MOX-Report No. 53/2023

Functional Boxplot Inflation Factor adjustment through Robust Covariance Estimators

Rossi, A.; Cappozzo, A.; Ieva, F.

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

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

Functional Boxplot Inflation Factor adjustment through Robust Covariance Estimators

Annachiara Rossi[a], Andrea Cappozzo[a], Francesca Ieva[a,b]

[a]*MOX, Department of Mathematics, Politecnico di Milano, Via Edoardo Bonardi 9, Milan, 20133, MI, Italy*
[b]*Health Data Science Center, Human Technopole, Viale Rita Levi-Montalcini 1, Milan, 20157, MI, Italy*

Abstract

The accurate identification of anomalous curves in functional data analysis (FDA) is of utmost importance to ensure reliable inference and unbiased estimation of parameters. However, detecting outliers within the infinite-dimensional space that encompasses such data can be challenging. In order to address this issue, we present a novel approach that involves adjusting the fence inflation factor in the functional boxplot, a widely utilized tool in FDA, through simulation-based methods. Our proposed adjustment method revolves around controlling the proportion of observations considered anomalous within outlier-free replications of the original data. To accomplish this, state-of-the-art robust estimators of location and scatter are employed. In our study, we compare the performance of multivariate procedures, which are suitable for addressing the challenges posed by the "small N, large P" problems, and functional operators for implementing the tuning process. A simulation study and a real-data example showcase the validity of our proposal.

Keywords: Functional Outlier Detection, Robust Covariation Estimators, Adjusted functional boxplot

1. Introduction

In recent years, the field of functional data analysis (FDA) has gained growing attention from statisticians due to its ability to effectively represent data types that are being encountered more frequently, such as signals over time, space, and other continuum measures. A functional datum can be regarded as a realization of a functional random variable, namely an element defined over a probability space $(\Omega,\ F,\ \mathbb{P})$ with values in the Hilbert space H. In the latter, points are functions defined over a closed interval. Hence, they provide a convenient characterization for data presenting a dependence over time or space, which nowadays finds applications in a variety of fields thanks to the increased capabilities of data storage. The difference from tabular data relies upon the continuity of these observations, which also allows for the study of their differential properties. Suitable methods have been developed to analyze this data (see, for example, Ramsay and Silverman [28] and Ferraty and Vieu [8] for a detailed review) since classical multivariate tools might not always be appropriate in infinite dimensional spaces. More in detail, functional data analysis is a special case of Object Oriented Data Analysis (OODA) where the complex objects are functions. Marron and Alonso in [25] formalize the distinction between *object space* and *feature space*. The former can be, for example, the set of continuous or differentiable functions, where the data object comes from. Instead, the feature space is used to simplify our understanding of the object by representing them as *digitized vectors*. For our purposes, we consider each curve within the object space as a point in the feature space, which is assumed to have a Euclidean structure.

Within the framework of FDA, a problem that has recently drawn increasing interest is functional outlier detection. Simply put, an outlier is "an observation which deviates so much from other observations, to arouse suspicion that it was generated by a different mechanism" [11]. Therefore, outlying observations should be identified, inspected, and potentially removed before any modeling is carried out. Nevertheless, not all outliers arise from errors or noise. Caution is needed when discarding such samples from the analysis because, even if they deviate from the mechanism that generated the majority of the data, they might carry important information about the phenomenon under study. On top of that, spotting outliers in high dimensional data (large P, small N problem) is a very challenging task, because even a small proportion of contamination can easily corrupt the results. In the functional setting many different outlying behaviors can be observed: for a complete review, the interested reader is referred to Hubert, Rousseeuw, Segaert [13]. At the time of writing, the main distinction currently accepted is between magnitude and shape outliers [14]: the first refers to amplitude or vertical variability, while the second to phase or horizontal variability. Once the latter has been taken care of (e.g., through registration [28, 34]) only the dispersion in the vertical direction is left, and this will be the main focus of the present manuscript. A powerful tool for visualization and identification of amplitude outliers, i.e., the direct generalization of the traditional ones, is the functional boxplot introduced by Sun and Genton in [31]. It is inspired by the classical boxplot, firstly proposed by Tukey [33], and extended to functional data using Band Depths as a way of measuring the centrality of a signal (see [22], [23] for an introduction to functional depth measures). This representation allows for a straightforward understanding of the distribution of the set of curves. The *box* contains the central region of the data, between the first quartile Q_1 and the third quartile Q_3 of the empirical distribution. The fences are given by $[Q_1 - F \cdot IQR,\ Q_3 + F \cdot IQR]$ where IQR denotes the interquartile range. Observations outside of the fences are flagged as outliers: given the inflation factor $F = 1.5$, the probability of standing above the fences for a univariate gaussian population can be computed as $\mathbb{P}(Z > Q_3 + F \cdot IQR) = 2\Phi(4z_{0.25})$ which equals to a probability of 0.7%. Thus, the choice of this value is justified by the normality assumption. With the idea that the functional boxplot degenerates to a classical boxplot when each curve is a point, in the first version of the functional boxplot the factor F was set to 1.5 [31]. However, the authors soon realized the inherent limitations associated with this particular selection. Indeed, in [32] a simulation-based method to adjust the fences of functional boxplots based on a data-driven scheme was proposed. Particularly, the authors herein stated that *"a constant factor of 1.5 is too large when spatial correlation exists because usually, spatially correlated curves are more concentrated than independent ones"*. This necessitates the adjustment of the inflation factor within the functional framework. Motivated by this issue, the main objective of this article is to understand how several robust estimators perform in the adjustment of the fences for the generation of the bands. In particular, we want to assess whether there is a gain in implementing a procedure based on the functional form of the data and to provide a taxonomy of the available options that the final user has when tuning the inflation factor F.

The remainder of the manuscript is organized as follows: Section 2 explores more in detail the structure of the functional boxplot, the tuning procedure as introduced in [32] and the two kinds of robust covariance estimators which will be considered. Section 3 presents the simulation study, emphasizing pros and cons of the robust operators reported in Section 2 in tuning the inflation factor of the functional boxplot under diverse contamination processes. An application to real data is carried out in Section 4 and conclusions are drawn in Section 5, highlighting possible directions for future research. All the implementations devised for the manuscript have been made available in a GitHub repository at the following link https://github.com/annachiara-rossi/robust-adj-fbplot.

2. The adjusted functional boxplot

As introduced in Section 1, the functional boxplot can be employed both as a visualization tool to explore the distribution of the signals over time/space and as an advanced outlier detection mechanism. To do so, and likewise in the case of univariate data, an ordering is required to classify curves. Depth measures come in handy in this situation, as they induce a center-outward ranking; they describe how *deep* a data point is compared to the data cloud. Functional depths lay the foundation for the construction of the functional boxplot: in what follows, we will make use of the Band Depths (BD) and their Modified version (MBD) proposed by López-Pintado-Romo in [22]. Recalling the structure of the univariate scalar boxplot, the concept of *box* is formalized as the region containing the deepest 50% of the samples. In this context, the $\alpha 100\%$ central region is given by:

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

where $X_i(t)$, $i = 1, \dots, N$ is the sample of curves evaluated in $t \in I$ with I a compact interval, and $X_{(i)}$ denotes the curve associated with the i-th largest depth value relative to the dataset, so $X_{(1)} = \text{argmax}_{X \in \{X_1, \dots, X_N\}} MBD(X)$ represents the median (i.e., the deepest and most central curve). According to Equation (1), $C_{0.5}$ gives an idea of the behavior of the clean data, as we expect at least 50% of the population to be free of outliers. The fences are computed by inflating $C_{0.5}$ by a factor $F > 1$: a curve lying outside of the fences for some $t \in I$ is considered to be an outlier. In this paper, we discuss several strategies to perform a data-driven adjustment for the optimal selection of the inflation factor F.

To grasp the compelling necessity for this adjustment, refer to Figure 1: the black curves are realizations of a stochastic process, whereas we generated the red curve using a different model. The functional boxplot built without performing the data-driven adjustment in Figure 2 fails to identify such a curve as an outlying signal, as it is contained in the fences. Instead, when tuning F through one of the estimators introduced in this article, it can be seen in Figure 3 that it is actually captured as an atypical observation. This was achieved employing a tuned $F^* = 1.1$, smaller than the default $F = 1.5$.

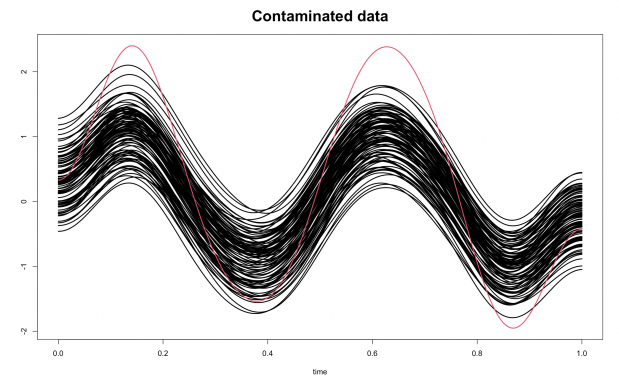


Figure 1: Simulated data with mean process $\mu = sin(4\pi t)$ and inflation of 1% of the curves by $u \sim U(1, 3)$.

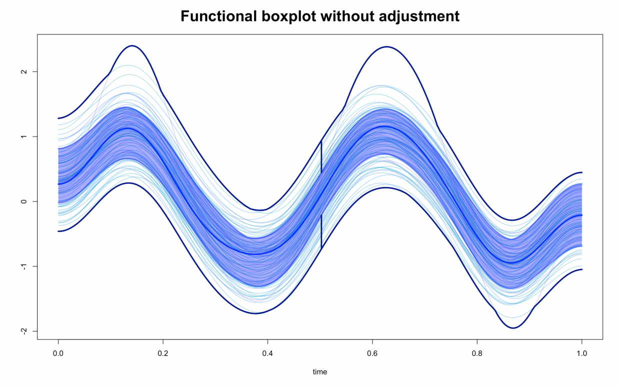


Figure 2: Functional boxplot without performing the data-driven adjustment.

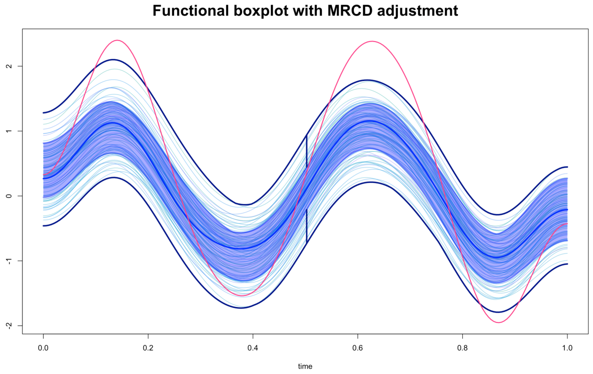


Figure 3: Functional boxplot using simulation-based adjustment employing MRCD (see Section 2.2.2).

2.1. Adjustment of the inflation factor F

We will hereafter present the scheme designed by Sun and Genton in [32] for the adjustment of the inflation factor F, to which our proposals are based upon. The idea is to control the probability of detecting no outliers, which depends on F, to be $(1 - 2\Phi(4z_{0.25}))\,100 = 99.3\%$ when the population is indeed outliers-free. The proposed solution is based on simulation. For a certain number of iterations, we generate a Gaussian population having the same mean and variance of the underlying process in our original, yet contaminated, data. We thus need an estimate of location and dispersion from the original data, which, however, must not be influenced by the extreme observations we want to spot. This is achieved through robust estimators of location and scatter. As we will explore in the following, we want the choice of the estimator to be well suited for the population under study. Given μ and Σ, representing respectively the mean vector and the scatter matrix, we sample N independent observations from a Gaussian population $\mathcal{N}(\mu, \Sigma)$ and compute: depths, $C_{0.50}$ - the 50% deepest region defined in Equation (1) - and the adjusted F for the current population i, referred to as F_i. We repeat this procedure N_{trials} times. Then, the optimal value F^* is identified as the mean value over all the N_{trials} iterations. Finally, the functional boxplot can be built for the original data with F^* as the inflation factor. In detail, such procedure can be summarized in the listing comprising Algorithm 1.

