Visualizing Statistical Analysis of Curve Datasets in ParaView

Alejandro Ribés*, Joachim Pouderoux+, Anne-Laure Popelin* and Bertrand Iooss*

*EDF R&D, +Kitware SAS

ABSTRACT

A method for the visualization of datasets of curves has been implemented in ParaView. The method is based on performing a PCA (Principal Component Analysis) of the curves; then quantiles and outliers are estimated in the plane defined by the two more relevant components of the PCA; finally this information is reprojected in the original curves space. We have created 3 new ParaView views, 3 filters and modified the existing PCA filter to also perform a robust PCA. Results of the analysis of curves coming from the nuclear industry are shown.

Keywords: Curve visualization, Uncertainty and sensitivity analysis, ParaView, Robust PCA.

1 INTRODUCTION

Visualizing a set of curves in the same plot can be a fruitless exercise. Indeed, when a large number of curves are superposed to one another, the overall perception of the graphs is lost, and the user cannot visually analyze the datasets. This is a known problem in the visualization community.

We propose a method that performs a statistical analysis of the curves and visualizes its quantiles and outliers. The curves are, in our context, functions and this is the reason why we present an example of time dependent data. We recall that the quantiles (or outliers) of a function are not the juxtaposition of quantiles (or outliers) of each individual sample of the function. In this work we then deal with functional quantiles and outliers that are estimated by use of the PCA (Principal Component Analysis).

2 ESTIMATING QUANTILES AND OUTLIERS THROUGH PCA

The estimation of the quartiles and outliers of the dataset of curves are performed in a plane defined by the first two vectors of its Principal Component Analysis (PCA). Our method is divided into 3 main steps:

1. Project the dataset into the PCA plane.
2. Perform the estimation of the Probability Density Function (PDF) on this plane, which allows for the estimation of quartiles and outliers.
3. Project the estimated quartiles and outliers back into the curves space.

2.1 Projecting on the PCA plane

Projecting the curves into the PCA plane is a technique of dimension reduction. The goal of dimension reduction is to represent the source data into a new space with lower dimensions, where it will be easier to study. The transformation should keep enough interesting information about the source data, while allowing simplifying the analysis. There are many methods for dimension reduction. Here, we focus only on the PCA in its classical and robust variants. Once in the new space, we can use different methods to estimate the quantiles and outliers.

In a classical linear PCA, the aim is to find the orthogonal axis on which the projection of the data matrix has a maximized variance. The disadvantage of this method is the lack of robustness regarding extreme values. A variant of this criterion is to use another variable to be maximized. For instance, the Median Absolute Variance criterion maximizes the median of the absolute differences between the samples and the median, which leads to the so-called *Robust PCA* that is used in our implementation.

Once the functional variable has been transformed to a fewer component space (two in our case that defines the PCA plane), the objective is to estimate quantiles in this space, then outliers can be detected.

2.2 Highest Density Regions method

The principle of this method is to assimilate observations in the space of principal components to the realizations of a random vector with density *f*. By calculating an estimate of the density *f*, the quantiles can then be computed.

A Gaussian smoothing kernel is used, as in the article of Hyndman & Shang [1]:

$$\hat{f}(X) = \frac{1}{n}\sum_{i=1}^{n} K_H(X - X_i) \qquad (1)$$

with $K_H(X) = |H|^{-1/2} K(H^{1/2}X)$,

$K(X) = \frac{1}{2\pi}\exp\left(-\frac{1}{2}\langle X, X\rangle\right)$ is the "standard" Gaussian kernel and H is the matrix containing the smoothing parameters. Depending on this matrix (diagonal or not), some preferential smoothing directions can be chosen.

Once the estimate of *f* is obtained, the Higher Density Regions (HDR) gives a description of important statistical information. We can see an example of the HDR shape of in the lower view of figure 1, which represents the PCA plane.

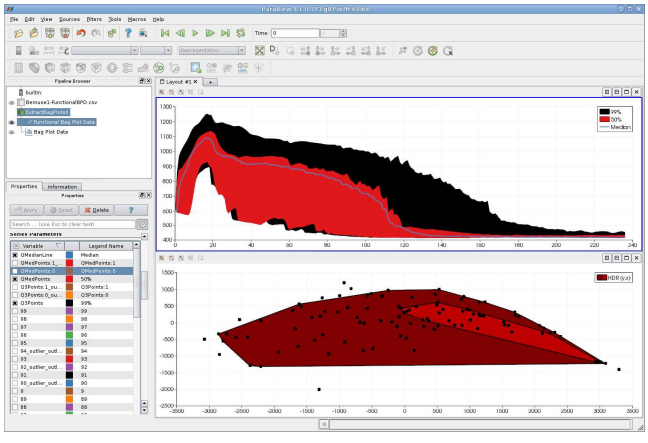

Figure 1: (top view) Overall shape of all transient curves, for the BEMUSE experiment, with their statistical bounds at 50%, in red, and 99%, in black. (bottom view) Plane defined by the first two vectors of the PCA of the BEMUSE curves. The two red regions are the statistical bounds on the plane at 50% and 99% estimated by use of the HDR method.

