MODELLISTICA E CALCOLO SCIENTIFICO

MODELING AND SCIENTIFIC COMPUTING

MOX-Report No. 03/2016

On Data Robustification in Functional Data Analysis

Tarabelloni, N.; Ieva, F.

MOX, Dipartimento di Matematica
Politecnico di Milano, Via Bonardi 9 - 20133 Milano (Italy)

mox-dmat@polimi.it http://mox.polimi.it

ON DATA ROBUSTIFICATION IN FUNCTIONAL DATA ANALYSIS

Nicholas Tarabelloni[♯] and Francesca Ieva[*]

[♯] MOX– Modeling and Scientific Computing
Department of Mathematics
Politecnico di Milano
Via Bonardi 9, 20133 Milano, Italy

nicholas.tarabelloni@polimi.it

[*] Department of Mathematics "F. Enriques"
Università degli Studi di Milano
Via Cesare Saldini 50, 20133 Milano, Italy

francesca.ieva@unimi.it

Keywords: Functional Data Analysis; Outlier Detection; Adjusted functional boxplot; Robust estimators; Depht Measures.

Abstract

The problem of outlier detection in high dimensional settings is nowadays a crucial point for a number of statistical analysis. Outliers are often considered as an error or noise, instead, they may carry important information on the phenomenon under study. If not properly identified, they may lead to model misspecification, biased parameter estimation and incorrect results, especially in those contexts where the number of available statistical units is lower than the number of parameters (for example, Functional Data Analysis).

In this paper we introduce a robustly adjusted version of the functional boxplot, which is the most common tool adopted to perform outlier detection in Functional Data Analysis. A crucial element of the functional boxplot is the inflation factor of the fences, controlling the proportion of observations flagged as outlier. After an overview of the methods and tools currently available in the literature, we will describe a robust method to compute a data-driven value for such inflation factor. In doing so, we will make use of robust estimators of variance-covariance operators and the corresponding eigenvalues and eigenfunctions.

Two simulation studies are proposed to give direct insights into the use of the proposed functional boxplot, and test both the robustness and accuracy of robust variance-covariance estimators, together with the performances of the functional boxplot in recognising truly outlying observations.

1 Introduction

During statistical analysis, outliers are often considered as an error or noise, instead, they may carry important information on the phenomenon under study. In fact, detected outliers are candidates for aberrant data that may otherwise adversely lead to model misspecification, biased parameter estimation and incorrect results. It is therefore important to identify them prior to modelling and data analysis (see [Wil+02] and [LSJ04]).

There is no general definition of outliers, since their presence often depends on assumptions regarding the hidden structure of data and the applied detection method. Yet, some definitions are general enough to cope with various types of data and methods. Hawkins in [Haw80] defines an outlier as *"an observation that deviates so much from other observations as to arouse suspicion* *that it was generated by a different mechanism"*. [BG05] proposes an interesting classification of outlier detection methods for both the univariate and multivariate case, distinguishing between parametric and nonparametric.

Outlier detection is particularly important in those contexts where contaminations may lead to consistent bias in estimation and inference. This is the case, among others, of Functional Data Analysis (FDA), the study field dealing with the statistical analysis of functions, which are regarded as sample realisations of suitable random processes (for an overview, see the monographs [RS05], [FV06] and [HK12]). Functional data are statistical units constituted by a sequence of measurements of a quantity of interest over a continuous variable (typically time or space). This concept suits well the rich output of scientific, technologic or economic processes considered nowadays, but requires special attention in its use since not all the statistical tools from multivariate statistics can straightforwardly be extended to this framework.
From a mathematical point of view, functional data can be seen as random functions, that is to say random elements of a functional space, i.e. $X : (\Omega, \mathcal{A}, P_X) \longrightarrow (\mathcal{V}, \|\cdot\|_{\mathcal{V}})$, where $\mathcal{V}$ is generally taken as a real, separable, infinite-dimensional Hilbert space, with norm $\|\cdot\|_{\mathcal{V}}$. The possible difficulties in the extension of classic multivariate tools to FDA are a consequence of the very rich structure of $\mathcal{V}$.

Outlier detection is of great importance in settings where inference is typically carried out with a limited amount of data compared to their dimensionality (a problem known as *large P - small N*), and it is even more the case with functional data, where observations are functions, i.e. infinite dimensional elements of function spaces. The presence of a few atypical observations may have dramatic consequences in the distortion of common sample (cross-sectional) estimators of location and dispersion parameters, which in turn are often used as basis for subsequent inferential processes. The FDA community has started debating only recently on the development of robust estimators and outlier detection techniques, able to restrain the fictitious variability induced by extremal observations in a dataset. This delay is probably due to the difficulty of pointing out an univocal definition of functional outliers as, given the very rich nature of functional data, it is difficult to point out extremality patterns that are sufficiently general to be independent of the particular dataset at hand. To this aim, an appreciable effort was carried out in [HRS15], where authors proposed a first taxonomy of functional outliers (see also the considerations in[FGGM08]). Despite that, a thorough and satisfactory definition of functional outlier is still far away from an operational point of view.

There are two ways, in general, to face outliers in a data sample: i) to apply outlier-detection tools and remove allegedly outlying observations from the dataset; ii) to robustify the estimators adopted for carrying out the inference. The first option brings about the methods for the effective identification of outliers with the aim of *robustifying* the sample (i.e. to purge it of outliers); the second directly targets the robustification of estimators. In this paper we address the problem of robustifying a dataset of curves, by enhancing a classic tool for functional outlier detection, namely the functional boxplot [SG11] (see Figure 2). Similarly to the case of real, univariate data, the functional boxplot is a visualization tool used to display the distribution of observed data, and to identify atypical curves. In particular, we will consider its adjusted version, an evolution of the standard tool proposed in [SG12] (see also [MBLR15]), where the adjustment allows to control the probability of rejecting atypical observations of a given family of genuine (i.e. not contaminated by outliers) gaussian data. We propose to integrate its adjustment process with natively functional robust estimators of location and variability, in order to overcome some shortcomings of the current technique and to have an appropriate, coherent and robust diagnostic tool for dataset robustification.

The paper is organized as follows: in Section 2 we will present and motivate the problem of outlier detection and dataset robustification in FDA, with an overview the available approaches. In Section 3 we will describe the notion of functional depths, which are a fundamental instrument of the functional boxplot. In Section 4 we will describe in detail the traditional functional boxplot and we will propose our version of the adjusted functional boxplot, based on the robust functional estimators recalled therein. Some simulation studies are presented in Section 5.
All the analyses have been carried out with R [R C15], and the implementation of the proposed method is available upon request at the BitBucket Repository https://bitbucket.org/ntarabelloni/rfda).

2 Outlier Detection and Robustification of functional data

The starting point of any functional data analysis is the reconstruction of data from noisy, pointwise observations and undergo a separation of *amplitude* from *phase* variability. The latter process is known as *registration* (or *alignment*). The first (amplitude) can be seen as a "vertical" variability in the function values, while the second (phase) is expressed by the "horizontal" dispersion of the same features across the dataset. A motivation for this separation stems from biological and medical applications, where the possible landmarks of signals can be dispersed along the horizontal axis following the so-called patient-specific variability, and entered the praxis of functional data analysis. In fact, the longitudinal dispersion of the same features would prevent from doing appropriate pointwise comparisons between signals. Registration is performed by using proper *warping* functions to map the timings of each observation to a common time (see, among others, [RS05], [Van12], [Mar+15] and references therein). As output we have a dataset of functions where the main features occur at the same reference time instant for all subjects, so that simple statistics as cross-sectional mean or covariance can be properly computed.

Although a formal and exhaustive definition of functional outlier is still missing in literature, the separation between amplitude and phase variability inspired the main distinction currently accepted between outlyingness patterns, i.e. *magnitude* and *shape* outliers. The first are related to amplitude, and are a direct analogue of the outlyingness concept in the multivariate context, while the second are related to phase variability, hence are completely new and does not have a counterpart in classic statistics. The different nature of these kinds of outliers motivates the need for different tool to detect and handle them.

In common practice, a dataset will be affected by both magnitude and shape outliers, and a first attempt to separate them is the registration step itself, where the synchronisation procedure may point at those data with degenerate warping. Alternatively, such data can be identified and possibly removed by applying detection methods tailored to shape outliers, such as the *outliergram* (see [AGR14]). At the end of this stage, the only remaining outliers will be of magnitude type, for which we propose a refined version of the classic depth-based functional outlier detection procedure based on the functional boxplot.
Our version of the functional boxplot, that fulfils both a graphical depiction of data and outlier detection, incorporates the strengths of the available estimators of location and dispersion for functional data in the traditional functional adjusted boxplot (which will be recalled in detail in Section 4). Thus, it can be seen as a synthesis of the two approaches to robust statistics for functional data: robust estimation and outlier detection.

Among the alternatives to the cross-sectional sample mean, we will consider the sample spatial median, whose definition was recently extended to functional data as a particular case of functional spatial quantile (see, e.g., [CCZ13] and [CC14b]). Alternatives to the sample covariance, instead, are proposed in applications involving PCA in [Loc+99] and [Ger08], where authors estimate covariances from data projected onto a unit sphere centred in the spatial median of data. A different approach to robust covariance estimation is followed in [KP12] and [CG15], where authors find estimators as solution of suitable (spatial median-like) L^1 minimisation problems.

