Uncertainty quantification and visualization for functional data

S. NANTY [1*], C. HELBERT [3], A. MARREL [1], N. PEROT [1], C. PRIEUR [2]

[1] CEA [2] Laboratoire Jean Kuntzman, Grenoble [3] Centrale Lyon

* simon.nanty@cea.fr

Introduction

Industrial context:

- Modelling of nuclear accident with thermal-hydraulic (T-H) and thermal-mechanical (T-M) workflow
- Study of the 3 functional outputs of T-H code, CATHARE2, which are inputs of T-M code, CAST3M

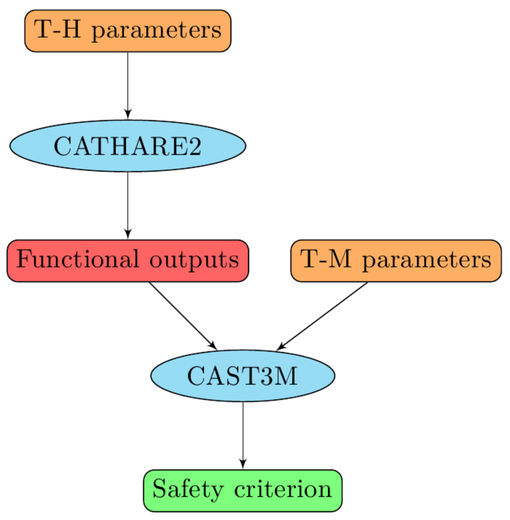
Data description:

Temperature, Pressure, Transfer Coefficient
- 1-dimensional functions depending on time
- Correlated outputs of CATHARE2
- Sample of 400 curves

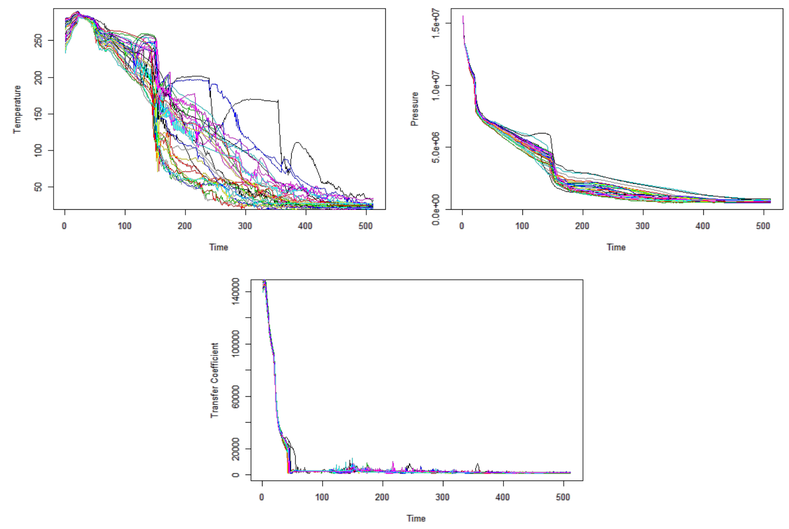
Safety Criterion
- Scalar positive variable
- Assess the security of the system

Objectives:

Extract the features of the functional variables in order to
- Generate new samples of curves following the same law
- Visualize curves



Studied T-H and T-M workflow



Sample of the functional outputs of CATHARE2

Methodology & results

Step 1: Functional Decomposition

1.a) Proposed method

Two objectives of the decomposition:
- Quality of approximation ⇨ assessed by computing L^2 norm between the curves and their approximations
- Link to the latent variable (the safety criterion)
- Taking into account the dependence between the three curves

