Outlier detection for training sets in an unsupervised functional classification framework: an application to ECG signals

F. Ieva[1]

[1] *MOX - Department of Mathematics, Politecnico di Milano (Italy)*
francesca.ieva@mail.polimi.it

Abstract

This work is concerned with a new method for definition of suitable training set for robustifying unsupervised classification techniques, in the challenging setting of multivariate functional data, i.e. data where statistical units consist of the joint observation of more than one curve. In order to do this, we generalize the concept of depth measure for functional data, defining composite indexes that enable us to define *multivariate functional outliers*.

Keywords: ECG signal, Depth Measures, Functional outlier detection, Functional k-means clustering

AMS subject classifications: 62P10, 92B15

1 Introduction

In this paper we propose a method for performing outlier detection within the framework of multivariate functional data, i.e. data where statistical units consist of the joint observation of more than one curve. In order to do this, we define suitable multivariate functional depth measure starting from the definition of functionl depth measure proposed in [3] and we make use of graphical tools like functional boxplots (see [4]), pointing out a procedure able to define *multivariate functional outliers* in the spirit of functional outlier detection of [1]. The necessity of dealing with multivariate functional data arises, for example, from biomedical context, in particular from the analysis of ECG signals. In this case, for each sunbject i, $(i = 1, \dots, n)$, a signal composed by 8 correlated curves (*leads*) is observed, in order to enhance clinical pattern recognition of physiological or pathological behaviours. Then, any automatic or semi-automatic procedure aimed at clustering (i.e. diagnosing) pathologies from the sole ECG signal morphology, has to deal with the choice of reliable training sets. In fact, classification as well as outlier detection are challenging tasks when data are curves. In general, it is mandatory a suitable training of the algorithm on data. This is particulary needed whenever a validation of an algorithm performance through cross-validation is requested. Functional boxplots [4] and a multivariate extension of depth measure for functional data [3] are then used to set a multivariate functional framework for detecting outliers and defining suitable training sets. In [2], a new method for unsupervised classification of multivariate functional data is proposed, and its application to ECG signals of PROMETEO (PROgetto sullarea Milanese Elettrocardiogrammi Teletrasferiti dall Extra Ospedaliero) project is performed and discussed. The main focus of this work is then the development of suitable techniques for defining a new concept of *outlier* for multivariate functional data, in order to decrease the misclassification cost of the semi-automatic diagnostic algorithm, robustifying selection criteria of data to be included in the training set.

2 Outlier detection for multivariate functional data

In this section we deal with the definition of *multivariate functional outlier*, starting from the concept of depth for functional data proposed in [3] and the graphical tool for functional boxplots described in [4]. In fact, a fundamental task in functional data analysis is to provide an ordering within a sample of curves that allows the definition of order statistics. A natural tool to analyse these functional data aspects is the idea of statistical depth, which measures the centrality of a given curve within a group of trajectories providing center-outward orderings of the set of curves. In general, several different definition of depth can be given (see [1]). In our case, we refer to the following definition: for any curve $f(t)$ in a sample of functions $f_1(t), \dots, f_n(t)$, $t \in I \subset \mathbb{R}$, *band depth* is defined as

$$BD_{n,J}(f) = \sum_{j=2}^{J} BD_n^{(j)}(f)$$

being $BD_n^{(j)}(f)$ the proportion of time that a curve is in the *band* determined by j different curves containing the whole graph of $f(t)$, i.e.

$$B(f_{i_1}(t), f_{i_2}(t), \dots, f_{i_j}(t)) = \left\{(t, y(t)) : t \in I, \min_r f_{i_r}(t) \leq y(t) \leq \max_r f_{i_r}(t)\right\}$$

In this work we will consider $J = 2$, so from now we will write for simplicity just $BD_n(f)$ to indicate $BD_{n,2}(f)$. Properties and consistency of the defined depth measure are discussed in [3]. Notice that the bigger is the value of band depth, the more central is the position of the curve. In fact, the *median* function in this context is the curve with greater depth, i.e. $f_{[1]} = \text{argmax}_{f \in \{f_1, \dots, f_n\}} BD_n(f)$.