Algorithm 1 Adjusted Functional Boxplot

```
Compute scatter estimate $\widehat{\Sigma}$
Define the cost function to be minimized: $c(F, X) = P\left(X \notin F \cdot C_{50\%}\right) - 2\Phi\left(4z_{0.25}\right)$
for $i \in 1, ..., N_{trials}$ do
    Generate Gaussian population $\tilde{X}_i$
    $F_i = argmin_F c(F, \tilde{X}_i)$
end for
$F^* = \sum_{i=1}^{N_{trials}} \frac{F_i}{N_{trials}}$
Build the functional boxplot on original data using $F^*$
```

The internal procedure of Algorithm 1 will be different according to the type of Covariance estimator that will be employed: in the following sections we will provide a taxonomy of the state-of-the-art estimators available in the literature to effectively accomplish the purpose.

2.2. Multivariate Robust Covariance estimators

Sun and Genton [32] used in their version of the functional boxplot a component-wise estimator, firstly proposed by Maronna and Zamar [24]. It is the Orthogonalized Gnanadesikan-Kettenring (OGK) estimator, widely used in spatial statistics and time series analysis. In the following, we will examine some novel estimators of the same nature as OGK coming from multivariate analysis, while the subsequent section will feature the introduction of functional operators.

2.2.1. Ledoit-Wolf

As a non-robust benchmark, we use the classical well-conditioned dispersion estimator, first introduced by Ledoit and Wolf in [19]. The rationale is to find a linear combination Σ^* of the identity matrix I and the sample covariance S that ensures invertibility and does not amplify estimation errors when inverted. As such, we expect it to be significantly influenced by outlying observations.

2.2.2. Minimum Regularized Covariance Determinant (MRCD)

The first considered robust estimator is the Minimum Regularized Covariance Determinant (MRCD) proposed by Boudt et al. [3]. This method searches for a subset H of all the observations in X such that the $h = \#\{H\}$

- subset of samples has the covariance matrix with the lowest possible determinant. The subset H is chosen in the space $\mathcal{H}$ being the collection of all the possible subsets of $\{1, ..., N\}$ such that $\hat{\Sigma}^H$ is of maximal rank. This estimate of the scatter matrix minimizes the generalized variance in the data by identifying the "least contaminated" h samples. Similarly to the shrinkage operator introduced in Section 2.2.1, the MRCD approach uses a convex combination of the sample covariance matrix of the h-subset, $\hat{\Sigma}^H$, with a well-conditioned, symmetric and positive definite target matrix T, see Equation (2). The constant ρ takes values in the interval $(0, 1)$ and is derived in a data-driven way such that the condition number k (ratio between the largest and smallest eigenvalue) of the final estimate $\hat{\Sigma}_{\text{MRCD}}$ is at most $k = 50$. The regularized covariance matrix $\hat{\Sigma}^H_{\text{reg}}$ is then employed for the minimization problem defined in Equation (3).

$$\hat{\Sigma}^H_{\text{reg}} = \rho T + (1-\rho)\hat{\Sigma}^H, \tag{2}$$

$$\hat{\Sigma}_{\text{MRCD}} = \underset{H \in \mathcal{H}}{\operatorname{argmin}} \left(\det \left(\hat{\Sigma}^H_{\text{reg}} \right) \right). \tag{3}$$

The resulting estimate is well-conditioned and does not need any transformation since it is positive and semi-definite by construction. The dimension of the data subset h should be set so that $N - h$ observations can potentially be outliers. As we are dealing with functional data we assume that the target matrix T in the regularization exhibits an equicorrelation structure, as suggested in [3]:

$$\mathbf{R}_c = c\mathbf{J}_P + (1-c)\mathbf{I}_P, \tag{4}$$

where $\mathbf{J}_P$ is a matrix of ones, $\mathbf{I}_P$ is the identity matrix of dimension P, and c is an average of the robust pairwise correlations. During the initial exploration of the behavior of this estimator, we noticed that using as a target the identity matrix would lead to very low values of the inflation factor F, causing a swamping problem.

2.2.3. Kernel MRCD

A kernel version of MRCD (kMRCD) has been proposed in [30]. The main contribution relies on abandoning the hypothesis of elliptically distributed observations by using the kernel trick [12]: the estimate of the scatter matrix is computed implicitly in a kernel-induced feature space. As such, the time complexity of the algorithm is no longer dependent on the number of variables but only on the sample size. Indeed, the $P \times P$ covariance matrix is replaced by the $N \times N$ Gram matrix. Unfortunately, the kMRCD methodology allows for the recovery of the covariance structure only when a linear kernel is used. Since the estimate of the scatter matrix in the original space is required in Algorithm 1, we make use of this estimator to compare its computatioanl burden with that of MRCD, whose runtime in contrast highly depends on the dimensionality of the problem as the covariance matrix inversion costs $\mathcal{O}(P^3)$. This could be a significant improvement in computational complexity when treating functional data. As a supplementary contribution to the present work we translated in R the Matlab implementation originally provided by the authors in [30]: the source code is freely available at the github.com/annachiara-rossi/kMRCD GitHub repository.

2.2.4. Adjustment procedure for multivariate estimators

In this setting, to simulate a Gaussian population that emulates the behavior of the uncontaminated samples during the adjustment procedure for the functional boxplot outlined in Section 2.1, we propose to employ a robust estimation of the dispersion matrix. One of the estimators discussed in Sections 2.2.1, 2.2.2, 2.2.3 can be used for this task. To this aim, we need an estimate of the median, provided by the curve of maximal depth, and the resulting covariance structure. In detail, the model considered for the generation of univariate functional data reads as follows:

$$\begin{aligned} X(t) &= m(t) + \epsilon(t), \quad t \in I = [a, b], \\ \mathrm{Cov}(\epsilon(s), \epsilon(t)) &= C(s, t), \quad \forall s, t \in I, \end{aligned} \tag{5}$$

where $m(t)$ it the centerline and $\epsilon(t)$ is a centered Gaussian process with covariance function C. Algorithm 2 extends Algorithm 1 for this specific type of estimators, minimizing the already defined cost function.

Algorithm 2 Adjustment using multivariate estimators

```
Compute scatter estimate $\widehat{\Sigma}$ exploiting any of Ledoit-Wolf, OGK, MRCD, or kMRCD estimators
Compute Cholesky factor $chol(\widehat{\Sigma})$
Compute centerline as the curve of maximum depth
for $i \in 1, ..., N_{trials}$ do
    $\tilde{X}_i = n$ realizations of the process as per Equation (5) with $m(t)$ estimated by centerline and $C$ estimated by $\widehat{\Sigma}$
    $F_i = argmin_F c(F, \tilde{X}_i)$
end for
$F^* = \sum_{i=1}^{N_{trials}} \frac{F_i}{N_{trials}}$
Build the functional boxplot on original data using $F^*$
```

2.3. Functional Robust Covariance estimators

As mentioned earlier, due to the infinite-dimensional nature of functional data, it is necessary to extend the approach described in Section 2.2 to methods that directly leverage the mathematical structure of the curves. A frequently employed tool to deal with this complication is functional principal component analysis [10], which allows for the reduction of the dimensionality while retaining the most significant features in the process. As a straightforward extension of the vectorial case, functional principal directions are the eigenfunctions of the covariance operator. Recall from previous sections that our objective is to simulate Gaussian data, given location and scatter estimates, which imitates the behavior of a clean population generated from the original one. In this context, given a stochastic process X, the Karhunen-Loève decomposition [21] provides a mechanism of data generation starting from a known process:

$$X = \mu + \sum\nolimits_{i=1}^{\infty} \sqrt{\lambda_i}\zeta_i\phi_i \quad \zeta_i \sim \mathcal{N}(0,1), \tag{6}$$

where $\{\lambda_i, \phi i\}_{i=1,\dots,\infty}$ are the eigencouples of the Covariance function, and $\{\lambda_i\}_{i=1,\dots,\infty}$ are the eigenvalues in decreasing order of magnitude. This expansion can be truncated at L components thanks to the results from Boente et al. [2]. Indeed, for a random variable with elliptical distribution, the linear space spanned by the first L eigenfunctions provides the best L-dimensional approximation of the process in terms of residual squared norm. These are associated with the L largest eigenvalues, capturing most of the data variability, while the remaining are vanishing as we add more dimensions. In the following, we will assume that the stochastic process under study X admits an expansion with finitely many terms. Given this formulation, we solely need to obtain a sound estimate of the eigenspace of the population Covariance function. Hence, the goodness of the simulation-based adjustment depends upon theoretical results on the equivalence of the underlying Covariance spectrum with that of the employed Covariance Operator. The property that interests us is that the eigenfunctions $\{\phi_i\}_{i=1,\dots,L}$ of the estimator coincide with those of the sample Covariance matrix. Usually, we do not have any guarantee on the eigenvalues $\{\lambda_i\}_{i=1,\dots,L}$. Nevertheless, they can be easily estimated as the variance of the data X projected over the eigenfunctions $\{\phi_i\}_{i=1,\dots,L}$ (see Section 2.3.4 for details). Theorems proving such a result are available for all three functional estimators that we will be subsequently described in the following subsections.

2.3.1. Spherical Covariance Operator

The first functional estimator considered is the Spatial Sign Covariance Operator [9], also called *Spherical* Covariance Operator. The estimator corresponds to the sample covariance operator of the centered curves, via a location functional $\widetilde{\mu}$, projected onto the unit sphere:

$$\mathcal{C}_S = \mathbb{E}\left[\frac{(X-\widetilde{\mu})\otimes(X-\widetilde{\mu})}{\|X-\widetilde{\mu}\|^2}\right], \tag{7}$$

where $\otimes$ denotes the tensor product in the Hilbert space $\mathcal{F}$. The centerline $\widetilde{\mu}$ generally identifies the deepest point in the data cloud, employing an appropriate depth notion. Alternatively, the functional geometric median or the spatial median can be used. The latter is obtained as the solution to the problem

$$\widetilde{\mu} = \underset{z\in\mathcal{F}}{\arg\min}\ \mathbb{E}[\|X-z\|-\|X\|], \quad \mathbb{E}\left[\frac{X-\widetilde{\mu}}{\|X-\widetilde{\mu}\|}\right] = 0, \tag{8}$$

i.e., it is the curve such that the mean distance from it to all the points of the distribution of X is minimum. By centering our data X with respect to the spatial median $\widetilde{\mu}$ and normalizing it, we obtain $\widetilde{X} = (X-\widetilde{\mu})/\|X-\widetilde{\mu}\|$. Hence, the estimation can be drawn by computing the sample covariance of this new object $\widetilde{X}$. As per Equation (8), its expected value is null by definition. Thus, the computation of the Covariance reduces to the expected value given in Equation (7). Boente et al. in [9] provide proof for the equivalence of the eigenfunctions of $\mathcal{C}_S$ with those of the sample covariance. Also, the order in the eigenvalues is preserved. Consequently, we can make use of the spectrum of the Spatial Sign Covariance $\mathcal{C}_S$ to generate new samples through Equation (6). Further details concerning the computation of $\mathcal{C}_S$ are reported in Appendix A.

