Robust analysis of functional data and adjusted wavelets frames

Background of the thesis

The scientific issue of this work is to gather and analyze, in a consistent formalism, data coming from heterogeneous but complementary sources corresponding to a single physical phenomenon. Our aim is to combine and classify them in meaningful groups and build pertinent models. Indeed, traditional statistical models need as input arrays or matrices of data. However, many problems are naturally expressed with more complex structures, as for example functional data. Functional data are very often called curves. Our main objective is to understand and develop new tools for studying their behavior. Functional data belong to high dimensional spaces. Classical statistical techniques are mainly linear and are not able to handle the whole characteristics of these observations. Other techniques like functional data analysis must be explored in order to improve the performance of the models and ease the interpretation. These types of data are very common within IFP Energies nouvelles (IFPEN), which will provide the cases to be studied. In order to be more specific, we now detail a problem where functional data are the key features.

NMR spectra and distillation curves

Nuclear magnetic resonance spectroscopy (NMR) data and distillations are very often measured on heavy petroleum products and can be seen as curves. NMR deal with chemical bounds, distillation is more concerned with physical properties of the samples. In the area of petroleum analysis, experts feel that these two aspects are complementary. Combining them in a single analysis would certainly help in extracting more meaningful information or better predictive models.

What are the problems? First, chemical measurements are prone to experimental errors, difficult to detect and leading to erroneous conclusions if not corrected. These data are called outliers. They are also present in functional data.

Secondly, the shapes of distillation curves and RMN spectra are very different and making them work together in the same mathematical formalism is not straightforward. RMN data present a lot of spikes, and could certainly be represented with wavelets while for distillations, splines seem suitable.

Additionally, as usual, the way the data are discretized can play a role. Additionally, the spectra can be deformed by pretreatments, and we expect the regression to be resistant to outliers.

State of art, and original contributions

Functional data have provoked a lot of studies. We can cite Ramsay's work [6] as a starting point in the field of statistical data analysis. Two more recent articles [5] and [7] describe the state of art. Nowadays, a few packages written in R are devoted to this topic, as can be seen on the website FunctionalData. Understanding how these tools can be put into practice is a challenge by itself which should permit IFPEN to accelerate and make the use of functional data more robust.

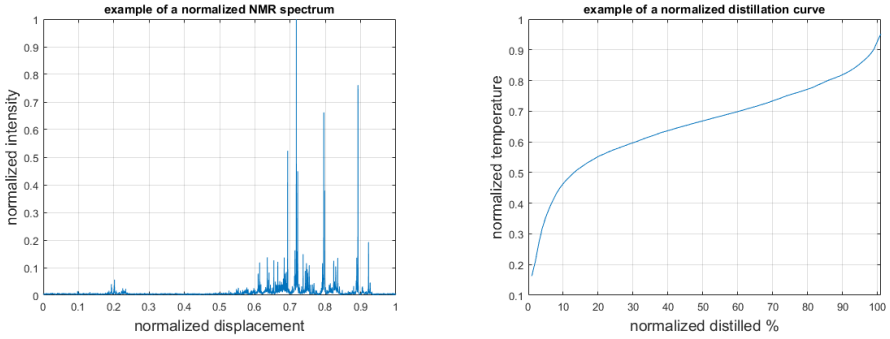

Figure 1: Examples of an NMR spectrum (left) and the corresponding distillation curve (right).

Combining methods for dimension reduction with signal decomposition techniques on well adapted bases or frames (Fourier, polynomials, splines, wavelets) is a growing tendency. These representations can highlight local characteristics of the signal, enhance subtle differences between very close signals, increase parsimony or resist to random perturbations. Some types of redundant representations are known to give high performances, especially in the case where the data differ mainly for experimental or phenomenological reasons like, for instance, shift variance. This type of pretreatment can be of great importance in unsupervised classification.

From the point of view of robustness or dimension reductions, the methods based on criteria like variance or energy very often lack of resistance, because of the use of second-moment statistics. Robust statistics, either by a judicious choice of the cost function or by well-adapted penalization (as for example in lasso algorithm) could be interesting in this context. Robust PCA with rotational invariance as proposed by [3] may find their place.

The expected original contribution of this thesis must reside in the development of more invariant and more robust methods, combining functional data analysis with convenient frames with robust regression methods in high dimension spaces. This topic does not seem to be much treated by the scientific community, maybe because of its place at the crossing of more than one technique. Note however that Han [4] steps in these directions, and cites the articles published previously by IFPEN on the design of wavelet frames [1], [2].

The new contribution of this PhD thesis will be also in the application of these recent methods to real and heterogeneous data, specifically in the field of chemometrics.

Details and Application

- PhD advisor: Clément Marteau, Université Lyon1
- Promoters:
  - François Wahl, Université Lyon1, IFP Energies nouvelles
  - Laurent Duval, IFP Energies nouvelles

The thesis will take place at Lyon, Université Lyon1.

If you are interested, send resume/CV with grades, and a short memo/letter explaining your interest and skill-match for this topic to Francois.Wahl(at)univ-lyon1.fr

References

[1] C. Chaux, J.C. Pesquet, and L. Duval. Noise covariance properties in dual-tree wavelet decompositions. *IEEE Transactions on Information Theory*, 53(12):4680–4700, 2007.

[2] C. Chaux, J.C. Pesquet, and L. Duval. A nonlinear stein based estimator for multichannel image denoising. *IEEE Transactions on on Signal Processing*, 2008.

[3] C. Ding, D. Zhou, X. He, and H. Zha. R1-pca: Rotational invariant l1-norm principal component analysis for robust subspace factorization. In *Proc. International Conference on Machine Learning*, 2006.

[4] X. Han, Z. Huang, S. Wang, S. Wang, X. Chen, K. Xu, and D. Chen. New insights to improve resolution and reliability of raman spectral analysis using higher-density multiscale regression. *Chemometrics and Intelligent Laboratory Systems*, 2017.

[5] J. Jacques and C. Preda. Functional data clustering: a survey. *Advances in Data Analysis and Classification*, 8(3):231–255, 2014.

[6] J.O. Ramsay and B.W. Silverman. *Functional Data Analysis*. Springer, 2006.

[7] J.L. Wang, J.M. Chiou, and H.G. Müller. Review of functional data analysis. *Annu. Rev. Stat. Appl.*, 2016.