In the development of robust estimators of statistical parameters, a key quantity in measuring their performances is the *breakdown* point, which we conveniently recall here. Given a dataset $\mathcal{X}$ and an estimator $T(\mathcal{X})$, its breakdown point is:

$$\varepsilon^* = \inf\left\{ \varepsilon : \sup_{\mathcal{X}'_\varepsilon} |T(\mathcal{X}) - T(\mathcal{X}'_\varepsilon)| = \infty \right\}, \tag{2.1}$$

where $\mathcal{X}'_\varepsilon$ indicates an $\varepsilon-$ contaminated/replaced/modified version of dataset $\mathcal{X}$ (see [HR09]). This means that the breakdown point measures the minimum data corruption required to bring estimates very far away from the correct one. For instance, the standard sample median has breakdown point 0.5, which is the highest value a translation-equivariant estimator can achieve [HR09]. On the contrary, just one observation can cause breakdown for the standard sample mean. This notion, that can be straightly translated to the case of functional data, makes no assumption on the particular functional form of the dataset corruption considered.

In order to exemplify the dangers of dealing with functional outliers, we show in Figure 1 the results of a simple simulation study. Here, a functional dataset of $N = 180$ observations with an outlier contamination proportion of $\varepsilon = 20\%$ (150 genuine observations and 30 outliying ones), generated according to models detailed in Subsection 5.1, is used to assess the empirical breakdown of standard sample estimators of mean and covariance. In particular, sample mean and covariance are computed on an increasing number of sample units, adding the observations one at a time, and leaving the outliers at the end, so that their effect on the estimation is identified. The sample mean is compared with the functional spatial median. In Figure 1 the logarithm of the estimation error is displayed with respect to the square root of the increasing sample size of the data used to compute the estimates.

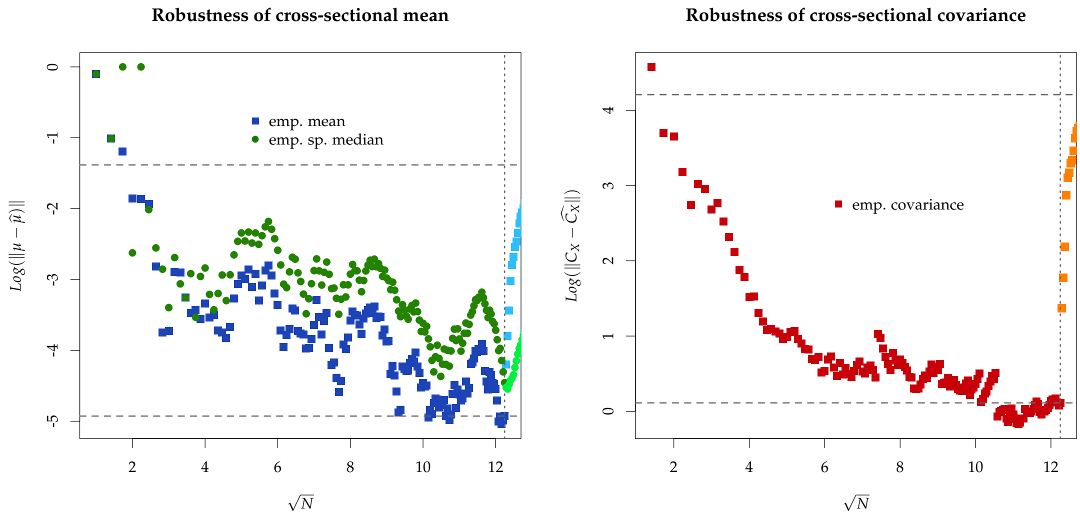

Figure 1: Corruption of standard sample estimators in presence of outliers, for data of Subsection 5.1. Sample mean and covariance estimators are computed for a dataset of 150 genuine observations to which an increasing number of outliers is added, up to 30. *Left:* Log-error norm w.r.t. exact sample mean versus square root of number of sample units, using the sample mean (squares), and using sample median (circles). *Right:* Log-error of Hilbert-Schmidt norm versus square root of number of sample units using sample covariance.

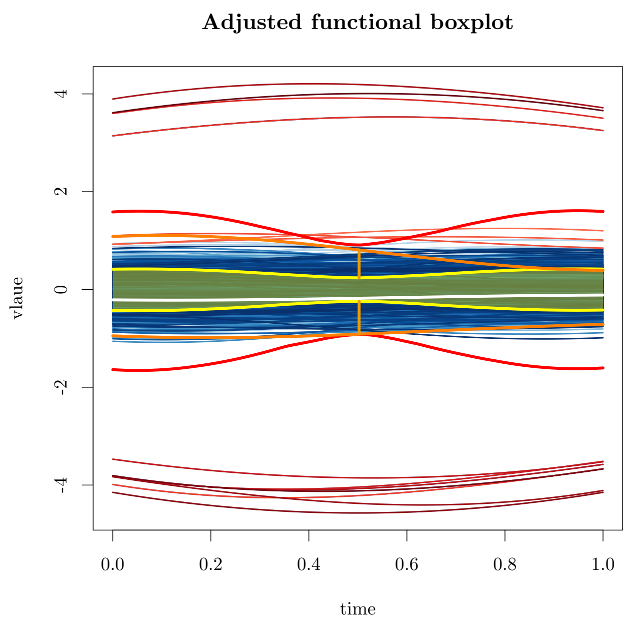

Figure 2: Adjusted functional boxplot, obtained by using the strategy proposed in the current paper on a subset of the synthetic population of the second simulation study of Subsection 5.2. The median is represented in white, the fences in orange, outliers are highlighted in red while genuine data are shown in shades of blue.

3 Statistical depth measures

Analogously to the standard, univariate boxplot for real random variables, the functional boxplot relies on a suitable ordering of functional data. Yet, high-dimensional spaces, even euclidean ones, do not have a natural order relation. Statistical depths, which were proposed in the framework of classic multivariate statistics (see the seminal paper by Tukey [Tuk75]) in order to introduce orderings in data clouds, represent a possible solution also for functional data. In particular, the statistical depth of a point with respect to a probability distribution is a measure of its centrality, hence depths offer a center-outward ordering relation for multi-dimensional data. Given a random vector $\mathbf{X} : (\Omega, \mathcal{A}, P) \longrightarrow \mathbb{R}^p$, the halfspace depth (or Tukey depth) of $\mathbf{z} \in \mathbb{R}^p$ w.r.t. $P_\mathbf{X}$ is $HD(\mathbf{z}; P_\mathbf{X}) = \inf_{a \in \mathbb{R}^p} P_\mathbf{X} \{x \in \mathbf{R}^p : \langle a, x - \mathbf{X} \rangle \geq 0\}$. Since the ideas of Tukey, much effort has been devoted to the development of a theory of depth measures, along with alternative definitions of depth for multivariate and functional data. Important examples are the *simplicial depth* (see [Liu90]), where the centrality is measured as the probability of $\mathbf{z} \in \mathbb{R}^p$ being included in a random simplex generated by the $p+1$ copies of $\mathbf{X}$; the *Mahalanobis depth* ([Liu92], [LS93]), which incorporates the Mahalanobis distance induced by $\mathbf{X}$ to define an intuitive notion of depth; the *spatial depths* [VZ00], defined starting from the spatial multivariate quantile theory as $D(\mathbf{z}; P_\mathbf{X}) = 1 - \|\mathbb{E}[(\mathbf{z} - \mathbf{X})/\|\mathbf{z} - \mathbf{X}\|]\|$; yet, many other definitions have been considered throughout the years. For an exhaustive review, see for instance [LPS99] and [ZS00].

3.1 Statistical depth for functional data

Starting from depths for multivariate data, generalisations to the functional and multivariate functional case have been recently proposed. The new definitions of depth are generally conceived so that a list of desired properties are satisfied. Among others, a very basic one (of which we will use a version to get the functional boxplot), is *affine invariance*. Given a family of functional data, $X : (\Omega, \mathcal{A}, P_X) \to (\mathcal{V}, \|\cdot\|)$, where $\mathcal{V}$ is a real, separable Hilbert space, and a generic depth function $D(\cdot, P_\mathcal{X}) \to [0,1]$, we have:

$$D(Lz, P_{LX}) = D(z, P_X), \qquad \forall z \in \mathcal{V}, \tag{3.1}$$

where $L \in \mathcal{F} \subseteq \mathcal{L}(\mathcal{V}, \mathcal{V})$, where $\mathcal{L}(\mathcal{V}, \mathcal{V})$ denotes the set of linear operators between $\mathcal{V}$ and itself. In the classic multivariate case, the linear operators generally considered for the proposition to hold are of type $L\mathbf{z} = A\mathbf{z} + \mathbf{b}$, where A is any nonsingular matrix and $\mathbf{b}$ is a vector [ZS00]. In the case of functional data, the characterisation of $\mathcal{F}$ is still an open problem, and different types of affine invariance properties are satisfied by the different depths available. Here we assume then a rather weak version of (3.1), namely *location-scale* invariance, corresponding to the particular choice of $\mathcal{F} = \{\mathcal{L} : \exists(\lambda, w) \in \mathbb{R}^+ \times \mathcal{V} : \forall x \in \mathcal{V},\ \mathcal{L}(x) = \lambda x + w\}$.