2.3.2. Median Covariation Operator

The next estimator considered for the robust estimation of the covariance operator is the Median Covariation Estimator introduced by Kraus and Panaretos [17]. It solves a more complex problem compared to the Spherical Covariance to suggest a median-type estimator of scatter:

$$\mathcal{C}^{\rho}_{\mathcal{M}} = \underset{\mathcal{M} \in HS}{\arg\min}\, \mathbb{E}\left[\rho\left(\|(X-\mu)\otimes(X-\mu)-\mathcal{M}\|_{HS}\right) - \rho\left(\|(X-\mu)\otimes(X-\mu)\|_{HS}\right)\right], \tag{9}$$

where ρ is a convex function and HS refers to the space of Hilbert-Schmidt operators over a space $\mathcal{F}$, such as L^2. Indeed, Equation (9) recalls the problem of the spatial median presented in Equation (8), with $(X-\mu)\otimes(X-\mu)$ in place of X. The idea relies on the fact that the sample covariance is a location estimate for the quantity $(X-\mu)\otimes(X-\mu)$, therefore we can generalize Equation (8) to get an estimate of the dispersion. See [17] for the assumptions on the existence and uniqueness of the defined quantity. In the following, we will employ $\rho(u) = u$ and $\mu = \widetilde{\mu}$ the spatial median as defined in Equation (8). Since no closed-form solution is available for this formulation of the problem, an iterative algorithm is employed: computational details are deferred to Appendix A.

2.3.3. Kendall's τ function

The last estimator taken into consideration is Kendall's τ function introduced by Zhong et al. in [36] for robust functional principal component analysis of non-Gaussian longitudinal data. Their version of Kendall's τ comes from the intuition behind Kendall's correlation coefficient [16] and the spatial sign covariance function [1], [9] introduced in Section 2.3.1:

$$K(s,t) = E\left[\frac{\{X(s)-\widetilde{X}(s)\}\{X(t)-\widetilde{X}(t)\}}{\|X-\widetilde{X}\|^2}\right], \tag{10}$$

where $\widetilde{X}$ is an independent copy of X. We can notice that it has a mathematical formulation similar to that in Equation (7). The difference is in the centering, which is not done by subtracting the spatial median $\widetilde{\mu}$ but using a duplicate of the realization of the same process inspired by the idea of correlation.

The equivalence of the eigenfunctions of Kendall's τ function with those of the sample covariance is proved in [36]. Even if we do not get an explicit estimate of the covariance, with this method we can directly estimate the eigenfunctions $\{\phi_i\}_{i=1,\dots,L}$ from the implementation provided in [36].

2.3.4. Adjustment procedure for functional estimators

This framework is more complex compared to the one presented in Section 2.2.1, as we aim at treating the functional dataset $\{X_1, \dots, X_N\}$ as a collection of continuous functions and not as simple multivariate vectors. That is, we aim at directly dealing with the object space without the need to resort to the feature space. For this reason, we need to associate the discrete observations with a geometry structure. This geometry defines a basis $\{\varphi_i\}_{i=1,\dots,\infty}$, for example, Fourier or B-splines, over a functional space such as L^2, H_1, etc. Hence, in case the user wants to build a functional boxplot using a functional covariation operator as a robust scatter estimate for the F adjustment procedure, the data needs to be at first projected over a lower dimensional space, with fixed dimensionality L depending on the amount of variability that we want to capture in the data. In our case, $L = 10$ is deemed to be enough to retain a substantial proportion of variance. Data is simulated using a truncated Karhunen-Loève generative model:

$$X = \mu + \sum_{i=1}^{L} \sqrt{\lambda_i}\zeta_i\phi_i \quad \zeta_i \sim \mathcal{N}(0,1), \tag{11}$$

where $\{\lambda_i, \phi_i\}_{i=1,\dots,L}$ are the first L eigencouples of the Covariance function. New functional observations can be generated by sampling $\zeta_i \sim \mathcal{N}(0,1)$. As mentioned in Section 2.3, while the eigenfunctions coincide for all the estimators defined in the previous sections, the eigenvalues must be approximated as the dispersion of our data X projected over the space of the eigenfunctions $\{\phi_i\}_{i=1,\dots,L}$. To do so, the robust estimator of scale Q_n is used for its breakdown properties [29]. However, this requires the tuning of a distribution-specific constant, which is not known since the original data does not satisfy any particular assumption. Thanks to the property of location-scale invariance of the Modified Band Depths [20], we can claim that translation and scale transformations of the quantities defined in Equation (11) do not influence the resulting depths. The $\alpha 100\%$ central region of X and its transformed version $X^* = \sqrt{\lambda_1}X + \mu$ are related by $C_\alpha(X^*) = \sqrt{\lambda_1}C_\alpha(X) + \mu$, hence $P_{X^*}(X^* \in FC_\alpha(X^*)) = P_X(X \in FC_\alpha(X))$. As a consequence, the width of the fences is simply rescaled while the value of the inflation factor F stays the same. It follows that we can rewrite Equation (11) subtracting the mean process and dividing by the first eigenvalue as follows:

$$X^* = \frac{(X-\mu)}{\sqrt{\lambda_1}} = \sum_{i=1}^{L} \sqrt{\frac{\lambda_i}{\lambda_1}} \zeta_i \phi_i, \quad \zeta_i \overset{i.i.d.}{\sim} \mathcal{N}(0,1). \tag{12}$$

Dividing by λ_1 allows us to estimate the ratio λ_l/λ_1 with $Q_n(\phi_l)/Q_n(\phi_1)$, thus avoiding the tuning of the distribution-specific constant that the Q_n estimator depends upon since it cancels out in the so-defined ratio. This is done at lines 6-12 of Algorithm 3, where the entire pipeline for the Adjustment procedure for functional estimators is reported in pseudo-code.

Algorithm 3 Adjustment using functional estimators

```
Set L
Define the geometry, given a basis $\{\varphi_j\}_{j=1,...,L}$
Project multivariate data over the basis
Compute scatter estimate in the basis $\widehat{\Sigma}$ with Spherical, Median or Kendall estimator
Estimate the eigenfunctions $\widehat{\Phi} = [\widehat{\phi}_1, ..., \widehat{\phi}_L]$
for $l \in 1, ..., L$ do
    for $j \in 1, ..., N$ do
        $p_{l,j} = \prod_{\widehat{\phi}_l} X_j$
    end for
    Compute robust estimate of variance of projected data: $q_l = Q_n(p_l)$
    $\rho_l = q_l / q_1$
end for
for $i \in 1, ..., N_{trials}$ do
    for $k \in 1, ..., n$ do
        $\tilde{X}_{i,k} = \sum_{l=1}^{L} \rho_l \tau_{k,l} \phi_l$ with $\tau_{k,l} \sim \mathcal{N}(0, 1)$
    end for
    Add geometry structure to the simulated data
    $F_i = argmin_F c(F, \tilde{X}_i)$
end for
$F^* = \sum_{i=1}^{N_{trials}} \frac{F_i}{N_{trials}}$
Build the functional boxplot on original data using $F^*$
```

Given all the algorithms introduced above, we aim at validating their performance in the tuning factor adjustment for the functional boxplot via a simulation study to observe the behavior of the various estimators under several contamination conditions.

3. Simulation study

This section is dedicated to the investigation of the empirical performance of the estimators discussed in Section 2 for the tuning of F in the functional boxplot. To understand how this can be affected by the presence of atypical curves, we propose one driving simulation study on the most commonly observed outlying behavior, that is amplitude outliers. All the simulations were run in parallel on a computer cluster with 32 processors Intel Xeon E5-4610 v2 @2.3GHz. This much computational power is needed specifically due to Kendall's τ function, whose implementation is particularly heavy in terms of memory, as we will see when discussing the results of our study. For each combination of the parameters below described $B = 50$ repetitions of the simulated experiment have been considered: results are reported in the next section.

3.1. Data generation and contamination

The experiment is carried out over samples in $\mathbb{R}^{N \times P}$, where N is fixed at 100 curves, while P assumes values in $\{200, 400\}$. This framework allows us to infer conclusions in the *small N, large P* context, which is the standard in FDA. The data-generating process is based on the spectrum of the Exponential Covariance matrix $C(s,t) = \alpha e^{-\beta|s-t|}$, with $s, t \in I = [0,1]$ and $\alpha = 0.12$, $\beta = 0.4$ leading to contained variability around the mean and high autocorrelation (non-noisy data). The custom basis for the dimensionality reduction to $L < P$ is built upon its eigenfunctions, L is set to 10, and the continuous space chosen is L^2. The generating process is Gaussian, with the covariance matrix having on its diagonal the eigenvalues and mean generating process defined as $\mu = sin(4\pi t)$. The datasets are then contaminated by replacing a proportion of observations with outliers. The fraction of corrupted data is selected to be 0% (clean population), 5%, 10%, 15%. We consider data contamination of the type shown in Figure 4. Here, a proportion of curves is inflated by multiplying the mean function coefficients by a random number $u \sim U(2,3)$. In detail, the data-generating mechanism for the clean and outlying data are respectively defined as follows:

$$\begin{aligned} X_i(t) &= sin(4\pi t) + \epsilon_i(t), \quad t \in [0,1], \\ X_i^{out}(t) &= sin(4\pi t) \times u + \epsilon_i(t), \quad u \sim U(2,3),\ t \in [0,1]. \end{aligned} \tag{13}$$

with $\epsilon_i(t)$ denoting a centered Gaussian process with covariance function $C(s,t)$.

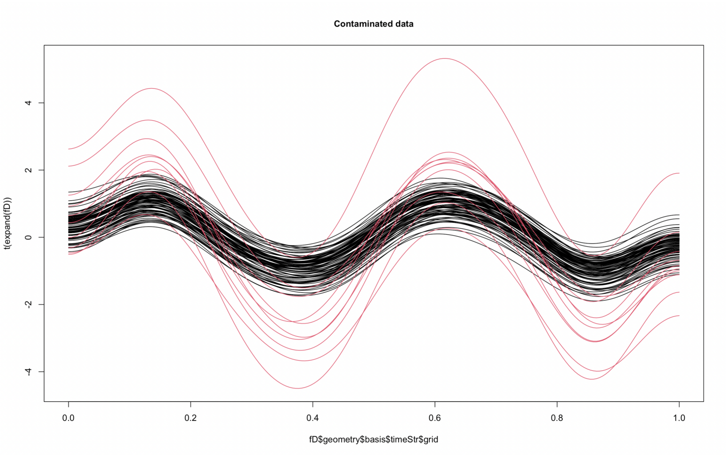


Figure 4: Simulated data with mean process $\mu = sin(4\pi t)$ and inflation of 5% of the curves by $u \sim U(2,3)$

3.2. Evaluation and discussion of results

In this subsection, we discuss the results obtained by running a simulation study using the data-generating mechanism presented in Section 2.3.1. We compute the False Positive Rate (FPR) and True Positive Rate (TPR, also named Recall) as metrics for the evaluation of the methods performance. The former is the fraction of cases that are wrongly flagged as outliers, among all the genuine observations. The latter, instead, is the fraction of uncontaminated cases that are correctly identified as such, among those that actually are. Ideally, we want to have TPR close to 1 and FPR close to 0. These values are related to the tuned inflation factor F^*: generally, the higher the estimated F^*, the higher the probability of missing an anomalous curve; the smaller the F^*, the higher the probability of observing a swamping effect. As already been observed in the literature [32] and justified in Section 2, the value $F = 1.5$ has been proved to be too large for functional data, which present a high correlation over time and/or space. Therefore, we expect our methods to end up with smaller values of the inflation factor, which will lead to a higher probability of identifying atypical observations, paying a price in terms of wrongly flagging some normal samples as to be outliers. Depending on the requirements of the application at hand, one might prefer low Precision (if there are many false positives) to get a high Recall (if there are very few false negatives), or vice versa. The first framework is very common in healthcare applications, where medical doctors would prefer to run an additional health screening for some uncertain subjects, instead of wrongly labeling a sick being as healthy.