IEEE Symposium on Visual Analytics Science and Technology 2014
November 9-14, Paris, France
978-1-4799-6227-3/14/$31.00 ©2014 IEEE

2.3 Estimated quartiles and outliers curves

Once the HDR is calculated, the reprojection is easily preformed by using the PCA matrices. Thus, on figure 1 (top) the 50% quantile is represented with red, and black is used for the 99% quantile zone. Points outside the HDR zone are considered as outliers, for our example, they are shown in figure 2. We note that these outliers are "functional outliers" then: they can be outside the bounds or present a extreme different shape from the rest of the curves.

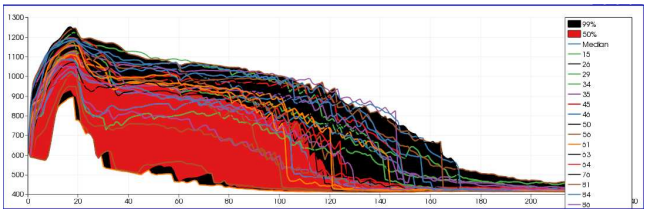

Figure 2: Transient curves, for the BEMUSE experiment, showing not only the quartiles but also all the outlier curves of this dataset.

3 Implementation in ParaView

The implementation, in ParaView, of the above described method consisted of:

- Three new ParaView views: "Box Chart View", "BagPlot Chart View", and "Functional Bag Chart View". In Figure 1 we show a screenshot of ParaView where a "Functional Bag Chart View" is visible on the top and a "BagPlot Chart View" in the bottom. We do not show the other view in this article.
- The filter "Principal Components Analysis" has been enriched with an option to calculate the robust PCA.
- Three new filters have been added, they are:

  1 / *ComputeQuartiles*, which takes a dataset or table as input and computes a new table containing the data used to display the box plots. The view Box Chart View automatically opens when applying this filter.

  2 / *Transpose table* allows the transposition of a table, it can be convenient for the user to reorder his/her data and it is also internally used by the filter *ExtractBagPlots*.

  3 / *ExtractBagPlots* takes a table as input. First, this filter typically performs a transpose but we can uncheck this option if we want to study the columns/series and not the lines of the table. Second, it applies the PCA (possibly robust if we chose it in a checkbox). Finally, the filter calculates the HDR and estimates the quartiles and outliers, as explained in section 2. This filter generates two output tables, one for each type of plots: Plot Data and Functional Bag Plot Data. Two corresponding views are automatically created when applying filter.

4 Results

In order to test the developed infrastructure we present an example coming from the nuclear industry. We consider the Benchmark for Uncertainty Analysis in Best-Estimate Modelling for Design, Operation and Safety Analysis of Light Water Reactors, [3]) proposed by the Nuclear Energy Agency of the Organization for Economic Co-operation and Development (OCDE/NEA). The study-case shown in our figures corresponds to the calculation of the LOFT L2.5 experiment, which simulated a large-break loss of primary coolant accident. It has been implemented on the French thermal-hydraulic computer code CATHARE2, developed at the Commissariat à l'Energie Atomique (CEA). One phase of the benchmark consists in applying on this study-case the so-called BEMUSE (Best Estimate Methods, Uncertainty and Sensitivity Evaluation) program in order to tests various uncertainty and sensitivity analysis methods, [3].

Our figures illustrates the BEMUSE data with 100 Monte Carlo simulations (by randomly varying around 50 uncertain inputs of the LOFT L2.5 scenario), given by CATHARE2, of the cladding temperature in function of time.

In this use-case we test that the tools described in section 3 can be helpful for thermal-hydraulical transient selection. Indeed, from a large number of curves, detecting which transients have a particular shape is not obvious. This question is particularly crucial in a sensitivity analysis approach. In figure 1 we can easily see the overall shape of all transient curves with their statistical bounds at 50% and 99%. It is also central to this work to see how this tool can easily detect outliers, which can be seen (for this use-case) in figure 2. The bottom part of figure 1 presents the two first components of the PCA analysis of the curves. In this plane the estimation of the PDF (Probability Density Function) is performed. We see that each point drawn in this plane represents a curve. It is easy to understand that the points outside the two red regions correspond to outliers. These outliers can be shown/hided interactively by the selection list in the left bottom of figure 1. In figure 2 all outliers are shown but the user could select only some of them; he/she could also display curves that are not outliers in order to compare their shapes.

5 Conclusion

In uncertainty studies, when analyzing a large number of results which are in a functional form (as time dependent curves), we are faced to difficult visualization problems. In this work, we have provided a method in order to answer to three questions found when dealing with a large number of one-dimensional curves:

1. What is the average curve?
2. Can we define some confidence interval curves containing most of the curves?
3. Can we detect some abnormal curves, in the sense of a strong difference from the majority of the curves?

The functional boxplot and bagplot tools implemented in ParaView allow answering to these three questions. Being ParaView an open-source software, these developments are already available to all the visualization community.

References

[1] R. J. Hyndman and H. L. Shang, "Rainbow plots, bagplots, and boxplots for functional data", *Journal of Computational and Graphical Statistics,* 19:29–45 (2010).

[2] A-L. Popelin, and B. Iooss, "Visualization tools for uncertainty and sensitivity analyses on thermal-hydraulic Transients", *Joint International Conference on Supercomputing in Nuclear Applications and Monte Carlo 2013* (SNA + MC 2013), Paris, France, October 27-31, 2013

[3] A. de Crécy et al., Uncertainty and sensitivity analysis of the LOFT L2-5 test: Results of the BEMUSE programme. *Nuclear Engineering and Design*, 12:3561–3578 (2008).