Other important properties of statistical depths, borrowed from multivariate statistics, are the *maximality at the center*, the *monotonicity w.r.t the deepest point* and the *vanishing at infinite* property (for their thorough description and review in the functional case, see [MP12]).

A first definition of depth for functional data, based on the time integration of a depth for scalar random variables, was given in [FM01]. The functional version of the halfspace depth (see among others [CC14a], where they show its structural inconsistency) and the functional version of spatial depth [CC14b], were recently proposed in the FDA community. Yet, among the most popular definitions is the *band-depth* (BD), which was proposed in [LPR07], [LPR09]. Given a random function $X \in C(I)$, where I indicates the dependent (e.g. time) variable, $X_1, X_2, \ldots, X_N$, i.i.d. copies of the process and given $J \in \mathbb{N}$, the band depth of $z \in C(I)$ is defined as:

$$\mathrm{BD}_X^J(z) = \sum_{j=1}^{J} \binom{N}{j}^{-1} \sum_{i_1<i_2<\ldots,<i_j} \mathbb{I}\left\{G(z) \in \mathrm{Env}(X_{i_1}, \ldots, X_{i_j}), \quad \forall t \in I\right\}, \tag{3.2}$$

where $G(z)$ denotes the graph of $z(t)$ and $\mathrm{Env}(X_{i_1}, \ldots, X_{i_j})$ indicates the envelope of $X_{i_1}, \ldots, X_{i_j}$'s graphs, i.e.

$$\mathrm{Env}(X_{i_1}, \ldots, X_{i_j}) = \left\{(t, y) : t \in I, \min_{l=1,\ldots,j} X_{i_l}(t) \leq y \leq \max_{r=1,\ldots,j} X_{i_r}(t)\right\}.$$

Of course, it must be $J \geq 2$, and its value controls the size of the tuples sampled from the observed data.
Due to the presence of the indicator function, in presence of real data with many crossings BD may yield low and similar values of depth to most of the observations, thus leading to the problem of heavy tails. To overcome this, authors proposed the Modified Band Depth (MBD), where the time interval that z spends in the envelope is weighted over I:

$$\mathrm{MBD}_X^J(z) = \frac{1}{J} \sum_{j=1}^{J} \binom{N}{j}^{-1} \sum_{i_1<i_2<\ldots,<i_j} \widetilde{\lambda}\left\{t \in I : \min_{l=1,\ldots,j} X_{i_l}(t) \leq z(t) \leq \max_{r=1,\ldots,j} X_{i_r}(t)\right\}, \tag{3.3}$$

where $\widetilde{\lambda}(A) = \lambda(A)/\lambda(I)$, and λ denotes the Lebesgue measure.
In [LPR09], authors state that while the choice of J clearly increases the magnitude of depth, it does not affect the induced ordering and therefore the ranks. This was supported in [Tar+15] by a simulation study involving an application with electrocardiograph (ECG) data. By setting $J = 2$, it is possible to greatly ease the computational effort required to compute depths, and exploit an exact and efficient algorithm proposed in [SGN12].
MBD fulfils property (3.1) over the subset of translation-scale functionals, yet in [LPR09], it was shown that affine invariance holds also for the linear functionals $T(x) = \lambda x + w$, with $a(t) \neq 0$ for all $t \in I$, $b \in C(I)$, and the functionals $H(x) = h(x)$, where $h(\cdot)$ is a continuous and strictly monotone mapping. Due to appealing properties and its popularity, in the following we will use MBD in the construction of the functional boxplot, but other possible depth definition can be applied, provided they fulfil (3.1).

4 The robust adjusted functional boxplot

In [SG11] authors suggest to use functional depths and the orderings they induce to build a functional boxplot, which serves both for visualisation and robustification purposes. In particular, let us denote by:

$$C_\alpha = \left\{(t, z(t)) : \min_{l=1,\ldots,\lceil \alpha N \rceil} X_l(t) \leq z(t) \leq \max_{r=1,\ldots,\lceil \alpha N \rceil} X_r(t)\right\}$$

the generic sample α-central region of data, i.e. the region containing the α% most central observations of the sample. A functional boxplot is obtained similarly to the case of univariate scalar data in the following steps: 1) take the region $C_{0.5}$, which contains the 50% of most central curves of the sample; 2) inflate it by a factor $F \geq 1$ and build the fences given by the envelope of the functions entirely contained inside the inflated region; 3) consider the observations crossing the fences or completely external as atypical curves. We prefer to use the term *atypical* rather than *outlier* to remark that these data are only observations that, according to the empirical distribution of data, should be rarely observed. Clearly, atypical observations can be either genuine but rare outcomes of the random process generating data, or corrupted data due to a possible contamination of the dataset. While the latter should be readily discarded, the former must be handled with care as, though not entirely representative of the generating law, they could still be useful for estimation. To point out this distinction, in standard univariate statistics, it is generally taken $F = 1.5$, so that in case of standard gaussian generating law, the proportion of data flagged as outliers is $\delta = 2\,\Phi(4\,z_{0.25}) \approx 0.698\%$.

In [SG12] authors argue that the choice $F = 1.5$ cannot be applied to boxplots for functional data, and suggest to select an optimal value for F such that only a fraction δ of the most outlying curves are discarded when data follow a gaussian process. This is completely in agreement with the univariate case. Yet, since no analytic expression for F can be derived, a suitable computational procedure to compute it must be devised. In [SG12] it is suggested to first estimate location and dispersion parameters from data, then to use them to simulate a synthetic population of gaussian functions without outliers, such that the optimal value of F can be computed numerically.

Despite this strategy being correct in principle, its practical fulfilment is at present not completely satisfactory, for a number of reasons that will be clear at the end of this subsection and in the following. The contribution of the present paper is to propose a valid, coherent and distribution-free alternative.

Since we are dealing with dataset potentially contaminated by outliers, suitable robust estimators must be used before simulating the gaussian population, so that a correct simulation and a coherent method is obtained. In fact, the synthetic gaussian population we would like to generate should have the same mean and covariance of the real dataset, which clearly are unknown and must be estimated.

The simulation strategy drives the choice of robust estimators. In fact, a classic simulation method is to exploit the Karhunen-Loève decomposition of X, $X = \mu + \sum_{i=1}^{\infty} \xi_i \sqrt{\lambda_i} \varphi_i$, where $\{\varphi_i, \lambda_i\}_{i \geq 1}$ are the eigen-couples of covariance function, and $\{\xi_i\}_{i \geq 1}$ is a collection of centred, uncorrelated random variables with unit variance. In practice, only the first L eigen-elements can be estimated from data, therefore we simulate a sample of gaussian functions with the truncated expansion $Z = \mu + \sum_{i=1}^{L} z_i \sqrt{\lambda_i} \varphi_i$, with $\{z_i\}_{i=1}^{L}$ i.i.d. standard normals.

The robust dispersion estimator should allow to estimate the required eigenvalues and eigenfunctions of covariance operator. Regardless of its robustness, the robust covariance estimator currently adopted in the functional boxplot, as it was suggested in [SG12], does not guarantee that its eigenvalues and eigenfunctions do actually estimate those of X's covariance. We propose, instead, two alternative robust estimators, recently proposed for functional data covariance. In Subsection 4.1 we will review both the currently employed robust covariance estimator, and the alternatives we suggest to use, highlighting their pros and cons, while in Subsection 4.2 we will make use of them in our version of the functional boxplot.

4.1 Robust estimators for functional data

In order to get an estimate of covariance operator from data, authors of [SG12] suggest to exploit a robust componentwise estimator of the dispersion matrix which was originally proposed in [MG01] in the context of random vectors. Given the random variables $X, Y : (\Omega, \mathcal{A}) \to \mathbb{R}$, observed at $\{X\}_N = \{X_1, \dots, X_N\}$ and $\{Y\}_N = \{Y_1, \dots, Y_N\}$, a highly robust estimator of their covariance is:

$$\widehat{\gamma}_{X,Y} = \frac{\alpha\beta}{4}\left[\mathcal{Q}_N^2\left(\{X\}_N/\alpha + \{Y\}_N/\beta\right) - \mathcal{Q}_N^2\left(\{X\}_N/\alpha - \{Y\}_N/\beta\right)\right] \tag{4.1}$$

where $\alpha = \mathcal{Q}_N(\{X\}_N)$, $\beta = \mathcal{Q}_N(\{Y\}_N)$. $\mathcal{Q}_N(\{X\}_N)$ is a classic estimator of the scale of X, very popular in robust statistics (see [RC92] and [RC93]), defined as:

$$\mathcal{Q}_N(\{X\}_N) = d\left\{|X_i - X_j|;\ i < j,\ i, j = 1, 2, \dots, N\right\}_{(k)}, \tag{4.2}$$

with $k = \lfloor(\binom{N}{2} + 2)/4\rfloor + 1$. The constant d is chosen depending on the probability distribution of X, in order to have Fisher consistency, and in the adjustment procedure of [SG12] is set to be equal to $(\sqrt{2}\,\Phi^{-1}(5/8))^{-1} = 2.2191$, in order to target Gaussian data (for a complete motivation, see [RC93]).

A robust estimator of covariance for a p-dimensional random vector $\mathbf{X} : (\Omega, \mathcal{A}) \to \mathbb{R}^p$ is given by applying estimator (4.1) to each couple of components of $\mathbf{X}$. A first drawback is that the matrix $\widehat{\Gamma}_Q$ obtained is in general only symmetric but not positive (semi-)definite, then it should be suitably transformed by exploiting quite complex methods discussed in [RM93] such as shrinkage.