For each combination of data dimensionality and outliers proportion, we observe the results for the considered estimators. Notice that MRCD is included in the cases $\alpha = 0.5$ and $\alpha = 0.75$ (see Section 2.2.2), to understand how this hyper-parameter could influence the final tuning. Something similar is done with OGK, for which we use both the Q_n and the MAD scale estimates. The situation in which the functional boxplot is built without making use of any covariance estimator to tune the inflation factor F is referred in the following as the "*No adjustment*" case. Also note that for the last robust estimator introduced in Section 2.3, Kendall's τ function, some bandwidth hyperparameters are needed in the evaluation. The R package presented in [36] selects them through generalized cross-validation (GCV). However, due to running time issues, we set `bwK = 0.045, bwmean = 0.03`, which are the optima found by GCV for a particular set of instances generated from the process in Equation (13). We first look at the empirical distribution of F^* over the repeated trials. In Figure 5 we can notice that functional operators behave similarly. Kendall's τ function tends towards higher values of F for higher outliers proportions, with respect to Median and Spherical. The cases $\alpha = 0.5$ and $\alpha = 0.75$ does not seem to make a difference when employing MRCD for the scatter estimate. Also, the different scale estimators in OGK, Q_n and MAD produce equivalent results. MRCD and kMRCD happen to lead to $F^* > 1.5$ in the case of a clean population. Such behavior is understandable when no outliers are present, as the resulting fences will be as wide as possible, leading to very few False Positives. kMRCD showcases the same behavior also when the proportion of outlying curves in the population increases, while MRCD tends to have smaller F^* values. Ledoit-Wolf becomes more biased as the number of outliers grows since its estimate is influenced by their presence. F^* values obtained by tuning the functional boxplot via OGK and Ledoit-Wolf estimators are always below the default value of 1.5. This behavior will lead to many False Positives, i.e., curves wrongly flagged as outliers. Moreover, the OGK estimate becomes more uncertain in the case of high dimensionality ($P = 400$) and high contamination (outliers proportion: 0.15).

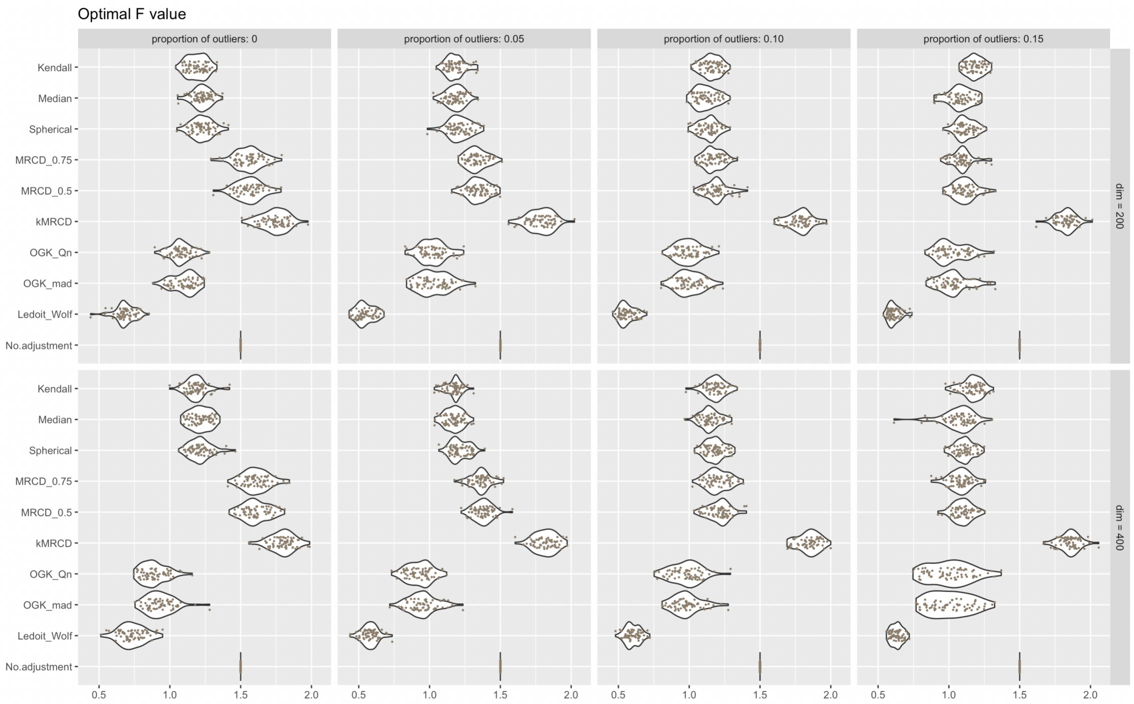


Figure 5: Violin plots of the optimal F values tuned with the adjustment procedures in Algorithms 2 and 3 using each one of the estimators investigated.

From these results, functional operators seem to have the most stable and robust behavior when it comes to the identification of the optimal F^* value.

The True Positive rate, whose violin plots for varying proportions of outliers and data dimensionality are displayed in Figure 6, gives a satisfying performance by the majority of the estimators. The same does not hold true for kMRCD, whose distribution reveals some difficulty in flagging the samples: many atypical observations are not identified as such. This is coherent with the high values of F^* showcased in Figure 5. Regarding the case in which F^* is fixed to 1.5, we notice an increasing difficulty in correctly identifying all outliers as their proportion increases, especially in the higher dimensional case (lower panel), which justifies the need for an adjustment of the inflation factor F adapted to the specific distribution of the data under study.

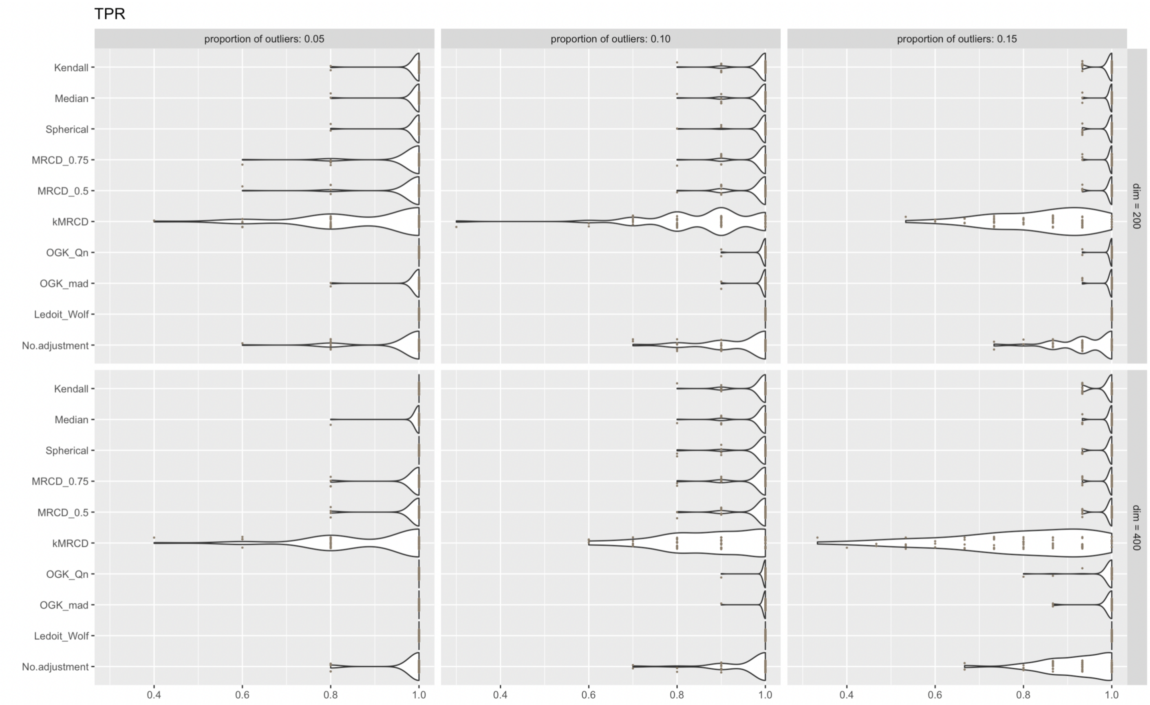


Figure 6: Violin plots of the TPR metric using each one of the estimators investigated.

In Figure 7, we display the False Positive rate. It ranges from 0 to a maximum of 0.3. It has an elongated distribution for the non-robust benchmark Ledoit-Wolf estimator, as we would have expected considering the previous plots. We can see that the currently employed estimator in the functional boxplot, OGK, also presents a more dispersed distribution for any case with respect to the newly presented ones, especially for $P = 400$. The kernel version of MRCD (kMRCD) is the best-behaving one when monitoring the FPR, but, as we have seen before, its performance is poor when it comes to identifying the inflated samples.

As mentioned in Section 2.2.3, we are interested in assessing the performance of the considered estimators in terms of computational efficiency and scalability. This task is very hard to accomplish in an absolute sense, as extensively discussed in [18], due to the inconsistencies encountered in different implementations of the same algorithm and/or the diverse levels of optimization achieved for the methods. Consequently, we limit our attention to the currently available versions of the algorithms in R for each of the estimators presented in Sections 2.2 and 2.3 and draw conclusions that are not independent on the type of implementation and software used.

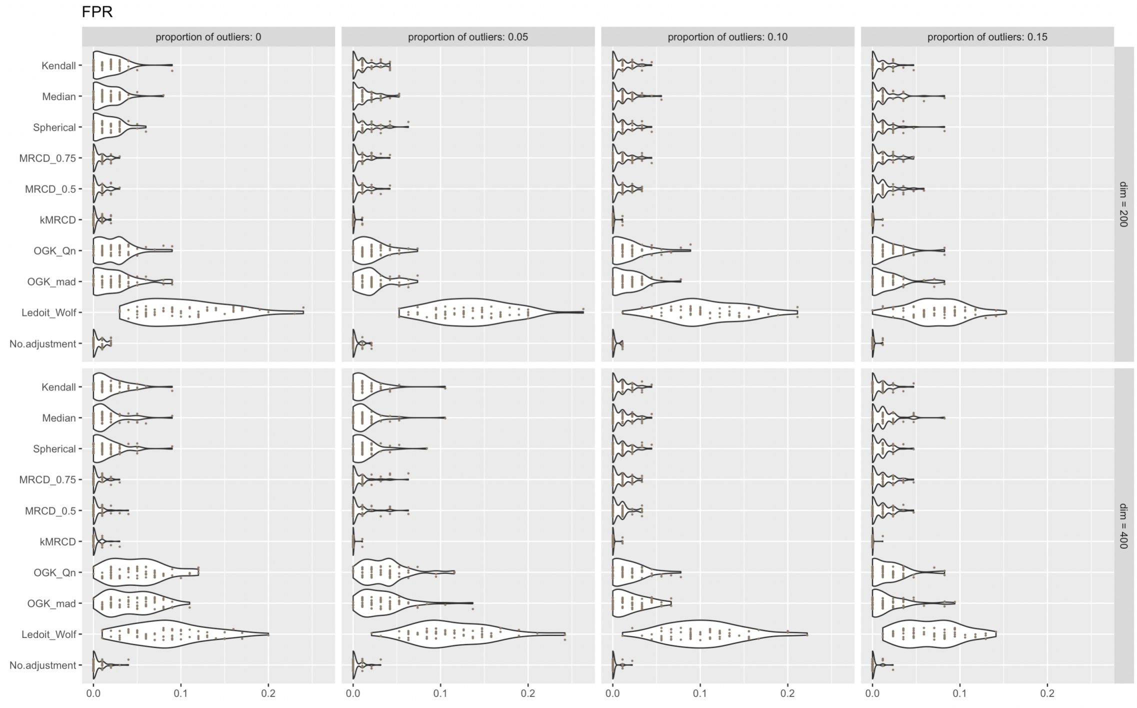


Figure 7: Violin plots of the FPR metric using each one of the estimators investigated.

