Robust functional principal components for irregularly spaced longitudinal data

Ricardo Maronna
University of La Plata and U.B.A.

Consider a data set x_{ij}, with $i = 1,...,n$ and $j \in J_i \subset \{1,...,p\}$, where x_{ij} is the $j-$th observation of the random function $X_i(.)$ observed at time t_j $(j = 1,...,p)$ and J_i is the set of non-missing values for case i. We propose a parsimonious representation of the data by a linear combination of a set of q smooth functions $H_k(.)$ $(k = 1,..,q)$ in the sense that $x_{ij} \approx \sum_{k=1}^{q} \beta_{ki} H_k(t_j)$, such that it (a) is resistant to atypical X_i's ("case contamination"), (b) is resistant to isolated gross errors at some t_{ij} ("cell contamination"), and (c) can be applied when the set J_i depends on i ("irregularly spaced data").

Among the abundant literature on this subject, Boente et al. (2015) Lee et al. (2013) and Cevallos Valdiviezo (2016) deal with item (a), and Yao et al (2005) deal with (c).

Our approach to deal with all three problems, which is similar to MM-estimation, is defined as follows. Let $B_l(.)$ be a basis of B-splines; for $\boldsymbol{\alpha} = \{\alpha_{kl}\}$, $\boldsymbol{\beta} = \{\boldsymbol{\beta}_{ki}\}$ and $\boldsymbol{\mu} = \{\boldsymbol{\mu}_j\}$ put

$$\widehat{x}_{ij}(\boldsymbol{\alpha}.\boldsymbol{\beta},\boldsymbol{\mu}) = \mu_j + \sum_{k=1}^{q} \beta_{ki} H_k(t_j)$$

with $H_k(t) = \sum_{l=1}^{m} \alpha_{kl} B_l(t)$. Then the estimator is given by

$$\left(\widehat{\mathbf{a}}, \widehat{\boldsymbol{\beta}}, \widehat{\boldsymbol{\mu}}\right) = \arg\min_{\boldsymbol{\alpha},\boldsymbol{\beta},\boldsymbol{\mu}} \sum_{i=1}^{n} \sum_{k=1}^{q} \widehat{\sigma}_j^2 \rho\left(\frac{x_{ij} - \widehat{x}_{ij}(\boldsymbol{\alpha},\boldsymbol{\beta}) - \mu_j}{\widehat{\sigma}_j}\right).$$

where $\widehat{\sigma}_j$ are previously computed local scales.

The parameters are computed by an iterative algorithm starting from deterministic initial values, which are the most complex part of the procedure.

Besides, a simple and fast estimator is proposed for complete data with both types of contamination, which consists of first imputing the cell outliers by means of a robust smoother, then applying a standard robust principal components estimator, and finally smoothing the resulting components.

Simulations and real data examples indicate that these procedures outperform their competitors in most cases as respects efficiency, resistance and computing speed.

References

Boente, G. and Salibian-Barrera, M. (2015). S-Estimators for Functional Principal Component Analysis. *Journal of the American Statistical Association,* **110,** 1100-1111.

Cevallos Valdiviezo, H. (2016). On Methods for Prediction Based on Complex Data with Missing Values and Robust Principal Component Analysis, PhD thesis, Ghent University (supervisors Van Aelst S. and Van den Poel, D.).

Lee, S., Shin, H. and Billor, N. (2013). M-type smoothing spline estimators for principal functions. Computational Statistics and Data Analysis **66,** 89-â€“100.

Yao, F., Müller, H-G. and Wang, J-L. (2005). Functional Data Analysis for Sparse Longitudinal Data. *Journal of the American Statistical Association,* **100,** 577-590.