Even if in [SG12] the resulting estimator was studied only in the multivariate case for $N \geq p$, they argue that it can be also applied in the functional case to estimate a pointwise approximation of the covariance operator $\mathcal{C}_X$ from the pointwise discretised version of data (hence in the case $N \leq p$). Anyway, estimation problems are likely to occur in this case. A possible solution could be to start by reducing the dimensionality of data. Clearly, a Karhunen-Loève reduction is not possible, because the eigen-decomposition of covariance operator is unknown without distributional assumptions, and it cannot be computed according to the cross-sectional sample estimators of covariance and mean, which can be corrupted by the outliers we want to remove. Therefore, a basis representation on a general functional basis must be used. This basis, though, needs not give a parsimonious representation of data, then the sample size could still be less than data dimension.

Beyond this and above all, the relations between the spectrum of the transformed version of covariance $\widehat{\Gamma}_Q$ and that of the original covariance operator is not clear. Finally, due to the tuning constant d, at this stage the procedure of estimating the covariance's entries is consistent only for data which are gaussian.

It is not among the scopes of the present paper to delve deeper in the study of $\widehat{\Gamma}_Q$, but in view of the drawbacks in the estimation of $\mathcal{C}_X$'s eigenvalues and eigenfunctions, we now introduce two alternative estimators that, besides being robust, can be safely employed to this aim. By using them, the estimation of $\mathcal{C}_X$'s eigenfunctions will be straightforward, while the estimation of eigenvalues will require some additional work through the use of Q_N and the properties of functional depths, and will be described directly in Subsection 4.2, where we will explain in detail our proposed robust adjusted functional boxplot.

Spherical covariance estimator

Robust estimators that directly exploit the functional nature of data can be used instead of $\widehat{\Gamma}_Q$ to compute *modal* (i.e. projective) surrogates of $\mathcal{C}_X$, and have been recently studied as a part of a general effort to spread out robust statistics techniques to FDA. In [Ger08], expanding an idea firstly advanced in [Loc+99], the author suggests the following *spherical* covariance estimator (and the corresponding sample version) for a random function with values in $L^2(I)$:

$$\mathcal{C}_S = \mathbb{E}\left[\frac{(X-\widetilde{\mu})\otimes(X-\widetilde{\mu})}{\|X-\widetilde{\mu}\|^2}\right], \qquad \widehat{\mathcal{C}}_S = \frac{1}{N}\sum_{i=1}^{N}\frac{\left(X_i-\widehat{\widetilde{\mu}}\right)\otimes\left(X_i-\widehat{\widetilde{\mu}}\right)}{\left\|X_i-\widehat{\widetilde{\mu}}\right\|^2}, \tag{4.3}$$

where $\widetilde{\mu}$ indicates the spatial (or geometric) median of X (see [Kem87] and [CC14a]), i.e. the solution $\widetilde{\mu} \in L^2(I)$ to the problem:

$$\widetilde{\mu} = \arg\min_{z\in L^2(I)} \mathbb{E}\left[\|X-z\| - \|X\|\right], \qquad \mathbb{E}\left[\frac{X-\widetilde{\mu}}{\|X-\widetilde{\mu}\|}\right] = 0, \tag{4.4}$$

that exists and is unique whenever X has a nonatomic distribution and is not entirely supported on a line. Clearly, the empirical version $\widehat{\widetilde{\mu}}$ can be obtained by addressing this M-estimation problem:

$$\widehat{\widetilde{\mu}} = \arg\min_{z\in L^2(I)} \sum_{i=1}^{N}\left(\|X_i-z\| - \|X_i\|\right), \qquad \sum_{i=1}^{N}\frac{X_i-\widehat{\widetilde{\mu}}}{\left\|X_i-\widehat{\widetilde{\mu}}\right\|} = 0. \tag{4.5}$$

The solution of equations (4.5) can be carried out with specific methods and implementations. Here we used the averaged stochastic gradient algorithm proposed in [CCZ13], suitably modified to account for a modal expansion of functional data.

The spherical covariance in (4.3) can be interpreted by considering that $(X-\widetilde{\mu})\,/\,\|X-\widetilde{\mu}\|$ is the projection of X on the unit sphere with centre in the spatial median. Then the spherical covariance $\mathcal{C}_S$ is simply the covariance of the projected data.
In [Ger08] $\mathcal{C}_S$ is used to perform robust principal component analysis on functional data, by exploiting its robustness and the spectral similarities to sample covariance $\mathcal{C}_X$. In fact it is shown that $\mathcal{C}_S$ possesses the same eigenfunctions $\{\varphi_i\}_{i=1}^{L}$ of $\mathcal{C}_X$, $\forall L \in \mathbb{N}$, with breakdown depending on the spacing of eigenvalues and generally decreasing with L. Assuming a $L-$component generative model for X (which can result after the KL truncation of original data),

$$X = \mu + \sum_{k=1}^{L}\zeta_k\sqrt{\lambda_k}\varphi_k, \tag{4.6}$$

where $\{\lambda_k, \varphi_k\}_{k=1}^{L}$ are the eigen-couples of $\mathcal{C}_X$ and $Z = (\zeta_1, \dots, \zeta_L)$ have symmetric and exchangeable marginals (e.g. Z is spherical), we have the following breakdown points for the estimated eigenfunctions of $\mathcal{C}_S$:

$$\begin{aligned}
\varepsilon^*_{\varphi_1} &\leq \frac{\tilde{\lambda}_1-\tilde{\lambda}_2}{1+\tilde{\lambda}_1-\tilde{\lambda}_2}, \\
\varepsilon^*_{\varphi_k} &\leq \min\left\{\frac{\tilde{\lambda}_{k-1}-\tilde{\lambda}_k}{1+\tilde{\lambda}_{k-1}-\tilde{\lambda}_k}, \frac{\tilde{\lambda}_k-\tilde{\lambda}_{k+1}}{1+\tilde{\lambda}_k-\tilde{\lambda}_{k+1}}\right\}, \qquad k = 2,\dots,L-1, \\
\varepsilon^*_{\varphi_L} &\leq \min\left\{\frac{\tilde{\lambda}_{L-1}-\tilde{\lambda}_L}{1+\tilde{\lambda}_{L-1}-\tilde{\lambda}_L}, \frac{\tilde{\lambda}_L}{1+\tilde{\lambda}_L}\right\}.
\end{aligned} \tag{4.7}$$

Here $\tilde{\lambda}_i$ indicates the i-th eigenvalue of $\mathcal{C}_S$ that, in general, does not coincide with λ_i, the eigenvalue of $\mathcal{C}_X$. In fact, it is shown that $\tilde{\lambda}_k = \lambda_k\Omega_{kk}$, $k = 1,\dots,L$, where $\Omega = \mathbb{E}\left[ZZ^T/Z^T\Lambda Z\right]$,

$Z = (\zeta_1, \dots, \zeta_L)$ and $\Lambda = \text{diag}(\lambda_1, \dots, \lambda_L)$. From equation (4.7) we deduce that the breakdown of φ_k depends on the spacing of $\mathcal{C}_S$'s eigenvalues, but since $\tilde{\lambda}_i \to 0$ as $i \to \infty$, the high-order eigenfunctions will have nearly zero breakdown. This is quite natural, since high-order modes are also those where variance is smaller and smaller, therefore are very easy to corrupt. Luckily, this is not a big issue, as we are interested only in the first modes of variation, because those generally capture most of X variability.

Given the relation with $\{\tilde{\lambda}_i\}$, in order to estimate eigenvalues $\{\lambda_i\}$ it is instead recommended to first compute the eigenfunctions and then to robustly estimate the variance of projected data, as it will be discussed in Subsection 4.2.

Median covariation estimator

A different robust estimator of $\mathcal{C}_X$ was proposed in [KP12], where authors formulated the problem of robust estimation in a similar way to that of the spatial median. In other words, they point out that sample covariance is a sample location estimator of the quantity $(X-\mu) \otimes (X-\mu)$, thus suggested a median-type estimator for dispersion, much in the same way as the spatial median is a robust alternative to the sample mean. The spatial median of X solves equation (4.4), therefore they suggest the estimator $\mathcal{C}_M^\rho$ solving:

$$\mathcal{C}_M^\rho = \underset{\mathcal{T} \in HS}{\arg\min}\, \mathbb{E}\left[\rho\left(\|(X-\mu) \otimes (X-\mu) - \mathcal{T}\|_{HS}\right) - \rho\left(\|(X-\mu) \otimes (X-\mu)\|_{HS}\right)\right], \tag{4.8}$$

where ρ is a real, convex function with $\rho(0) = 0$ and $\mu \in L^2(I)$ indicates a location parameter for X. Here HS indicates the space of linear Hilbert-Schmidt operators from $L^2(I)$ to itself, which, endowed with the Hilbert-Schmidt norm $\|\mathcal{T}\|_{HS} = \left(\sum_{i=1}^\infty \lambda_i^2\right)^{1/2}$ becomes a Hilbert space itself. The choice of this space is motivated by the fact that $\mathcal{C}_X$ is a Hilbert-Schmidt operator [Bos00]. We consider the case $\rho(u) = u$ and $\mu = \widetilde{\mu}$, yielding the estimator $\mathcal{C}_M$ a spatial median-type covariance operator of X.