Given the difference in scale in the empirical distribution of the computational time experienced in the simulations between Kendall's τ function and the remaining estimators, the computational time, graphically reported in Figure 8, is in log-scale. It appears clear that the time required for the adjustment of the inflation factor is comparable among all the estimators but Kendall's τ function. Indeed, to get the estimate of the eigenfunctions from the latter, high computational power and resources are required, especially when the dimension is high.

Some interesting differences can be highlighted. OGK is in any case the slowest among all methods and is very influenced by the dimensionality of the data. As we were expecting, kMRCD is faster than the non-kernelized version since its computational complexity depends on N instead of P, despite the additional computation due to the refinement step. This gap is more evident in the high-dimensional case. Given that the target matrix employed in kMRCD does not take into account correlation over the horizontal axis, as mentioned in Section 2.2.3, it is possible that extending the kernel version to this framework could lead to improved results. Overall, the Spherical and Median Covariation Operators are the most efficient, as they are the least affected by the dimensionality of the data.

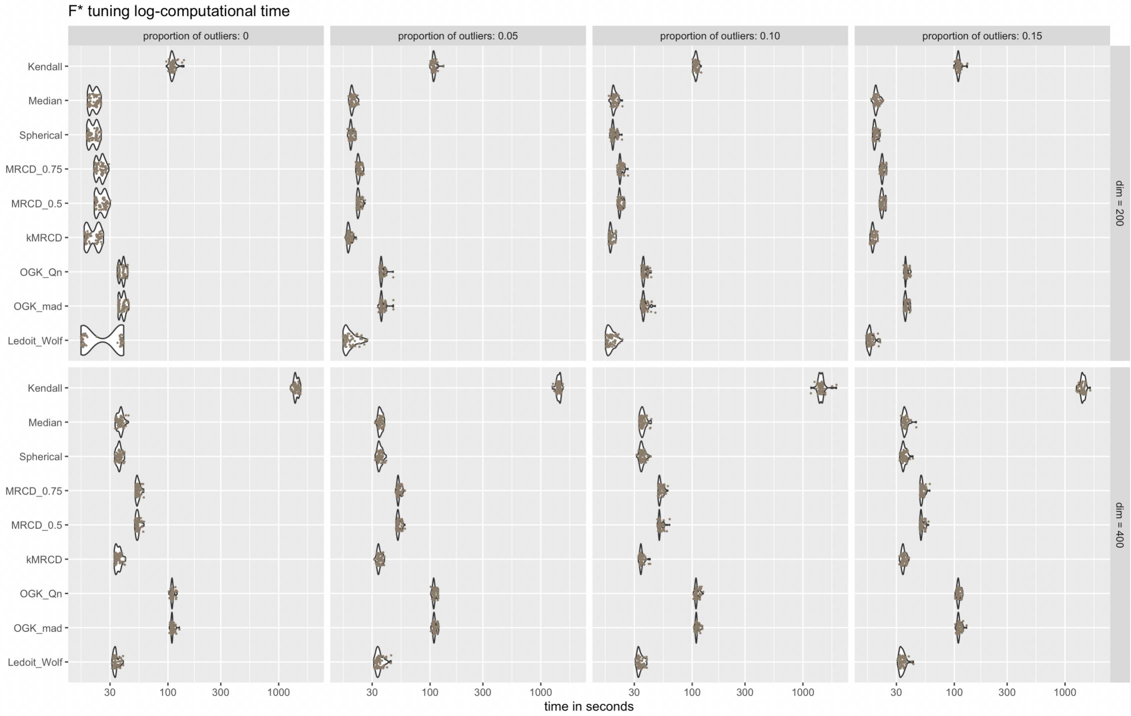


Figure 8: Boxplots of the Log-Computational time of all estimators to perform the adjustment in the different simulation settings.

To wrap up, in a scenario where the data is contaminated by amplitude outliers functional operators seem to provide the right balance between identifying the true anomalous curves whilst not producing too many false positives. In particular, the Spherical and Median Covariation Estimators bring similar performance in a satisfactory amount of time and resources, while Kendall's τ scalability seems to be heavily influenced by growth in dimensionality. Other simulated scenarios that may be encountered in functional outlier detection are reported in Appendix B.

4. Application on real data

This section is devoted to the illustration of the methodologies introduced in Section 2 on a real data example. The Electrocardiographic (ECG) data set, available in the `roahd` CRAN package, was collected for the PROMETEO (PROgetto sull'area Milanese Elettrocardiogrammi Teletrasferiti dall'Extra Ospedaliero) project, aimed at spreading the intensive use of ECGs as a pre-hospital diagnostic tool. The database comprehends eight leads I, II, V1, V2, V3, V4, V5, and V6 for every statistical unit, each one describing the heart dynamics of the patient. As our methods have been developed to handle univariate data, the analysis will be led on one of them. A review of the literature on the importance of the leads in ECG interpretation reveals no clear predominance of one over the others (e.g., see [35]). We highlight that the developed procedure is reproducible on any of the leads, and we report the results employing the first one. The signals have been registered and smoothed over an evenly spaced grid of 1024 time points at 1kHz. The registration landmark-based procedure, outlined in Ieva et al. [15], identifies as landmarks those time points that can be associated with a specific biological event.

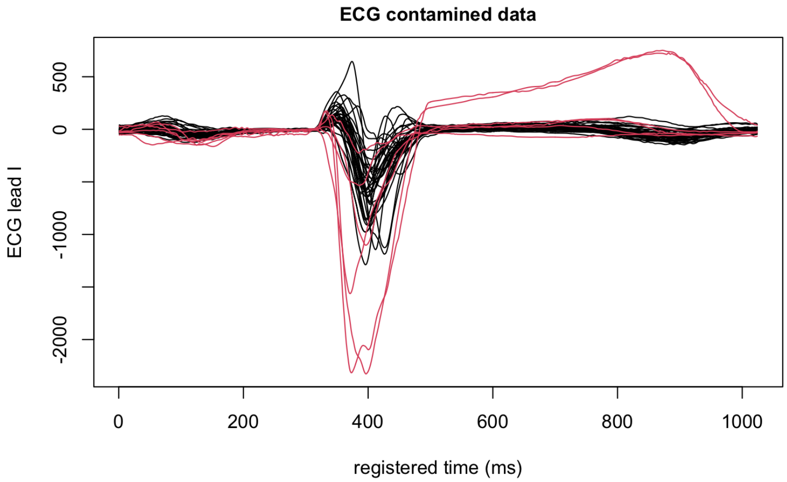


Figure 9: ECG healthy data contamined with LBBB data

This allows us to separate amplitude variability from phase variability: the duration of each ECG interval will not influence the final estimates. To mimic the outlier detection process, we randomly sampled $N_1 = 34$ ECG traces for healthy patients and $N_2 = 6$ for subjects suffering from a cardiac pathology called Left-Bundle-Branch-Block (LBBB), leading to a 15% contamination in the resulting dataset. Thus, the total number of units $N = 40$ is far less than the dimension $P = 1024$. Figure 9 shows in black the *physiological* signals and in red the curves for which Left-Bundle-Branch-Block is diagnosed. Our objective is to detect pathological ECG traces. Within this example, it is clearly of paramount importance to effectively identify patients with the disease, thus accepting some False Positives. This means being able to diagnose illness in more subjects, to further analyze the patient-specific situation. To this aim, we applied each one of our proposals for tuning F and obtained the results shown in table 1. Notice that the case in which no adjustment is performed, displayed in Figure 10, is only able to capture 67% of the actual outlying samples. The multivariate approaches presented in Section 2.2 improve the TPR metric while suffering more on the FPR side.

	No adjustment	**Ledoit-Wolf**	**OGK_Qn**	**MRCD**	**Spherical**	**Median**
TPR	0.67	0.833	0.833	0.833	1	1
FPR	0	0.147	0.147	0.205	0.235	0.294
Time [m]	0.005	0.605	4.163	1.986	12.900	12.750

Table 1: True Positive Rate, False Positive Rate and Computational time in minutes of each estimator for the data at Figure 9.

The functional techniques seen in Section 2.3 lead to the complete recognition of patients affected by LBBB. Figure 11 shows the functional boxplot which results from our implementation employing the Spherical Covariance Operator, which is the best performing to our aim, as it represents the best trade-off between TPR and FPR. This procedure would allow for the quick identification of the most concerning ECG leads and raise a warning for the most alarming ones, thus allowing for more efficient diagnosis and treatment of the condition. This means that, among the negative cases, 23% will undergo a further health check even if they are not affected by the disease. At any rate, atypical behavior in the ECG curve of healthy patients may be a premonitory signal of an upcoming hearth-related pathology which is best to submit to physicians and experts in the domain for further evaluation.

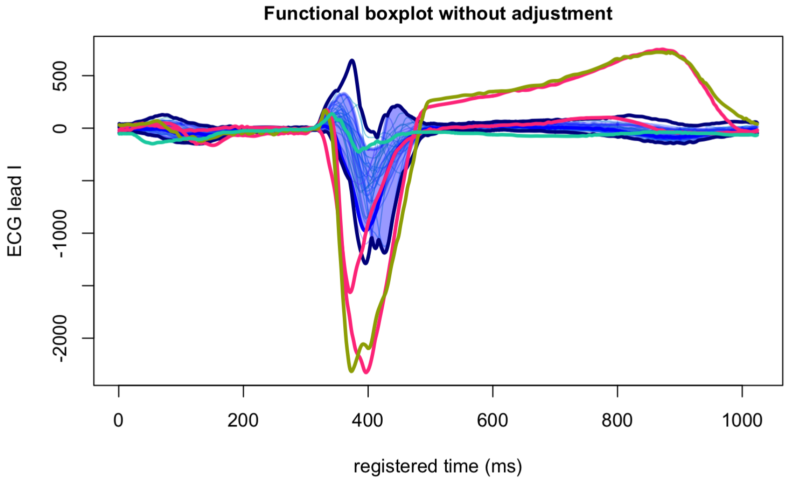


Figure 10: Anomaly detection of ECG unhealthy signals without carrying out the adjustment of the inflation factor F.

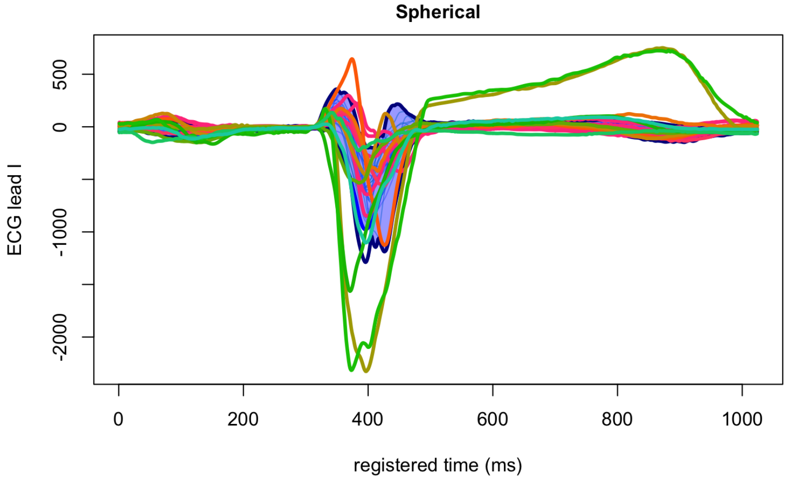


Figure 11: Anomaly detection of ECG unhealthy signals employing the Spherical Covariance Operator for the simulation-based tuning of the inflation factor F.

The functional boxplots resulting from the usage of the other estimators on the ECG data are reported in Appendix C.