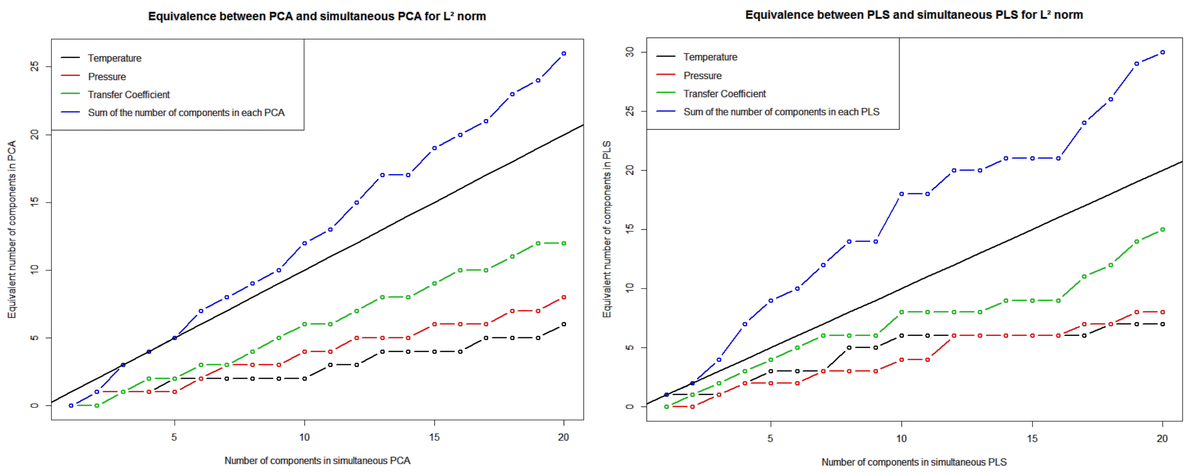
Two retained decompositions:
- Principal Components Analysis (PCA) which minimizes the L^2 norm between the initial curves and their approximations.
- Partial Least Squares (PLS) which provides a basis of components with the greatest covariance with the latent variable.

Applying PCA or PLS to the three standardized functions simultaneously with a standardization process:
- Takes into account the dependence between them
- Reduces the number of selected components

1.b) Component selection

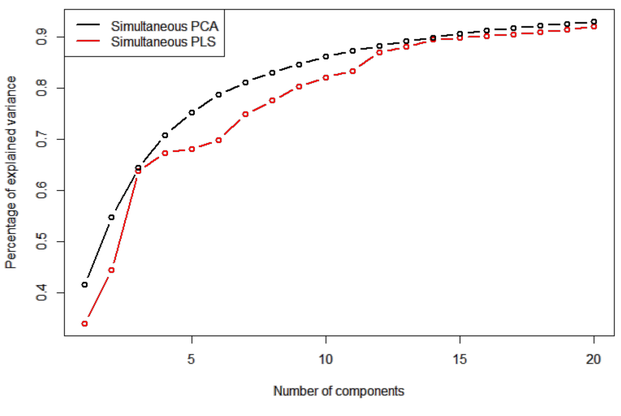
For both PCA and PLS, the first components are kept. For PCA, most of the variance lies in the first components. For PLS, these are the most correlated to the latent variable.

Number *n* of selected components is limited by the density estimation in a *n*-dimensional space in Step 2.

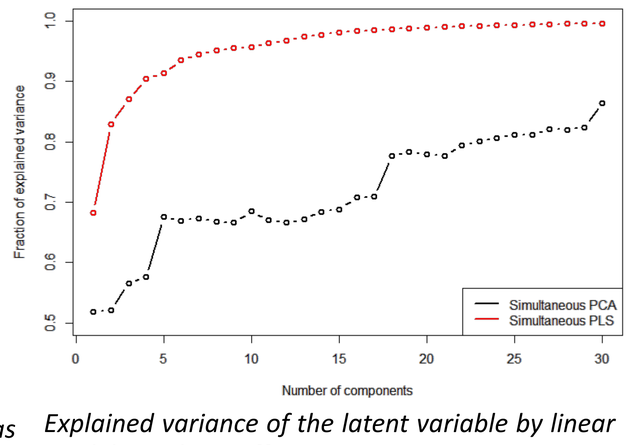


For each number of selected components in the simultaneous PCA (respectively PLS), the numbers of components needed in the PCA (resp. PLS) of each curve to get approximately the same quadratic error. The sum of the 3 other curves is drawn in blue.

Simultaneous PCA (respectively PLS) produces better approximation in L^2 norm than PCA (resp. PLS) on each type of curve, with the same number of components.



Explained variance of simultaneous PCA and PLS as functions of the number of components



Explained variance of the latent variable by linear models on the coefficients

Step 2: Density Estimation

2.a) Gaussian Mixtures

The density of the coefficients of Step 1 is approximated by Gaussian Mixtures (GM) and estimated thanks to Expectation-Maximization algorithm.

⇨ Estimation of the number of clusters in GM thanks to criteria based on likelihood: AIC (Akaike Information Criterion) or BIC (Bayesian Information Criterion)…

2.b) Sparse Gaussian Mixtures

Number of parameters in GM model increases with the number of clusters

⇨ Risk of overfitting

⇨ Proposed solution: Gaussian mixtures with sparse covariance matrices [3] based on a L^1 penalization (penalization parameter chosen by cross-validation)

2.c) Validation

The quality of the estimated density can be evaluated thanks to a test for equality of multidimensional distributions [1], which is an extension of the Cramér-Von Mises test.