Provided that the distribution of X is not concentrated on a line, equation (4.8) has an unique solution. Moreover, if we replace μ with a consistent estimator, the covariance estimator obtained is consistent for the corresponding operator defined around μ.

Equation (4.8) does not have an analytical solution, thus a proper computational method must be used to solve it numerically. In their original paper [KP12], authors propose to use a Quasi-Newton BFGS algorithm (see [NW99]) to directly solve the optimisation problem, while here we adapted the much simpler, efficient and reliable averaged stochastic gradient algorithm described in [CG15], suitably adapted to cope with a basis-representation of functional data.

For what concerns the eigen-decomposition of $\mathcal{C}_M$, under the assumption that the distribution of KL scores (i.e. the vector $Z = \{\zeta_i\}_{i \geq 1}$ in the un-truncated version of model (4.6)) is invariant under the change of sign of any component, $\mathcal{C}_M$ has the same eigenfunctions as $\mathcal{C}_X$. Unfortunately, similarly to $\mathcal{C}_S$, the two set of eigenvalues are not coincident and linked in a non-trivial way, in particular:

$$\tilde{\lambda}_k = \lambda_k \frac{\mathbb{E}\left[\zeta^2 / \sqrt{\sum_{i=1}^\infty (\tilde{\lambda}_i - \lambda_i \zeta_i^2)^2 + \sum_{j \neq i} \lambda_i \lambda_j \zeta_i^2 \zeta_j^2}\right]}{\mathbb{E}\left[1 / \sqrt{\sum_{i=1}^\infty (\tilde{\lambda}_i - \lambda_i \zeta_i^2)^2 + \sum_{j \neq i} \lambda_i \lambda_j \zeta_i^2 \zeta_j^2}\right]}$$

where $\tilde{\lambda}_k$ denotes the $k-$th eigenvalue of $\mathcal{C}_M$ and λ_k the $k-$th eigenvalue of C_X. Therefore, also in this case, a robust estimation from the scores of data projected on estimated eigenfunctions is recommended (see Subsection 4.2 for a discussion about this).

In Subsection 4.2, we will employ both spherical covariance and median covariation to obtain our proposed functional boxplot.

4.2 Robust adjusted functional boxplot

A property common to the robust estimators showed in the previous subsection is that, under similar hypotheses, their spectral structure is similar to that of $\mathcal{C}_X$. In particular, eigenfunctions are the same, while eigenvalues in general are not.
In order to recover the correct set of eigenvalues of $\mathcal{C}_X$, that we recall are used together with eigenfunctions to generate the gaussian population for the tuning of F, it is possible to robustly estimate them from data, once the corresponding eigenfunctions have been computed from either $\mathcal{C}_S$ or $\mathcal{C}_M$. In practice, the computation of eigenvalues turns into a set of robust univariate scale estimation problems, a very well established branch of the statistical research.

Some of the classic estimators used in this context are the MAD (median absolute deviation), $\mathcal{Q}_N$ or $\mathcal{S}_N$ (proposed together with $\mathcal{Q}_N$ in [RC93]). They are compared in [RC93], where authors seem to advise the use of $\mathcal{Q}_N$ as it has 50% breakdown and an asymptotic efficiency of 88.27% for Gaussian distributions, consequently, our choice fell on it. MAD, $\mathcal{S}_N$ and $\mathcal{Q}_N$ require all to specify a tuning, multiplicative constant (parameter d in (4.2)) which is necessary to make them consistent for the distribution of interest. They are not, in this sense, fully distribution-independent.

This could be a major issue for their use in the procedure of tuning the parameter F in the functional boxplot. In particular, we recall that eigenvalues and eigenfunctions of X should be used to generate a genuine family of gaussian data through a KL-type generative model with Gaussian scores, but if we want the procedure to be general, it should not depend on the particular latent distribution of data. A solution can come by directly exploiting the translation-scale invariance of the statistical depths (property (3.1)) used to build the functional boxplot. In particular, for a random function X, in order to generate the associated gaussian family Y, instead of using model:

$$Y_i = \mu_X + \sum_{j=1}^{\infty} \sqrt{\lambda_j}\, \zeta_{i,j}\varphi_j, \qquad \zeta_{i,j} \overset{i.i.d.}{\sim} \mathcal{N}(0,1), \tag{4.9}$$

we use model:

$$Y_i^* = \frac{(Y_i - \mu_X)}{\sqrt{\lambda_1}} = \sum_{j=1}^{\infty} \sqrt{\frac{\lambda_j}{\lambda_1}}\, \zeta_{i,j}\varphi_j, \qquad \zeta_{i,j} \overset{i.i.d.}{\sim} \mathcal{N}(0,1). \tag{4.10}$$

Owing to property (3.1), we get $D(Y|P_Y) = D(Y^*|P_{Y^*})$. As a consequence, their induced order relation is the same. Moreover, the α% central regions are related by $C_\alpha(Y) = \sqrt{\lambda_1}C_\alpha(Y^*) + \mu_X$, hence $P_{Y^*}(Y^* \in FC_\alpha(Y^*)) = P_Y(Y \in FC_\alpha(Y))$, and the value of F is the same in the two cases.

The advantage in the procedure is that the ratios $\sqrt{\lambda_j}/\sqrt{\lambda_1}$ can be conveniently estimated with $\mathcal{Q}_N(\varphi_j)/\mathcal{Q}_N(\varphi_1)$, without the need to determine the distribution-specific constant d (see Subsection 4.1). As a consequence, the whole procedure of estimating eigenfunctions and eigenvalues, and to use them in order to simulate the gaussian synthetic population (4.10) to tune F, is coherent and distribution-free. At the same time, the robustness of estimates is enforced by the use of either spherical covariance or median covariation. We come then to the algorithm in Figure 3 to compute the optimal value F^* of the robust adjusted functional boxplot:

Algorithm 1: Robust adjusted functional boxplot

Input: Functional dataset $D = \{X_1, \dots, X_N\}$
Compute $\widehat{\mathcal{C}}_S$ or $\widehat{\mathcal{C}}_M$;
for $i \in 1, \dots, L$ **do**
    compute $\widehat{\varphi}_i$;
    for $j \in 1, \dots, N$ **do**
        compute projections $p_{i,j} = \Pi_{\widehat{\varphi}_i}(X_j)$;
    compute $q_i = Q_N(\{p_{i,1}, \dots, p_{i,N}\})$;
    if $i == 1$ **then**
        $\rho_1 = 1$;
    else
        $\rho_j = q_j / q_1$;
sample M realisations of $Y_k^* = \sum_{j=1}^{L} \rho_j \zeta_{k,j} \phi_j$, with $\zeta_{k,j} \overset{i.i.d.}{\sim} \mathcal{N}(0,1)$, $k = 1, \dots, M$;
compute $\mathrm{MBD}_k = \mathrm{MBD}(Y_k^* | Y^*)$ for $k = 1, \dots M$;
compute $\widehat{C}_{0.5}(Y^*)$;
find F^* s.t. $\widehat{P}_{Y^*}(Y^* \in F^* C_{0.5}(Y^*)) = 2\Phi(4\, z_{0.25})$;

Figure 3: Pseudo-code of algorithm to compute the tuning constant F^*.

5 Simulation Studies

In this section we will present some simulation studies concerning the tools employed in the construction of the proposed functional boxplot. In particular, two aspects will be studied, namely the ability of the robust functional estimators to provide robust estimates of eigenfunctions and eigenvalues (see Subsection 5.1), and the ability of the strategy proposed to compute the optimal value F^* of reproducing the true (yet in practice unobservable) value (see Subsection 5.2).

5.1 A comparison of robust covariance estimators

While some robustness properties of the theoretical study of the breakdown properties of eigenfunctions and eigenvalues arising from $\mathcal{C}_S$ were explained before, we find particularly interesting to complete the paper with a numerical study whose aim is to assess and compare the robustness of $\mathcal{C}_S$ and $\mathcal{C}_M$ in practice.

We fix a generative model for a family of functional data with realistic features, then we artificially create a number of outliers, and study the empirical corruption of standard and robust estimators as more and more outliers are added to the dataset. As reference model we choose a gaussian process with mean $\mu_X = \sin(4\pi t)$, $t \in I = [0, 1]$, and exponential Matérn covariance function $C_X(s, t) = \alpha \exp(-\beta |s - t|)$, $s, t \in I$. In the following we will use $\alpha = 0.12$ and $\beta = 0.4$.