5. Conclusions and future developments

This paper has focused on the crucial challenge of detecting outliers in functional data. Building upon the well-established functional boxplot methodology, we have extended the simulation-based adjustment technique initially proposed in [32] to enhance the capability of outlier detection. Particularly, we have concentrated on motivating the significance of performing a distribution-free adjustment of the inflation factor F by repeatedly simulating some datasets of Gaussian functional observations with the same mean and covariance as the original dataset, but not influenced by the anomalous observations that we want to spot. Making use of several robust estimators of location and scatter we compared the performance of multivariate procedures and functional operators for implementing the tuning process. Through a comprehensive simulation study, we demonstrated the superiority of our method over the original proposal. Lastly, a favorable application in the healthcare field has cast light on the promising usefulness of such an approach in the nonparametric inference for vital signs.

As possible direction for future research one can consider the exploration of a variety of depth measures to set up a flexible procedure that behaves differently depending on the type of outliers that need to be identified. Another possible development is the generalization of the proposed procedure to the multivariate functional case. While functional boxplots for multivariate curves have been recently introduced in the literature [7], [27], the generalization of the tuning procedure to this framework remains an open issue. Some proposals are currently being explored and they will be the object of future studies.

Appendix A. Computational details for Spherical Covariance and Median Covariation operators

Spherical Covariance Operator

The Spherical Covariance Operator can be implemented following the approach in Algorithm 4. The function takes in input the functional data already projected over $L < P$ basis elements. The basis can be chosen by the user over a space $\mathcal{F}$ defining the geometry of the data object. The Mass Matrix or Gram Matrix $M \in \mathbb{R}^{L\times L}$ represents the projection matrix over the space $\mathcal{F}$, by means of the basis functions $\{\varphi_j\}_{j=1,\dots,L}$. The element k, j of M is the scalar product in the space $\mathcal{F}$ of the corresponding basis functions.

Algorithm 4 Spherical Covariance

```
Input: Functional data $\{X_1, ..., X_N\}$ represented in the basis with L components
Compute median $\widetilde{\mu} \in \mathbb{R}^L$
for $i \in 1, ..., N$ do
    Standardize wrt median according to the geometry $[M]_{k,j} = \langle \varphi_k, \varphi_j \rangle_{\mathcal{F}}$
    $\widetilde{X}_i = (X_i - \widetilde{\mu})/\sqrt{(X_i - \widetilde{\mu})^T M (X_i - \widetilde{\mu})}$
end for
Compute Sample Covariance $\mathcal{C}_S = Cov(\widetilde{X})$
```

Median Covariation Operator

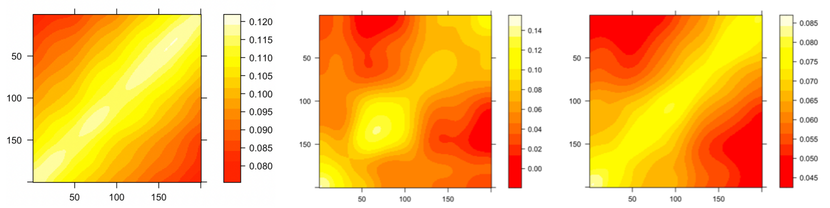
The Median Covariation Operator can be implemented by means of the Averaged Stochastic Gradient (ASG) optimization procedure [4]. It requires a sequence of learning weights $\gamma_n = c/(\max\{n-1, 1\})^\alpha$, decreasing with the number of iterations n, to allow faster convergence to the optimum. The first equation in (A.1) defines the stochastic gradient step since the direction on the right-hand side is an approximation of the gradient of the functional to be minimized in Equation (9), with respect to $\mathcal{M}$. We will call M_n the value of $\mathcal{M}$ at iteration n. At each iteration, we average the newly updated value M_{n+1} with the previous mean $\bar{M}_n$. This is an efficient modification that removes the need to store the value M_{n+1} at each step.

$$\begin{aligned} M_{n+1} &= M_n + \gamma_n \frac{(X_{n+1} - \widetilde{\mu})(X_{n+1} - \widetilde{\mu})^T - M_n}{\left\| (X_{n+1} - \widetilde{\mu})(X_{n+1} - \widetilde{\mu})^T - M_n \right\|_F}, \\ \bar{M}_{n+1} &= \bar{M}_n - \frac{1}{n+1}\left(\bar{M}_n - M_{n+1}\right) \end{aligned} \tag{A.1}$$

The pseudo-code to compute $\mathcal{C}_M$ following this reasoning can be found in Algorithm 5. Differently from the initially developed algorithm described in [17], our proposal makes use of a non-negative modification for the update step at line 7 in Algorithm 5, as suggested by Cardot et al. in [5]. As explained in the latter, M_{n+1} could not be positive semi-definite when the ratio $\gamma_n / \left\| (X_{n+1} - \widetilde{\mu})(X_{n+1} - \widetilde{\mu})^T - M_n \right\|_F \leq 1$. To adjust for this case, we simply truncate the learning rate at 1 when the ratio would actually be smaller.

Some trials have been conducted on a set of functional data with known covariance structure, to highlight the contrast between these two approaches.

Figure A.12 proves that the non-negative modification brings a consistent improvement in the approximation.



(a) Sample Covariance.

(b) Median Covariation.

(c) Median Covariation with non-negative modification.

Figure A.12: Differences of estimation of the Covariance matrix with and without non-negative adjustment.

As before, the function takes in input the data already in the reduced dimensionality L. Some default values for the hyperparameters are suggested in [17]. Notice that in this case, the norm is in the Frobenius sense: this is different from a row-wise norm as it is computed as the sum of the norms over rows. Also for this estimator, proof of the correspondence with the eigenfunctions of the underlying covariance structure is given in [5].

Algorithm 5 Median Covariance

```
Set $\alpha \in (0.5, 1)$ and $c > 0$, default $\alpha = 3/4, c = 2$
Compute median $\widetilde{\mu} \in \mathbb{R}^L$
Initialize $M = \mathbf{0}, \overline{M} = \mathbf{0} \in \mathbb{R}^{L \times L}$
for $i \in 1, ..., N$ do
    Compute $\gamma_i = c/(\max\{i-1, 1\})^\alpha$
    Compute $T = (X_i - \widetilde{\mu})^T (X_i - \widetilde{\mu})$
    Average Stochastic Gradient step:
    $M = M + (T - M)\ min(1, \frac{\gamma_i}{\|T-M\|_F})$
    $\bar{M} = \bar{M} - (\bar{M} - M)/i$
end for
$\mathcal{C}_M = \bar{M}$
```

Appendix B. Further simulation settings

In this appendix, we account for more contamination setups to better explore the capabilities of the proposed methods and understand their limitations. In the taxonomy study by Hubert at el. [13], anomalous functional observations are categorized into isolated, amplitude, shift and shape outliers. Isolated outliers contain a spike or peak in a limited interval over the domain. The remaining ones are designated as persistent outliers since the atypical behavior is all over the domain. Shift outliers are generated from the same process assumed for the genuine curves, but are moved away from the bulk of the data. They are of no interest in our application since they can be easily treated by means of registration. Shape outliers, generally speaking, are curves presenting a different structure from the majority of the samples. Lastly, amplitude outliers, have already been in-depth discussed in Section 3.1 as they are the main focus of the present article. Hereafter we will briefly comment on the impact that shape outliers (Figure B.13), amplitude outliers of various intensities (Figure B.14), and isolated outliers (Figure B.16) produce on the devised procedure.

Shape outliers can be constructed by defining a new mean trend, e.g., $\widetilde{\mu} = cos(4\pi t)$, for the outlying curves:

$$\begin{aligned} X_i(t) &= sin(4\pi t) + \epsilon_i(t), \quad t \in [0,1], \\ X_i^{out}(t) &= cos(4\pi t) + \epsilon_i(t), \quad t \in [0,1]. \end{aligned} \tag{B.1}$$

Figure B.13 displays the resulting behavior. This case study led to very similar outcomes to those discussed in Section 3.2 for amplitude outliers.

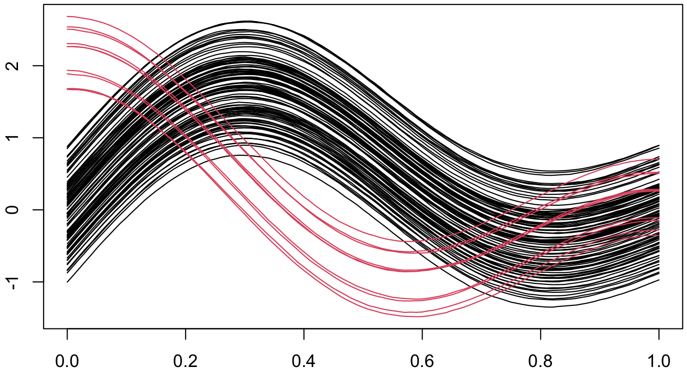


Figure B.13: Simulated data with mean process $\mu = sin(4\pi t)$ and outlying observations coming from $\widetilde{\mu} = cos(4\pi t)$. The covariance structure is the same for both genuine and anomalous curves.

For defining the remaining two contamination settings, we will make use of the library `fdaoutlier` introduced by Ojo et al. in [26] where several techniques for detecting functional outliers are presented. The authors provide the implementation of some convenience functions for the generation of contaminated datasets of functional data, which have been useful for testing the performance of our methods under diverse scenarios. Figure B.14 shows the case in which our data is contaminated by amplitude outliers of different intensities.

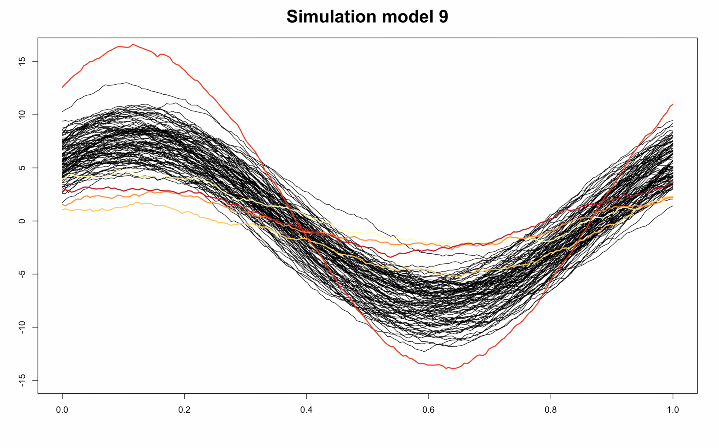


Figure B.14: Sample of Amplitude outliers of different intensities from `fdaoutlier` R package.

In details, the data generating mechanism is given as follows:

$$\begin{aligned} X_i(t) &= a_{1i} \sin \pi + a_{2i} \cos \pi + e_i(t), \\ X_i^{out}(t) &= (b_{1i} \sin \pi + b_{2i} \cos \pi)\,(1 - u_i) + (c_{1i} \sin \pi + c_{2i} \cos \pi)\, u_i + e_i(t), \end{aligned} \tag{B.2}$$

with $t \in [0,1]$, a_{1i}, a_{2i} following a Uniform distribution over an interval $[a_1, a_2]$, b_{1i}, b_{2i} following a Uniform distribution over an interval $[b_1, b_2]$, and c_{1i}, c_{2i} following a Uniform distribution over an interval $[c_1, c_2]$. In the considered case, $[a_1, a_2] = (3, 8)$, $[b_1, b_2] = (1.5, 2.5)$, and $[c_1, c_2] = (9, 10.5)$. Instead, u_i follows the Bernoulli distribution with $p = 0.5$. The covariance structure of $e_i(t)$ is of the same type as previously defined in Section 3.1, with $\alpha = 1, \beta = 1$. The outcome of this analysis is summarized in Figure B.15. We can observe the trade-off between TPR and FPR for each estimator, represented in different colors. The size of the dots is proportional to the uncertainty on the FPR metric. Focusing on the panels where there is a non-zero percentage of contamination in the data, it is clear that not performing an adjustment will never lead to the identification of the atypical observations (notice the very small and bright yellow dot on the bottom-left of each plot). The non-robust Ledoit-Wolf estimator has the same tendency as in the other scenarios, considering many samples as outlying thus having both high TPR and FPR. MRCD seems to be the best compromise between identifying outliers whilst not overestimating their presence in the dataset. Indeed, all the other estimators struggle in identifying anomalous curves, as they lead to higher values of the inflation factor F.

