design*, **238**, no. 12, pp. 3457- 3467 (2008).
8. I. BORG and P. GROENEN, "Modern Multidimensional Scaling: theory and applications" (2nd ed.), Springer-Verlag, New York (2005).
9. I. T. JOLLIFFE, "Principal Component Analysis" (2nd ed.), Springer-Verlag, New York (2002).

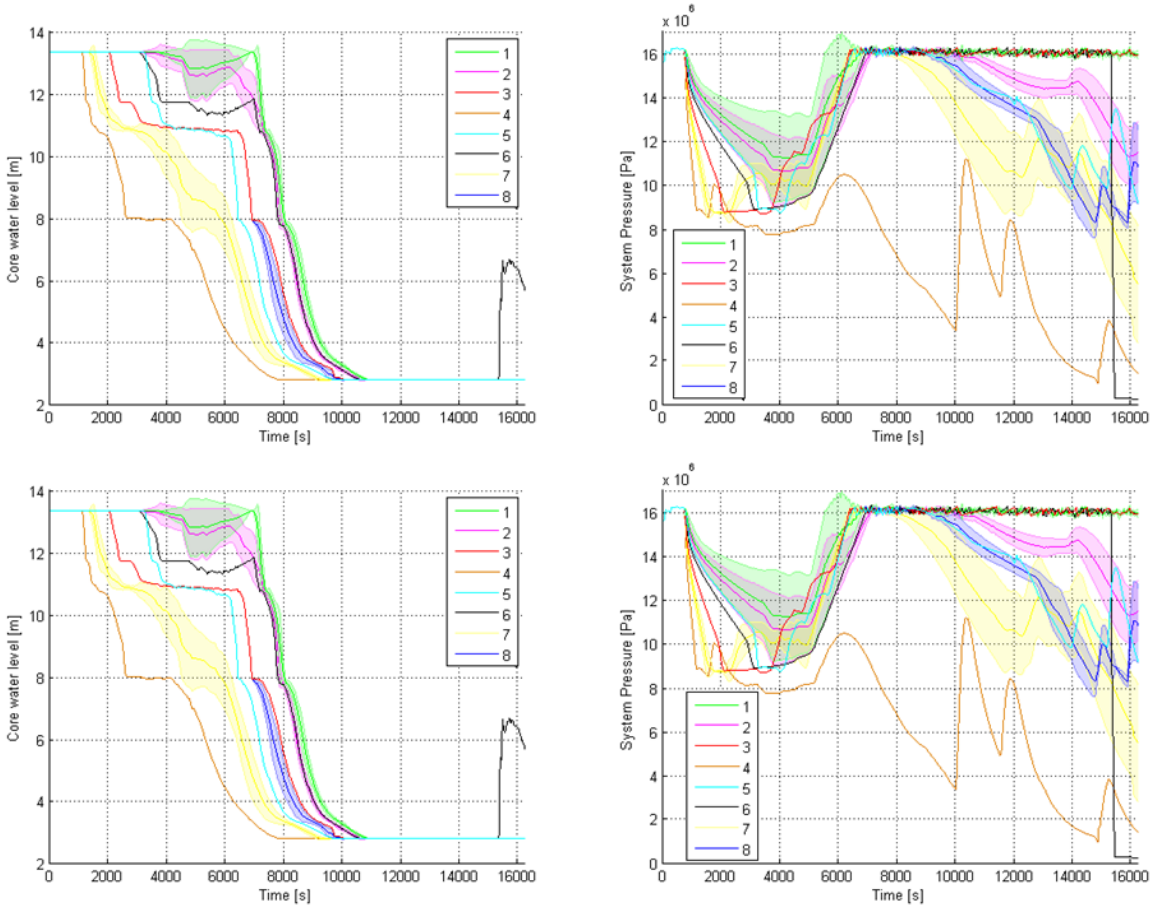

Fig. 2 Plots of the cluster centers and cluster envelopes derived from all contained within. The top two figures are for cluster obtained from the original dataset and the bottom figures are from the reduced dataset.