This simple choice produces as output functional realisations which, upon suitably choosing parameters, show the typical features of real functional data, i.e. an identifiable "shape" subjected to variability and some roughness (due for instance to the presence of unfiltered noise). Beside this, the analytical form of the covariance function leads to a more accurate computation of the exact eigenfunctions of the covariance operator. In fact, solving the KL decomposition leads to

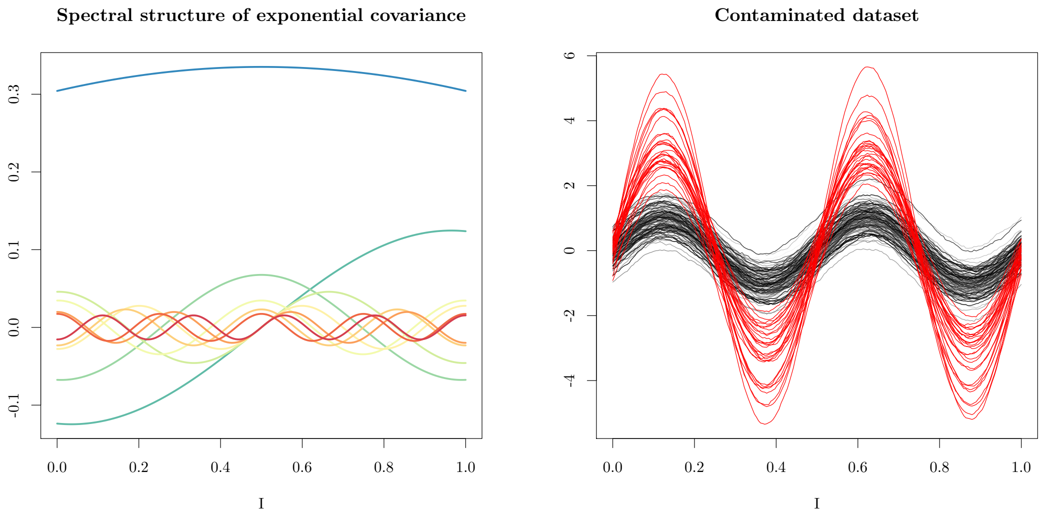

Figure 4: Illustration of the model and data for the simulation study in Subsection 5.1. *Left:* Plot of the first $L = 10$ terms $\sqrt{\lambda_i}\,\varphi_i(t)$, $i = 1, \ldots, L$, of the exponential covariance function ($\alpha = 0.12$, $\beta = 0.4$). *Right:* Plot of the genuine dataset (shades of blacks) of $X_i(t) \sim P_X$, $i = 1, \ldots, 150$ and contaminating outliers (red), $Y_i(t) \sim P_Y$, $i = 1, \ldots, 30$, with an overall contaminating proportion of $\delta = 20\%$.

the following second-order, Fredholm integral equation:

$$\int_0^1 C_X(s,t)\varphi_i(s)\mathrm{d}s = \lambda_i \varphi_i(t), \qquad \forall t \in [0,1], \qquad \forall i = 1, 2, \ldots$$

which in case of the exponential covariance can be rearranged in such a way to yield the eigenfunctions form:

$$\varphi_i(t) = \begin{cases} \cos\left(\omega_i\left(t - \frac{1}{2}\right)\right)\left(\frac{1}{2} + \frac{\sin(\omega_i)}{2\omega_i}\right)^{-\frac{1}{2}} & \text{if } i \text{ is even} \\ \sin\left(\omega_i\left(t - \frac{1}{2}\right)\right)\left(\frac{1}{2} - \frac{\sin(\omega_i)}{2\omega_i}\right)^{-\frac{1}{2}} & \text{if } i \text{ is odd} \end{cases} \tag{5.1}$$

and with eigenvalues $\lambda_i = 2\alpha/(\beta + \omega_i^2/\beta)$. Here ω_i are the (ordered) positive roots of the equation $(\beta - \omega\tan(\omega/2))(\omega - \beta\tan(\omega/2))$. Such equation cannot be solved analytically, and a root-finding algorithm should be applied. We used the univariate Brent method, available in package `stats` of `R` [R C15] through `uniroot` command, which is reasonably fast and globally convergent.

We generated a dataset of $N = 150$ functional observations, and we added up to 20% of outliers (30 functional observations) to contaminate it. The outliers were generated according to the model: $Y_i(t) = (5/2 + w_i)\mu_X(t) + Z_i(t)$, $i = 1, \ldots, 30$, where $\{w_i\}_{i=1}^{30} \sim \mathcal{E}(2)$ is an i.i.d exponential sample, μ_X is the mean of the model for X and $Z_i(t)$ is a realisation from a centred stochastic gaussian process with the same exponential covariance as X. This modelling choice

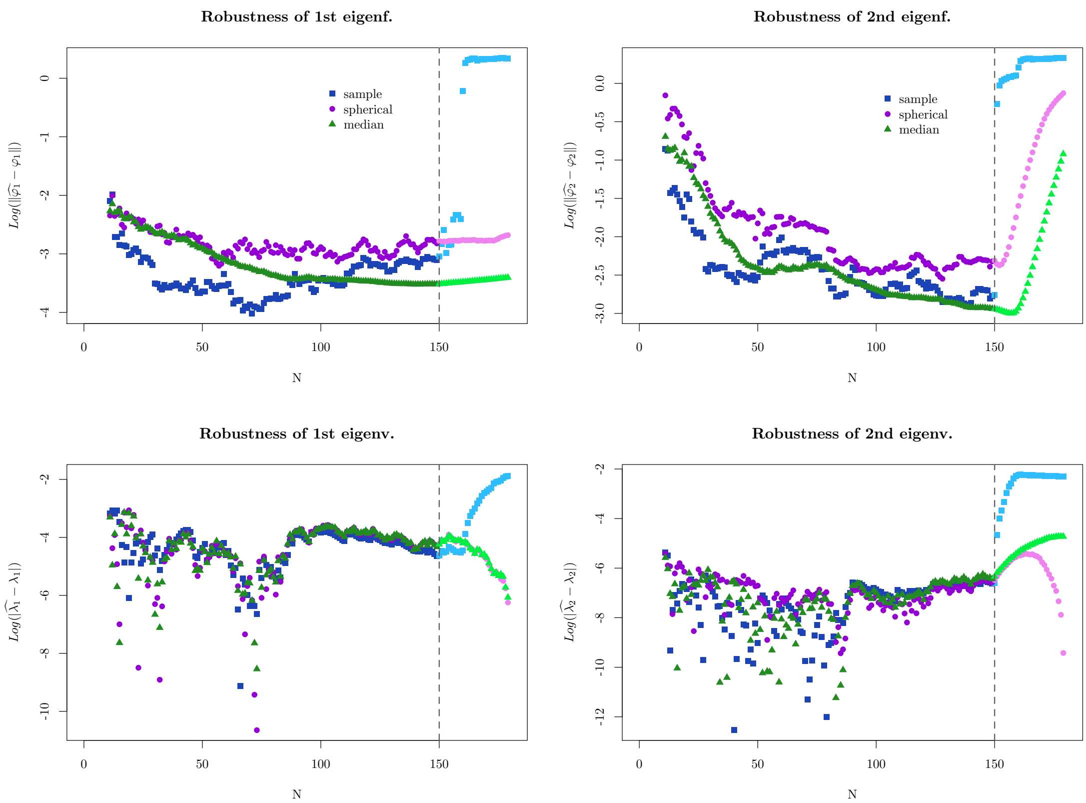

Figure 5: Comparison of eigenvalues and eigenfunctions estimation through sample, spherical and median covariance, for a functional dataset generated according to the models in Subsection 5.1. The estimation is repeated on an increasing sample size, and 20% outliers are added one at a to the dataset. *Left:* Log-error norm of first eigenfunction (*top*) and eigenvalue (*bottom*). *Right:* Log-error norm of second eigenfunction (*top*) and eigenvalue (*bottom*).

allows to have outliers with the very same shape of X, and at the same time with a magnitude undoubtedly outlying. We report two illustrative plots of the models for data and outliers in Figure 4. In the left plot are shown the quantities $\sqrt{\lambda_i}\,\varphi_i$, $i = 1, \ldots, 10$, that highlight the spectral structure of the chosen exponential covariance; in the right plot, instead, we report an example of the simulated, genuine data from P_X, and the outliers generated according to P_Y.

In order to measure the robust estimation performances of both spherical and median covariance estimators, we compute empirical eigenvalues and eigenfunctions the outlying observations adding outliers one at a time to the dataset to which the robust estimators are applied. In particular, as explained in Subsection 4.2 eigenvalues are computed by using $\mathcal{Q}_N$ to robustly estimate the standard deviation on the robustly estimated eigenfunctions. Since we know the exact expression of eigenvalues and eigenfunctions arising from the spectral decomposition of exponential covariance, we can compute the exact estimation error due to the gradual increment in data contamination. A visual summary of the performances is shown in Figure 5, where the sample, spherical and median covariance are compared. It is clear how the robust alternatives perform better both for what concerns eigenfunctions and eigenvalues. In particular, the two robust estimators seem to be quite equivalent in their performances, with perhaps a better stability in the estimation of the second eigenfunction for the median covariance.

5.2 Approximation of F^*

In this subsection we present a simulation study to assess the use of the two robust estimators in the procedure of determining the exact value of F^* in the robust, adjusted functional boxplot. In particular, in order to isolate the effect of estimators' accuracy from that of outliers, we design two schemes with progressive complexity.

In the first one, depicted in top panel of Figure 6, we will consider populations without outliers, and measure the quality of the approximation of F^* through the robust estimation of eigen-couples, with either spherical covariance $\mathcal{C}_S$ or median covariation $\mathcal{C}_M$. Hence, we expect the values $F^*_{\mathcal{S}}$ and $F^*_{\mathcal{M}}$ to be sufficiently close to the reference value F^*.
Such value, F^*, is computed from a dataset of simulated functional observations, generated by exploiting a truncated KL decomposition with $L \in \mathbb{N}$ terms of the gaussian process with the same exponential covariance of Subsection 5.1 (with $\alpha = 0.12$ and $\beta = 0.4$). The computation of the adjusted value of F makes use of a numerical optimisation method (namely, `optim` from package `stats`, which for 1D problems exploits Brent algorithm [R C15]).