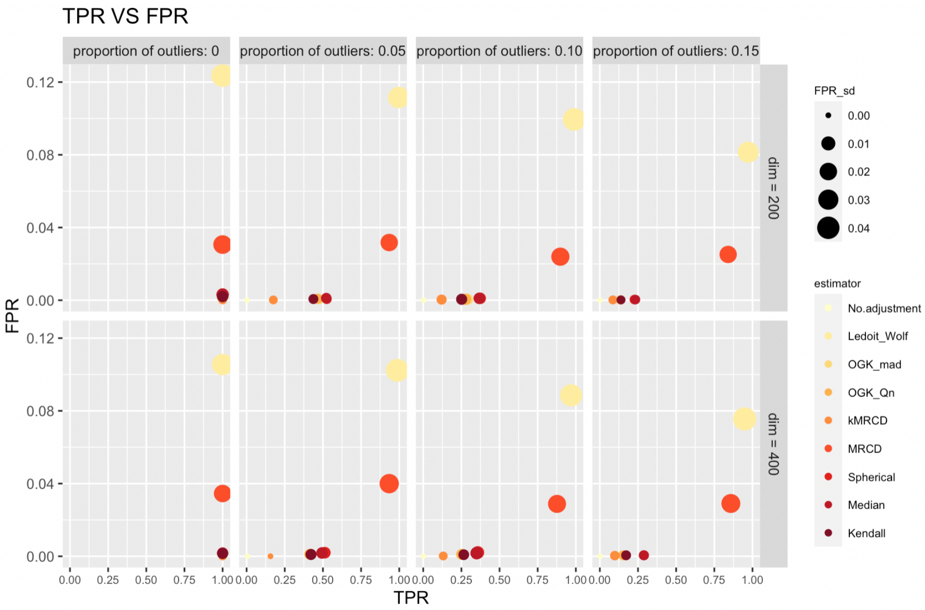


Figure B.15: Representation of True Positive Rate VS False Positive Rate average results for simulations run from Equation (B.2).

The last analysis is carried out employing magnitude-isolated outliers. The contamination process reads as follows

$$\begin{aligned} X_i(t) &= \mu t + e_i(t), \\ X_i^{out}(t) &= \mu t + q k_i I_{T_i \leq t \leq T_i + l} + e_i(t), \end{aligned} \tag{B.3}$$

with $t \in [0, 1]$, $k_i \in \{-1, 1\}$ and $P(k_i = -1) = P(k_i = 1) = 0.5$, while $q = 8$ defines the height of the peak. The constant l is set to 0.05 and defines the proportion of the interval over which the observation deviates from the majority of observations in the sample. The covariance structure of $e_i(t)$ is still the same as formerly described. An example of the so-devised data-generating process is graphically displayed in Figure B.16.

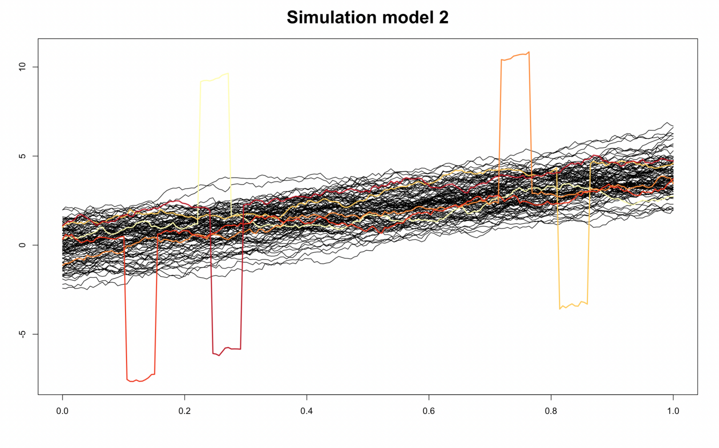
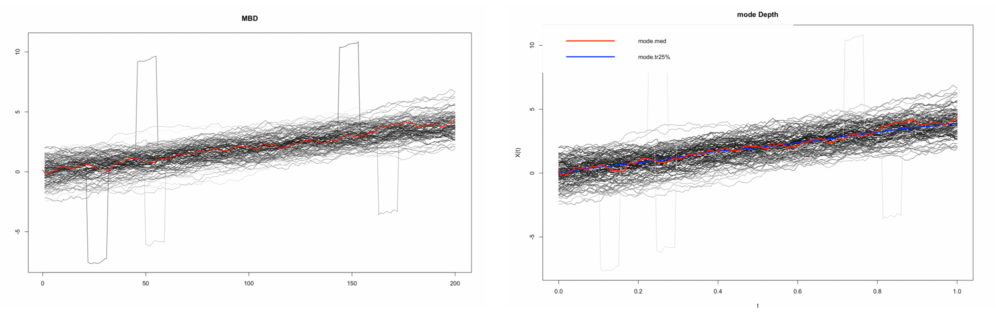


Figure B.16: Sample of magnitude isolated outliers from `fdaoutlier` R package.

This study led to unexpected results. Since these samples have a noticeable anomalous behavior, we were expecting even the simplest of the proposed methods to effectively identify the majority of them. Surprisingly, no matter the adjustment procedure implemented, we were not able to flag all outlying observations as such, and also many false positives arose. Due to this unforeseen outcome, we carried out a deeper examination of the depths used in the functional boxplot, which turned out to be the cause of the somewhat unexpected issue. In detail, this result is coherent with the definition of Modified Band Depth and their interpretation: the *normality* of a sample is judged based on the *amount of time* spent inside the band defined by all combinations of other two curves. It is thus expected that curves showcasing an outlying behavior only for a limited proportion of the time domain are not effectively identified as such. A more appropriate definition of depth could be employed, as the Modal Depth introduced by Cuevas et al. in [6]. The idea is that the depth of a curve is computed as a function of the number of curves in its neighborhood. Figure B.17 shows one realization of the contamination process outlined above, with a grayscale palette going from very light (curves of low depth) to very dark (curves of high depth).

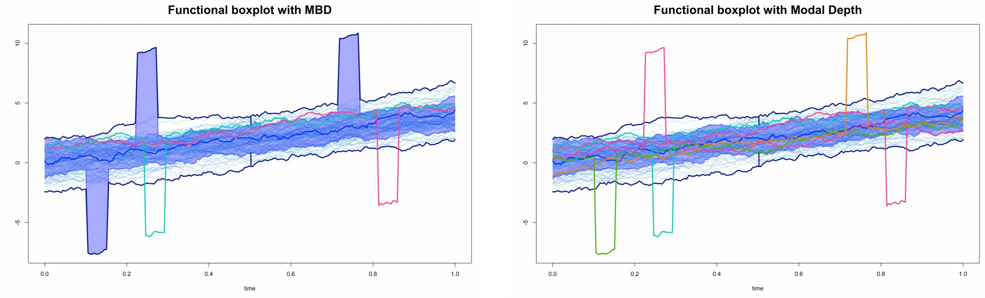


(a) MBD grayscale plot.

(b) Modal Depth grayscale plot.

Figure B.17: Comparison of depths measured by MBD (a) and Modal Depth (b).

It is clear from Figure B.17(b) that the Modal depth can recognize all isolated outliers as those which deviate significantly from the general trend, even if they spend most of their time inside the boundaries of the mass of functions, while MBD struggles in this task. Indeed, some of the spikes in Figure B.17(a) are marked in darker colors.



(a) Functional boxplot using Modified Band Depth.

(b) Functional boxplot using Modal Depth.

Figure B.18: Non-adjusted functional boxplot comparison by using: MBD (a) and Modal Depth (b).

The functional boxplot implementation allows for the employment of user-defined Depth functions. Two functional boxplots which employ respectively MBD and Mode depths are showcased in Figure B.18 to support our argument. This analysis is out of the scope of the present article but classifies as a possible further development of the adjusted functional boxplot.

Appendix C. Additional figures

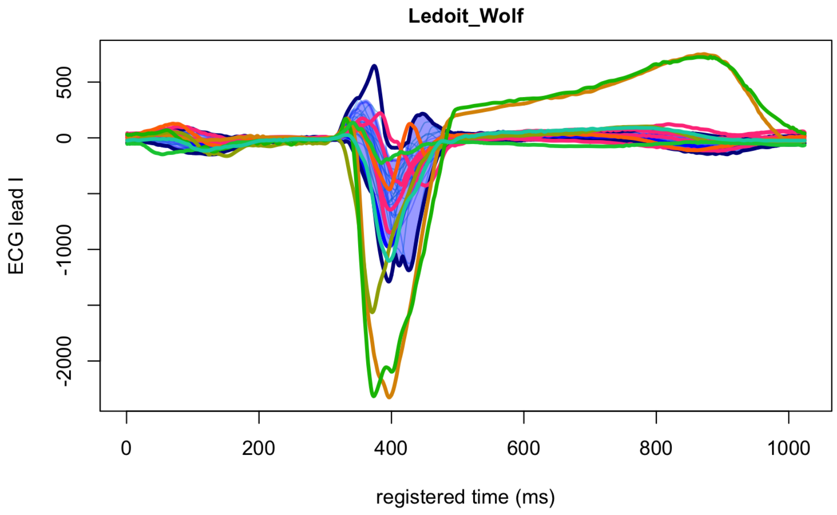


Figure C.19: Anomaly detection of ECG unhealthy signals employing the Ledoit-Wolf estimator for the simulation-based tuning of the inflation factor F.

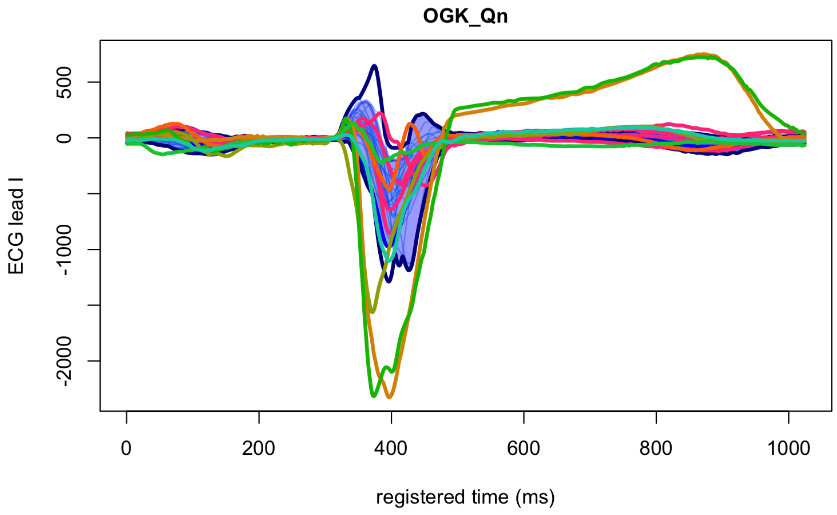


Figure C.20: Anomaly detection of ECG unhealthy signals employing the OGK Covariance estimator for the simulation-based tuning of the inflation factor F.

