manuscript No.
(will be inserted by the editor)

Comments on: Multivariate Functional Outlier Detection

L.A. García-Escudero · A. Gordaliza · A. Mayo-Iscar

Received: date / Accepted: date

1 Introduction

First of all, we would like to congratulate M. Hubert, P. Rousseeuw and P. Segaert for this very interesting and stimulating work. It is clear that functional data are becoming ubiquitous in many disciplines and the development of appropriate statistical techniques is clearly needed. Moreover, outliers are very likely to occur in this type of data, where many measurements are taken by applying mostly unsupervised procedures. The authors provide several tools that can be successfully applied for detecting outliers when dealing with (even multivariate) functional data. They are very intuitive graphical tools based on suitable depth notions for functional data. We consider that these graphical tools are clearly useful specially in the multivariate setting where it is virtually impossible to visualize directly data curves in order to detect anomalous patterns.

In our comment, we will focus on explaining how trimming principles can be also taken into account in the detection of functional outliers.

2 Trimming and functional outlier detection

Fixed a trimming level α, trimming methods try to discard the proportion α of the "most outlying" observations in the sample. Additionally, trimming can be seen as a way to provide some kind of depth notion in data sets. The higher is the trimming level needed to remove an observation the higher is its "depth". Thus, $(1-\alpha)$-depth regions can be so defined by considering

L.A. García-Escudero · A. Mayo-Iscar · A. Gordaliza
IMUVA and Departamento de Estadística e Investigación Operativa. Facultad de Ciencias. Universidad de Valladolid. 47011, Valladolid. Spain.
Tel.: +34-983-185878
Fax: +34-983-185861
E-mail: lagarcia@eio.uva.es

those observations that "survive" after applying an α proportion trimming level. Of course, depending on the chosen trimming approach, some common assumptions for the corresponding $(1-\alpha)$-depth regions may be lost, as those having to do with convexity or the fact that regions are nested for decreasing α values.

Trimming ideas can also be applied when dealing with functional data. In fact, many procedures designed to deal with functional data can be easily adapted by including trimming in them. In this comment, we will just focus on a very simple (and well-known) approach that relies on projecting curves onto a space of functions generated by a functional basis $\{\phi_1, ..., \phi_p\}$. When those bases are properly chosen, this projection serves to reduce the dimensionality and to smooth curves by removing disturbing noise. Additionally, this does not require that curves are observed at the same evaluation points.

Given a single data curve $\{(t_i^j, x_i(t_i^j))\}_{j=1}^{J_i}$ (result of recording curve x_i at $t_i^1 < t_i^2 < ... < t_i^{J_i}$), we assume that

$$x_i(t_i^j) = \sum_{s=1}^{p} \beta_i^s \phi_s(t_i^j) + \varepsilon_i^j, j = 1, ..., J_i, \tag{1}$$

and where ε_i^j are some error terms.

Thus, by fitting n ordinary least squares regression to our n data curves $\{x_i\}_{i=1}^n$, we obtain n vectors of fitted coefficients $\{\beta_i\}_{i=1}^n$ with $\beta_i = (\beta_i^1, ..., \beta_i^p)$. For instance, cubic B-splines (with $p-4$ interior knots), Fourier, (orthogonal) polynomials, wavelets bases,... can be applied when x_i are real-valued curves. Multivariate regression in (**??**) and more sophisticated $\{\phi_1, ..., \phi_p\}$ are needed when $\{x_i\}_{i=1}^n$ are functions taking values in $\mathbb{R}^d$.

From this finite-dimensional representation of the curves, we can then apply standard trimming methods (as, for instance, the MCD or MVE) and those trimmed β_i coefficients are automatically translated into a set of trimmed x_i curves.

This trimming approach can be also applied for robust clustering. With this in mind, trimmed k-means and cubic B-splines were considered in García-Escudero and Gordaliza (2005). That approach with a large k value can also provide more flexibility when trimming the $\{\beta_i\}_{i=1}^n$ set of coefficients than the single use of trimming based on ellipsoids (and addressing non-convexity issues in this set).

Another possible extension of these ideas is motivated by the fact that trimming a complete curve may be too drastic when we only have outlying behavior during a short time interval (or only affecting few coefficients of the chosen basis). This is the analogous to the "isolated outliers" case appearing in the authors' work. In this case, "cellwise" trimming (instead of "casewise") trimming is highly advised. This can be done by applying "snipping" trimming methods as those described in Farcomeni (2014a, b).

Figure 1 shows a simple example illustrating the use of these trimming ideas in a synthetic data set where a cubic B-splines basis is considered. Curves "1" and "2" are completely trimmed ones as persistent/shift outliers and curves

"3" and "4" are only locally trimmed (affected parts of the curves are highlighted by using "•" symbols). Note that we can use the terminology "local trimming" there because, when using cubic B-splines, ϕ_s is only nonzero over a span of at most five distinct knots.

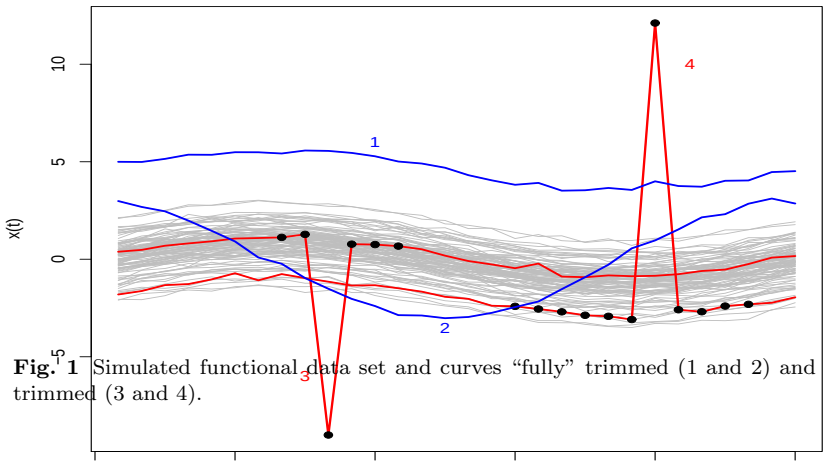

Fig. 1 Simulated functional data set and curves "fully" trimmed (1 and 2) and "locally" trimmed (3 and 4).

The ideas presented in this comment are under current investigation and more work is clearly needed but we firmly believe that trimming tools can be also useful to detect outliers in functional data sets.

Acknowledgements Research partially supported by the Spanish Ministerio de Economía y Competitividad, grant MTM2014-56235-C2-1-P, and by Consejería de Educación de la Junta de Castilla y León, grant VA212U13.

References

1. Farcomeni, A. (2014a) Snipping for robust k-means clustering under component-wise contamination, *Statistics and Computing*, **24**, 909-917.
2. Farcomeni, A. (2014b) Robust constrained clustering in presence of entry-wise outliers, *Technometrics*, **56**, 102-111.
3. García-Escudero, L.A. and Gordaliza, A. (2005), "A proposal for robust curve clustering", *Journal of Classification*, **22**, 185-201.