We simulate a dataset of $M = 10^4$ curves with the selected generating law, observed on a grid of $P = 200$ time points. To make comparisons more fair, the size of the simulated sample generated after the estimation of eigenvalues/eigenvectors is the same as the original one, which will be used to determine the benchmark value F^*. We then compute $F^*_{\mathcal{S}}$ and $F^*_{\mathcal{M}}$ and repeat the whole procedure 50 times. A visual comparison of the obtained distributions is provided in Figure 7 (left). We can notice that the three boxplots are completely in accordance, and a Wilcoxon test for the equality of the distributions gives a p-value of about 50% for benchmark-spherical samples, and about 25% in the benchmark-median case.

In the second design, whose workflow is depicted in bottom panel of Figure 6, we add a percentage γ of outliers gto a genuine dataset, and assess the ability of our enseble estimation/simulation method to reproduce the correct F^* using either spherical covariance or median covariation. Having established that the values of F^* and $F^*_{\mathcal{S}}$ or $F^*_{\mathcal{M}}$ are in accordance in absence of outliers, here the outcome will describe mainly the effect of outliers on the estimation of the tuning factor. The generating law we use is the same as before, and we set the percentage of outliers (see the bottom panel of Figure 6) $\gamma = 2\,\Phi(4z_{0.25})$. The outliers are generated according to a symmetric law $Y_i(t) = (4 + Z_i(t))(2B_i - 1)$, where $B_i \sim \mathcal{B}(1/2)$. The symmetric design of outliers is chosen in order to select a distortion effect on the magnitude of covariance, which we expect will be reflected in the values of eigenvalues.

We remark that, the specific law for the generation of outliers we chose, is devised in such a way to produce observations undoubtedly recognisable as outlying. In fact, being the definition of functional outliers only operative and dependent on the employed outlier detection tool (above all, the functional boxplot, which we are now manipulating), we had to be sure to work with observations universally recognisable as outliers.

Again, we start with a population of $M = 10^4$ functions, with a chosen fraction γ of outliers. We compute the values $F^*_{\mathcal{S}}$ and $F^*_{\mathcal{M}}$ with a simulated population of $M = 10^4$ signals, and repeat the procedure 50 times. As a result, we compute first the fraction of outliers correctly identified

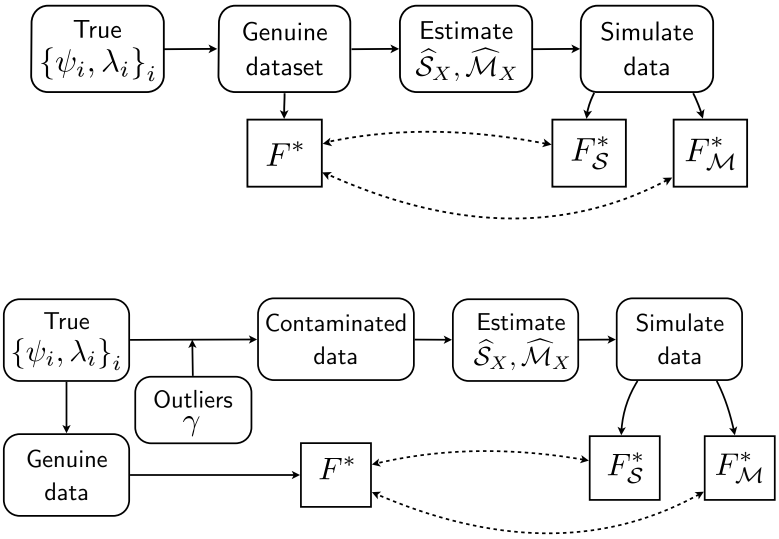

Figure 6: Analysis workflow of the first (above) and second (below) simulation designs in Subsection 5.2.

Case	Mean **TPR** ($\pm$ std. dev.)	Mean **FPR** ($\pm$ std. dev.)
Sample	1 (-)	$6.46 \cdot 10^{-3}$
Spherical	1 (-)	$6.85 \cdot 10^{-3}$
Median	1 (-)	$7.09 \cdot 10^{-3}$

Table 1: Table summarising the results of the second simulation study.

as outliers using the adjusted value of F just computed, which we call *true-positive-rate* (TPR). In all the cases, the TPR is always 1, meaning that all the real outliers are correctly identified. Then, we computed the proportion of observations incorrectly identified ad outliers, which we called false-positive-rate (FPR). The distribution of FPR across all the repetitions and for all the cases (benchmark, spherical covariance and median covariation) is shown in Figure 7 (right), while results in tabular form are reported in Table 1.
We deduce that the distributions are in complete accordance, and two Wilcoxon tests for the equality of the distributions (benchmark-spherical covariance and benchmark-median covariation) in the two cases give p-values of about 25% and 7%.

An example of functional boxplot, obtained with the mean value of the quantities F^* computed in the 50 trials, is shown in Figure 2, where for graphical reasons only a subset of $M = 10^3$ of the original dataset is used. The results are visually equivalent among the sample, spherical and median cases, thus only the result for the sample case is shown.

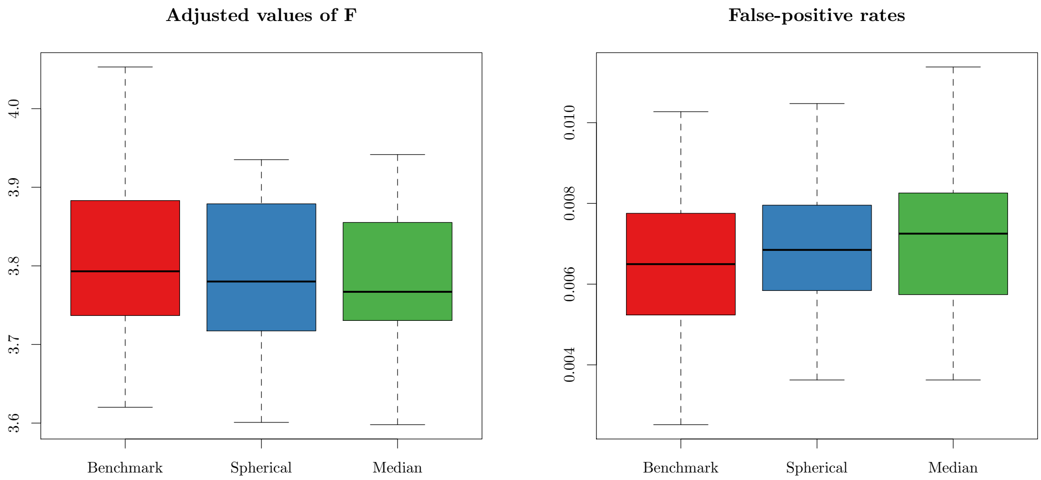

Figure 7: Outcomes of the simulation study in Subsec 5.2. *Left:* boxplots of F^*, $F^*_{\mathcal{S}}$ and $F^*_{\mathcal{M}}$. *Right:* false positive rates in outlier detection using F^*, $F^*_{\mathcal{S}}$ and $F^*_{\mathcal{M}}$.

6 Discussion and Conclusions

In this paper we focused on robust statistics and outlier detection in FDA, a branch of statistics that, despite its maturity, saw little effort in the development of robust methods. Being the definition of outlier in infinite dimensional settings still far from being exhaustive, we felt a more comprehensive overview and discussion of the topic was needed. In view of this, first we gave an overview of the scattered contributions available in literature to define and deal with outliers (both magnitude and shape ones), trying to identify a common toolchain to employ them in a general data analysis.

So far, the most important instrument to perform outlier detection is the (adjusted) functional boxplot, which we described and analysed in detail, reporting some concerns in its actual computation. In particular, these are related to the tuning of the inflation factor, F, which is based on the simulation of a dataset of gaussian functional data with same mean and covariance as the original dataset, and were described in detail.

Our proposal, instead, incorporated two alternative robust methods to estimate variance-covariance operator's spectrum, in order to take advantage of a fully functional setting and to build a coherent tuning process. These were based on the use of the recently proposed spherical covariance and median covariation estimators, whose robustness and estimation accuracy properties have also been investigated in a simulation experiment. To complete the tuning procedure, a different probabilistic law generating the family of gaussian functional data used to compute the optimal value F^* was proposed. By combining it with the property of most of functional depths of being translation-scale invariant, and with the particular robust estimator chosen to estimate functional scores, we were able to obtain the desired F^* robustly and without distribution assumptions on original data. The performances obtained by our proposed method to compute F^* and the consequent functional boxplot were studied in an ad-hoc computer experiment, were we established that results obtained with either spherical covariance or median covariation are completely satisfactory.

This work is the seminal step for a wide range of further developments. The most intuitive one is the generalisation to the multivariate functional case, which is a context where each statistical unit is a vector of functions. Moreover, the flexibility of the method allows for different tools for ranking curves to be used within the functional boxplot, provided the fulfil the translational-scale invariance property.

Acknowledgements

The authors wish to thank Prof. Anna Paganoni for the interesting discussions and the constant support provided to this work.

References

[AGR14] A. Arribas-Gil and J. Romo. "Shape outlier detection and visualization for functional data: the outliergram". In: *Biostatistics* 15.4 (2014), pp. 603–619.