In general, for a sample of functions $f_1(t), \dots, f_n(t)$, a graphical tool for visualizing outlier is proposed in [4]. In order to produce the *functional boxplot*, a depth index is computed for each curve, then curves are ranked and outlier are labelled as those curves that, for at least one t, are outside the fences obtained inflating the envelope of the 50% central region by 1.5 times the range of the 50% central region. In particular the α% central region determined by a sample of curves is defined as $\mathcal{C}_\alpha = \left\{(t, y(t)) : \min_{r=1,\dots,\lceil \alpha n \rceil} f_{[r]}(t) \leq y(t) \leq \max_{r=1,\dots,\lceil \alpha n \rceil} f_{[r]}(t)\right\}$ where $\lceil \alpha n \rceil$ is the smallest integer greater than or equal to αn. This definition clearly depends on the depth measure adopted. In our case, we do not deal with a single sample of curves, but with a sample of mathematical objects whose components are curves. Let define then a *multivariate function* as $\mathbf{f}_i : t \mapsto (f_{i1}(t), \dots, f_{ip}(t)) \in \mathbb{R}^p$, $t \in I$, $i = 1, \dots, n$, which describes parametric curves in p dimensions. We will refer to each $f_{ik}(t)$ as k-th *component* of $\mathbf{f}_i$. Notice that each component is a functional data itself. The idea is then to generalize the framework of depth measure and functional boxplot to such kind of objects, obtaining a new definition of *outlier* which takes into account simultaneously the behaviour of all the p components of $\mathbf{f}$. So, in order to define (and visualize) outlier in a family of multivariate functions, the following steps should be implemented:

1. For each statistical unit i, compute the rank of the k-th component of $\mathbf{f}_i(t)$ of the i-th curve, according to a suitable depth index;
2. Starting from the depth index of each component of $\mathbf{f}_i(t)$, compute a suitable *multivariate index of depth* for i-th unit;
3. Rank the multivariate functions $\mathbf{f}_i(t)$ according to the multivariate index of depth defined in the previous step and define *multivariate outlier* each multivariate function which is outside the fences pointed out using the multivariate index of depth and the procedure defined above, for at least one t;

4. Visualize the functional boxplot of each component, building the envelope of the 50% deepest function and then the functional boxplot according to the ranking arising from the multivariate index previously pointed out.

Notice that this algorithm defines outliers according to a multivariate index of depth, which takes into account simoultaneously the depth of all components of the multivariate function. The definition of a *multivariate outlier* will then depend on the multivariate index of depth we consider, then different proposals can be done starting from the definition of $BD_n(\mathbf{f}) = g(BD_n(f_1), \dots, BD_n(f_p))$.

For example, we can rank $\mathbf{f}_i(t)$, $i = 1, \dots, n$ using as multivariate index of depth the average depth computed on all components, i.e.

$$BD_n(\mathbf{f}_i) = \frac{1}{8}\sum_{k=1}^{p} BD_n(f_{ik}) \tag{1}$$

In this case, once the band depth in (1) has been computed, multivariate functions $\mathbf{f}_i$ can be ordered according to decreasing values of depth. Another choice is computing the Euclidean distance in $\mathbb{R}^p$ of the vector of depth of each component $(BD_n(f_{i1}), \dots, BD_n(f_{ip}))$ from the "deepest" curve, whose index of depth is $\mathbf{1} = (1, \dots, 1) \in \mathbb{R}^p$, i.e.

$$BD_n(\mathbf{f}_i) = \sum_{k=1}^{p} |BD_n(f_{ik}) - 1|^2 \tag{2}$$

In this case, multivariate functions $\mathbf{f}_i$ should be ordered according to increasing values of BD_n. Using multivariate indexes like those proposed in (1) and (2), a ranking of $\mathbf{f}_i$s is provided. Neverthless, since we defined *multivariate outliers* as in step 3, a curve can be an outlier only for some components, i.e. there could be time istances where it results outside the fences of 1.5 times the width of the central region only in some components. That is because the shape of the region depends on the shape of the curves in that component, even if they are sorted with the same order on all components. Then functional boxplots of each component of the multivariate function let us to visualize this, since they perform a sort of projection of the joint 50% central region on each component.