⇨ Compare simulated data from the estimated distribution to the initial data

Step 3: Simulation and visualization

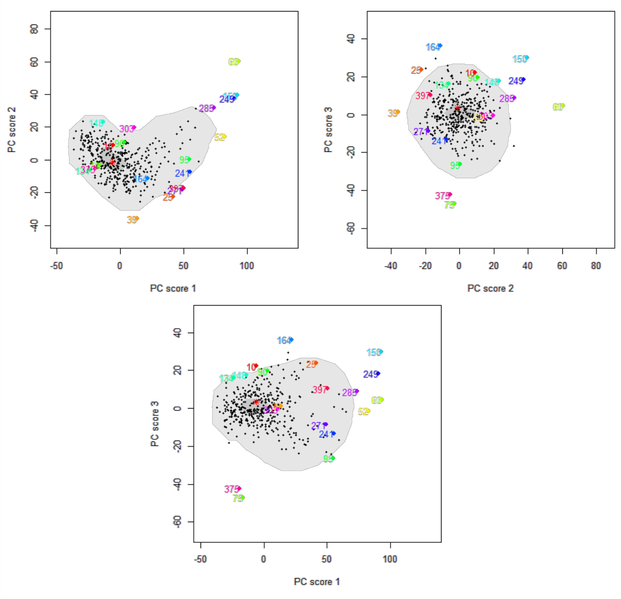
3.a) Simulation

New curves can be simulated:
- Simulate coefficients from the estimated distribution in Step 2
- Reconstruct new curves from the simulated coefficients

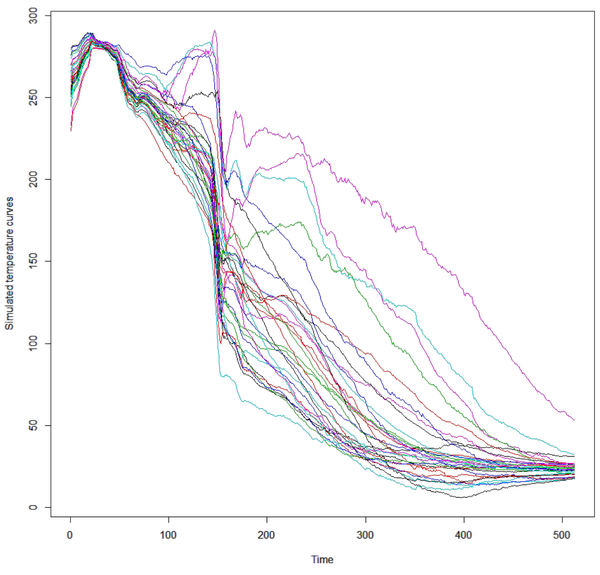
3.b) Visualization

Several methods can be applied to the data to visualize their distribution and detect outliers among the curves

- High density region boxplot [2]: initially defined with the first two components
  ⇨ Extension to more components with estimated density (Step 2)
- Functional boxplot [4] extends boxplots to functional data. Directly applied to the curves available or to simulated curve population (Step 3.a)



Projections of High density region boxplot [2] for the 1st three components of the simultaneous PLS.
Colored points: outliers
Red star: point of highest density
Gray shapes: contain 95% & 50% of the points



Sample of 30 simulated temperature curves.
Decomposition: simultaneous PLS
Number of components: 15
Number of clusters in GM: 4

Conclusion & perspectives

- A method to characterize and visualize the features of functional data has been developed.
- Simulations (Step 2) will be used to conduct uncertainty quantification and sensitivity analysis studies on CAST3M.
- The efficiency of the method to recover the main features of the data will be further studied. The safety criterion distributions obtained by running CAST3M with the initial curves and with simulated curves will be compared.

References

[1] Baringhaus, L. & Franz, C. (2004). On a new multivariate two-sample test. Journal of Multivariate Analysis, 88(1), 190-206.

[2] Hyndman, R. J. & Shang, H. L. (2010). Rainbow plots, bagplots, and boxplots for functional data. Journal of Computational and Graphical Statistics, 19(1).

[3] Krishnamurthy, A. (2011). High-Dimensional Clustering with Sparse Gaussian Mixture Models.

[4] Sun, Y. & Genton, M. G. (2011). Functional boxplots. Journal of Computational and Graphical Statistics, 20(2).