[BG05] I. Ben-Gal. "Outlier Detection". In: *Maimon O. and Rockach L. (Eds.) Data Mining and Knowledge Discovery Handbook: A Complete Guide for Practitioners and Researchers*. Kluwer Academic Publishers, 2005.

[Bos00] D. Bosq. *Linear processes in function spaces: theory and applications*. Springer-Verlag, 2000.

[CC14a] A. Chakraborty and P. Chaudhuri. "On data depth in infinite dimensional spaces". In: *Ann Inst Stat Math* 66 (2014), pp. 303–324.

[CC14b] A. Chakraborty and P. Chaudhuri. "The spatial distribution in infinite dimensional spaces and related quantiles and depths". In: *The Annals of Statistics* 42.3 (2014), pp. 1203–1231.

[CCZ13] H. Cardot, P. Cénac, and P.-A. Zitt. "Efficient and fast estimation of the geometric median in Hilbert spaces with an averaged stochastic gradient algorithm". In: *Bernoulli* 19.1 (2013), pp. 18–43.

[CG15] H. Cardot and A. Godichon. "Robust principal components analysis based on the median covariation matrix". In: *arXiv preprint arXiv:1504.02852* (2015).

[FGGM08] M. Febrero, P. Galeano, and W. González-Manteiga. "Outlier detection in functional data by depth measures, with application to identify abnormal NOx levels". In: *Environmetrics* 19.4 (2008), pp. 331–345.

[FM01] R. Fraiman and G. Muniz. "Trimmed means for functional data". In: *Test* 10.2 (2001), pp. 419–440.

[FV06] F. Ferraty and P. Vieu. *Nonparametric functional data analysis: theory and practice*. Springer, 2006.

[Ger08] D. Gervini. "Robust functional estimation using the median and spherical principal components". In: *Biometrika* 95.3 (2008), pp. 587–600.

[HK12] L. Horváth and P. Kokoszka. *Inference for functional data with applications*. Springer, 2012.

[HR09] P.J. Huber and E. M. Ronchetti. *Robust Statistics, second edition*. Wiley Series in Probability and Statistics. Wiley, 2009.

[HRS15] M. Hubert, P. Rousseeuw, and P. Segaert. "Multivariate functional outlier detection". In: *Statistical Methods and Applications* (2015).

[Haw80] D. M. Hawkins. *Identification of outliers*. Vol. 11. Springer, 1980.

[KP12] D. Kraus and V. M. Panaretos. "Dispersion operators and resistant second-order functional data analysis". In: *Biometrika* 99.4 (2012), pp. 813–832.

[Kem87] J.H.B. Kemperman. "The median of a finite measure on a Banach space". In: *Statistical data analysis based on the L1-norm and related methods* (1987), pp. 217–230.

[LPR07] S. Lopez-Pintado and J. Romo. "Depth-based inference for functional data". In: *Computational Statistics & Data Analysis* 51.10 (2007), pp. 4957–4968.

[LPR09] S. Lopez-Pintado and J. Romo. "On the Concept of Depth for Functional Data". In: *Journal of the American Statistical Association* 104.486 (2009), pp. 718 –734.

[LPS99] R. Liu, J. Parelius, and K. Singh. "Multivariate analysis by data depth: descriptive statistics, graphics and inference". In: *The Annals of Statistics* 27.3 (1999), pp. 783–858.

[LS93] R. Liu and K. Singh. "A quality index based on data depth and multivariate rank tests". In: *Journal of the American Statistical Association* 88.421 (1993), pp. 252–260.

[LSJ04] H. Liu, S. Shah, and W. Jiang. "On-line outlier detection and data cleaning". In: *Computers & chemical engineering* 28.9 (2004), pp. 1635–1647.

[Liu90] R. Liu. "On a notion of data depth based on random simplices". In: *The Annals of Statistics* 18.1 (1990), pp. 405–414.

[Liu92] R. Liu. "Data depth and multivariate rank tests". In: *L1-Statistical Analysis and Related Methods* (1992), pp. 279–294.

[Loc+99] N. Locantore et al. "Robust principal component analysis for functional data". In: *Test* 8.1 (1999), pp. 1–73.

[MBLR15] B. Martin-Barragan, R.E. Lillo, and J. Romo. "Functional boxplots based on epigraphs and hypographs". In: *Journal of Applied Statistics* (2015).

[MG01] Y. Ma and M. G. Genton. "Highly robust estimation of dispersion matrices". In: *Journal of Multivariate Analysis* 78.1 (2001), pp. 11–36.

[MP12] K. Mosler and Y. Polyakova. "General notions of depth for functional data". In: *arXiv preprint arXiv:1208.1981* (2012).

[Mar+15] J.S. Marron et al. "Functional data analysis of amplitude and phase variation". In: *Statistical Science* 4.30 (2015), pp. 468–484.

[NW99] J. Nocedal and S.J. Wright. *Numerical optimization*. Springer-Verlag, 1999.

[RC92] P. J. Rousseeuw and C. Croux. "Explicit scale estimators with high breakdown point". In: *L1-Statistical analysis and related methods* (1992), pp. 77–92.

[RC93] P. J. Rousseeuw and C. Croux. "Alternatives to the median absolute deviation". In: *Journal of the American Statistical association* 88.424 (1993), pp. 1273–1283.

[RM93] P. J. Rousseeuw and G. Molenberghs. “Transformation of non positive semidefinite correlation matrices”. In: *Communications in Statistics–Theory and Methods* 22.4 (1993), pp. 965–984.

[RS05] J. O. Ramsay and B.W. Silverman. *Functional Data Analysis*. Second. New York: Springer, 2005.

[SG11] Y Sun and M. G. Genton. “Functional boxplots”. In: *Journal of Computational and Graphical Statistics* 20.2 (2011).

[SG12] Y. Sun and M.G. Genton. “Adjusted functional boxplots for spatio-temporal data visualization and outlier detection”. In: *Environmetrics* 23.1 (2012), pp. 53–64.

[SGN12] Y. Sun, M. G. Genton, and D. W. Nychka. “Exact fast computation of band depth for large functional datasets: How quickly can one million curves be ranked?” In: *Stat* 1 (2012), pp. 68–74.

[Tar+15] N. Tarabelloni et al. “Use of Depth Measure for Multivariate Functional Data in Disease Prediction: An Application to Electrocardiograph Signals”. In: *The international journal of biostatistics* (2015).

[Tuk75] J. Tukey. “Mathematics and the picturing of data”. In: *Proceedings of the International Congress of Mathematicians, Vancouver*. Vol. 2. 1975, pp. 523 –531.

[VZ00] Y. Vardi and C.-H. Zhang. “The multivariate L1-median and associated data depth”. In: *Proceedings of the National Academy of Sciences* 97.4 (2000), pp. 1423–1426.

[Van12] S. Vantini. “On the definition of phase and amplitude variability in functional data analysis”. In: *Test* 4.676–696 (2012).

[Wil+02] G. Williams et al. “A comparative study of RNN for outlier detection in data mining”. In: 2002.

[ZS00] Y. Zuo and R. Serfling. “General notions of statistical depth function”. In: *The Annals of Statistics* (2000), pp. 461–482.

[R C15] R Core Team. *R: A language and Environment for Statistical Computing*. R Foundation for Statistical Computing. Vienna, Austria, 2015. URL: `http://www.R-project.org/`.

MOX Technical Reports, last issues

Dipartimento di Matematica
Politecnico di Milano, Via Bonardi 9 - 20133 Milano (Italy)

02/2016 Crivellaro, A.; Perotto, S.; Zonca, S.
Reconstruction of 3D scattered data via radial basis functions by efficient and robust techniques

01/2016 Domanin, M.; Buora, A.; Scardulla, F.; Guerciotti, B.; Forzenigo, L.; Biondetti, P.; Vergara, C.
Computational fluid-dynamic analysis of carotid bifurcations after endarterectomy: closure with patch graft versus direct suture

62/2015 Signorini, M.; Zlotnik, S.; Díez, P.
Proper Generalized Decomposition solution of the parameterized Helmholtz problem: application to inverse geophysical problems.

63/2015 Lancellotti, R.M.; Vergara, C.; Valdettaro, L.; Bose, S.; Quarteroni, A.
Large Eddy Simulations for blood fluid-dynamics in real stenotic carotids

61/2015 Tagliabue, A.; Dedè, L.; Quarteroni, A.
Fluid dynamics of an idealized left ventricle: the extended Nitsche's method for the treatment of heart valves as mixed time varying boundary conditions

59/2015 Menafoglio, A.; Guadagnini, A.; Secchi, P.
Stochastic Simulation of Soil Particle-Size Curves in Heterogeneous Aquifer Systems through a Bayes space approach

60/2015 Perotto, S.; Reali, A.; Rusconi, P.; Veneziani, A.
HIGAMod: A Hierarchical IsoGeometric Approach for MODel reduction in curved pipes

58/2015 Iapichino, L.; Rozza, G.; Quarteroni, A.
Reduced basis method and domain decomposition for elliptic problems in networks and complex parametrized geometries

57/2015 Wilhelm, M.; Dedè, L.; Sangalli, L.M.; Wilhelm, P.
IGS: an IsoGeometric approach for Smoothing on surfaces

55/2015 Fumagalli, A.; Zonca, S.; Formaggia, L.
Advances in computation of local problems for a flow-based upscaling in fractured reservoirs