3 Application to ECG signals

In [2], a statistical framework for analysis and classification of ECG curves starting from their sole morphology is proposed. The main goal of the paper is then to identify, from a statistical perspective, specific ECG patterns which could benefit from an early invasive approach. In fact, the identification of statistical tools capable of classifying curves using their shape only could support an early detection of heart failures, not based on usual clinical criteria. In order to do this, a real time procedure consisting of preliminary steps like reconstructing signals, wavelets denoising and removing biological variability in the signals through data registration is tuned and tested. Then, a multivariate functional k-means clustering of reconstructed and registered data is performed. Since when testing new procedures for classification the performances of classification method are to be validated through cross validation, it is mandatory a suitable training of the algorithm on data. This would lead to robustify classification algorithm and would improve reliability in prediction. The procedure proposed in Section 2 is an effective way to reach this goal. In this case we selected for the training set the proportion of multivariate curves whose depth is greater. The ECG of the i-th patient is the multivariate function $\mathbf{f}_i$, and $p = 8$, i.e f_{ik} are the eight leads I, II, V1, V2, V3, V4, V5 and V6. Figures 1 shows 8 functional boxplots (one for each lead of the ECG), where the 98 traces of healthy patients are plotted (see [2] for details of statistical analysis and procedures). Since there is an unique ranking, induced by the multivariate index of depth for each curve, the central band is defined with the same curves in each component.

Both the multivariate indexes proposed in (1) and (2) have been tested, leading to slightly different numbers of outliers; anyway in figure 1 only the case with $BD_n(\mathbf{f}_i)$ computed as in (2) is reported.

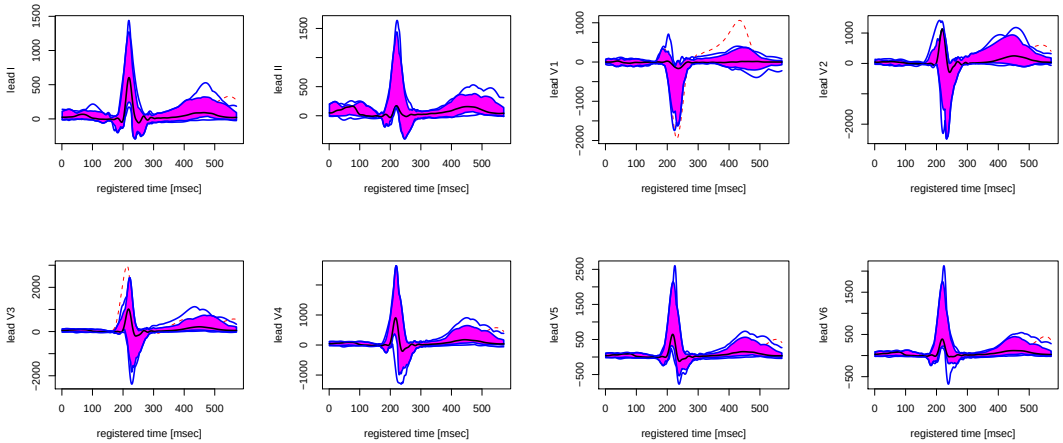

Figure 1: Functional boxplot of each component (*lead*) of the 98 phisiological ECG considered for the analysis presented in [2]. The central bands (purple coloured area) and outliers (red dotted lines) of each lead are defined as described in Section 2, according to the ranking induced by $BD_n(\mathbf{f}_i)$ computed in (2).

4 Conclusions and further developments

In this paper we proposed a general method to define outliers in multivariate functional case. We are now working on further refinements about the choice of suitable univariate depth measure in step 1 of the framework proposed in Section 2, as well as the definition of suitable multivariate index in step 2. We believe that it is important to provide general and flexible tools to deal with functional methods to detect outlier in multivariate functional datasets, since tha data output sophistication in emerging research fields requires advancing the statistical analysis of complex data.

Acknowledgements: This work is within PROMETEO (PROgetto sullarea Milanese Elettrocardiogrammi Teletrasferiti dallExtra Ospedaliero). The author wishes to thank Mortara Rangoni Europe s.r.l. for having provided data.

References

[1] Febrero M., Galeano P., Gonzalez-Manteiga W. (2008), Outlier detection in functional data by depth measures, with application to identify abnormal NOx levels, *Environmetrics* 19, 331-345

[2] Ieva F., Paganoni A.-M., Pigoli D., Vitelli V. (2011), Multivariate functional clustering for the analysis of ECG curves morphology, *Tech. Rep. MOX, Math. Dept.*, Politecnico di Milano.

[3] Lopez-Pintado S., Romo J. (2009), On the Concept of Depth for Functional Data, *Journal of the American Statistical Association* 104, 486, 718–734. Theory and Methods

[4] Sun Y., Genton M.-G. (2011), Functional Boxplots, *Journal of Computation and Graphical Statistics*, to appear.