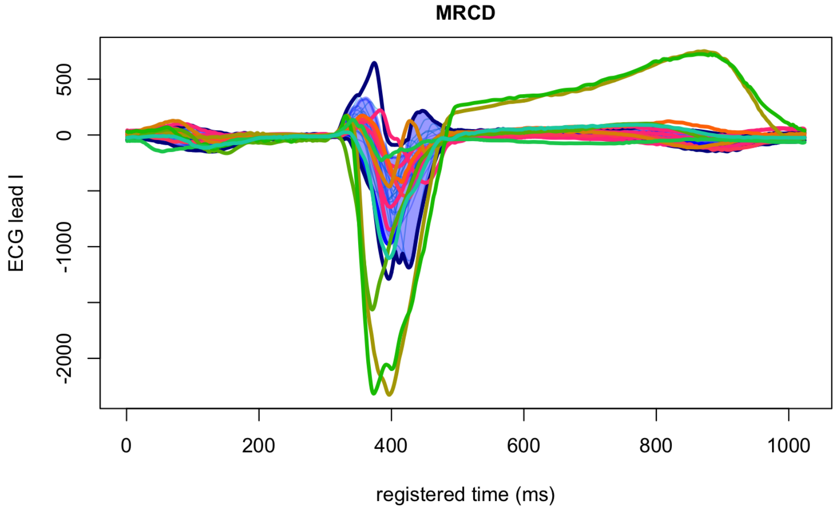


Figure C.21: Anomaly detection of ECG unhealthy signals employing the MRCD Covariance estimator for the simulation-based tuning of the inflation factor F.

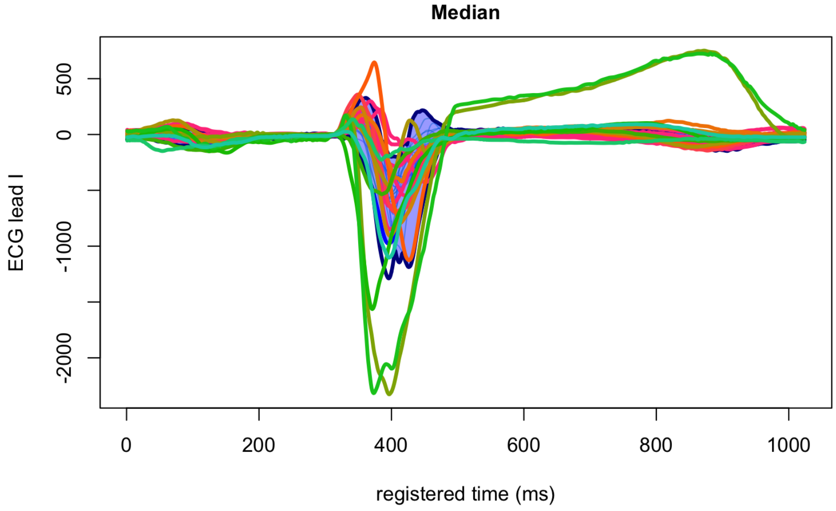


Figure C.22: Anomaly detection of ECG unhealthy signals employing the Median Covariation Operator for the simulation-based tuning of the inflation factor F.

References

[1] Boente, G., Rodriguez, D., Sued, M., 2018. The spatial sign covariance operator: Asymptotic results and applications. Journal of Multivariate Analysis 170. doi:10.1016/j.jmva.2018.10.002.

[2] Boente, G., Salibián Barrera, M., Tyler, D.E., 2014. A characterization of elliptical distributions and some optimality properties of principal components for functional data. Journal of Multivariate Analysis 131, 254–264. URL: https://www.sciencedirect.com/science/article/pii/S0047259X14001638, doi:https://doi.org/10.1016/j.jmva.2014.07.006.

[3] Boudt, K., Rousseeuw, P., Vanduffel, S., Verdonck, T., 2020. The minimum regularized covariance determinant estimator. Statistics and Computing 30. doi:10.1007/s11222-019-09869-x.

[4] Cardot, H., Godichon-Baggioni, A., 2015. Robust principal components analysis based on the median covariation matrix .

[5] Cardot, H., Godichon-Baggioni, A., 2017. Fast estimation of the median covariation matrix with application to online robust principal components analysis. TEST 26, 461–480. URL: http://link.springer.com/10.1007/s11749-016-0519-x, doi:10.1007/s11749-016-0519-x.

[6] Cuevas, A., Febrero-Bande, M., Fraiman, R., 2007. Robust estimation and classification for functional data via projection-based depth notions. Computational Statistics 22, 481–496. doi:10.1007/s00180-007-0053-0.

[7] Dai, W., Genton, M., 2018. Functional boxplots for multivariate curves: Multivariate curves. Stat 7, e190. doi:10.1002/sta4.190.

[8] Ferraty, F., Vieu, P., 2006. Nonparametric functional data analysis: theory and practice. Springer series in statistics, Springer, New York. OCLC: ocm70261207.

[9] Gervini, D., 2008. Robust functional estimation using the median and spherical principal components. Biometrika 95, 587–600. URL: https://doi.org/10.1093/biomet/asn031, doi:10.1093/biomet/asn031.

[10] Hall, P., Müller, H.G., Wang, J.L., 2006. Properties of principal component methods for functional and longitudinal data analysis. The Annals of Statistics 34. doi:10.1214/009053606000000272.

[11] Hawkins, D., 2014. Identification of outliers. Springer. OCLC: 1154106149.

[12] Hofmann, T., Schölkopf, B., Smola, A.J., 2008. Kernel methods in machine learning. The Annals of Statistics 36. URL: https://projecteuclid.org/journals/annals-of-statistics/volume-36/issue-3/Kernel-methods-in-machine-learning/10.1214/009053607000000677.full, doi:10.1214/009053607000000677.

[13] Hubert, M., Rousseeuw, P.J., Segaert, P., 2015. Multivariate functional outlier detection. Statistical Methods & Applications 24, 177–202. URL: http://link.springer.com/10.1007/s10260-015-0297-8, doi:10.1007/s10260-015-0297-8.

[14] Hyndman, R.J., Shang, H.L., 2010. Rainbow Plots, Bagplots, and Boxplots for Functional Data. Journal of Computational and Graphical Statistics 19, 29–45. URL: http://www.tandfonline.com/doi/abs/10.1198/jcgs.2009.08158, doi:10.1198/jcgs.2009.08158.

[15] Ieva, F., Paganoni, A.M., Pigoli, D., Vitelli, V., 2013. Multivariate functional clustering for the morphological analysis of electrocardiograph curves: *Analysis of Electrocardiograph Curves*. Journal of the Royal Statistical Society: Series C (Applied Statistics) 62, 401–418. URL: https://onlinelibrary.wiley.com/doi/10.1111/j.1467-9876.2012.01062.x, doi:10.1111/j.1467-9876.2012.01062.x.

[16] Kendall, M.G., 1938. A new measure of rank correlation. Biometrika 30, 81–93. URL: https://doi.org/10.1093/biomet/30.1-2.81, doi:10.1093/biomet/30.1-2.81.

[17] Kraus, D., Panaretos, V.M., 2012. Dispersion operators and resistant second-order functional data analysis. Biometrika 99, 813–832. URL: http://www.jstor.org/stable/41720736.

[18] Kriegel, H.P., Schubert, E., Zimek, A., 2017. The (black) art of runtime evaluation: Are we comparing algorithms or implementations? Knowledge and Information Systems 52, 341–378.

URL: http://link.springer.com/10.1007/s10115-016-1004-2, doi:10.1007/s10115-016-1004-2.

[19] Ledoit, O., Wolf, M., 2004. A well-conditioned estimator for large-dimensional covariance matrices. Journal of Multivariate Analysis 88, 365–411. URL: https://linkinghub.elsevier.com/retrieve/pii/S0047259X03000964, doi:10.1016/S0047-259X(03)00096-4.

[20] López-Pintado, S., Romo, J., 2009. On the Concept of Depth for Functional Data. Journal of the American Statistical Association 104, 718–734. URL: http://www.tandfonline.com/doi/abs/10.1198/jasa.2009.0108, doi:10.1198/jasa.2009.0108.

[21] Loève, M., 1978. Probability theory. 2. Number 46 in Graduate texts in mathematics. 4th ed ed., Springer, New York, NY Heidelberg.

[22] López-Pintado, S., Romo, J., 2007. Depth-based inference for functional data. Computational Statistics & Data Analysis 51, 4957–4968. URL: https://linkinghub.elsevier.com/retrieve/pii/S0167947306003872, doi:10.1016/j.csda.2006.10.029.

[23] López-Pintado, S., Romo, J., 2009. On the Concept of Depth for Functional Data. Journal of the American Statistical Association 104, 718–734. URL: http://www.tandfonline.com/doi/abs/10.1198/jasa.2009.0108, doi:10.1198/jasa.2009.0108.

[24] Maronna, R.A., Zamar, R.H., 2002. Robust Estimates of Location and Dispersion for High-Dimensional Datasets. Technometrics 44, 307–317. URL: http://www.tandfonline.com/doi/abs/10.1198/004017002188618509, doi:10.1198/004017002188618509.

[25] Marron, J.S., Alonso, A.M., 2014. Overview of object oriented data analysis: An overview of object oriented data analysis. Biometrical Journal 56, 732–753. URL: https://onlinelibrary.wiley.com/doi/10.1002/bimj.201300072, doi:10.1002/bimj.201300072.

[26] Ojo, O., Lillo, R.E., Anta, A.F., 2021. Outlier Detection for Functional Data with R Package fdaoutlier. URL: http://arxiv.org/abs/2105.05213. arXiv:2105.05213 [stat].

[27] Qu, Z., Genton, M.G., 2022. Sparse Functional Boxplots for Multivariate Curves. Journal of Computational and Graphical Statistics 31, 976–989. URL: https://www.tandfonline.com/doi/full/10.1080/10618600.2022.2066680, doi:10.1080/10618600.2022.2066680.

[28] Ramsay, J.O., Silverman, B.W., 2005. Functional data analysis. Springer series in statistics. 2nd ed ed., Springer, New York.

[29] Rousseeuw, P.J., Croux, C., 1993. Alternatives to the Median Absolute Deviation. Journal of the American Statistical Association 88, 1273–1283. URL: http://www.tandfonline.com/doi/abs/10.1080/01621459.1993.10476408, doi:10.1080/01621459.1993.10476408.

[30] Schreurs, J., Vranckx, I., Hubert, M., Suykens, J.A.K., Rousseeuw, P.J., 2021. Outlier detection in non-elliptical data by kernel MRCD. Statistics and Computing 31, 66. URL: https://link.springer.com/10.1007/s11222-021-10041-7, doi:10.1007/s11222-021-10041-7.

[31] Sun, Y., Genton, M., 2010. Functional boxplot. Journal of Computational and Graphical Statistics 20. doi:10.2307/23110490.

[32] Sun, Y., Genton, M.G., 2012. Adjusted functional boxplots for spatio-temporal data visualization and outlier detection. Environmetrics 23, 54–64. URL: https://onlinelibrary.wiley.com/doi/10.1002/env.1136, doi:10.1002/env.1136.

[33] Tukey, J.W., et al., 1977. Exploratory data analysis. volume 2. Reading, MA.

[34] Vantini, S., 2012. On the definition of phase and amplitude variability in functional data analysis. TEST 21, 676–696. URL: http://link.springer.com/10.1007/s11749-011-0268-9, doi:10.1007/s11749-011-0268-9.

[35] Yang, T., Gregg, R.E., Babaeizadeh, S., 2021. Big data reveals insights for lead importance in ECG interpretation. Journal of Electrocardiology 69, 12–22. URL: https://linkinghub.elsevier.com/retrieve/pii/S0022073621000042, doi:10.1016/j.jelectrocard.2021.01.002.

[36] Zhong, R., Liu, S., Li, H., Zhang, J., 2022. Robust functional principal component analysis for non-gaussian longitudinal data.

Journal of Multivariate Analysis 189, 104864. URL: https://www.sciencedirect.com/science/article/pii/S0047259X21001421, doi:https://doi.org/10.1016/j.jmva.2021.104864.

MOX Technical Reports, last issues

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

51/2023 Bucelli, M.; Regazzoni, F.; Dede', L.; Quarteroni, A.
Preserving the positivity of the deformation gradient determinant in intergrid interpolation by combining RBFs and SVD: application to cardiac electromechanics

52/2023 Antonietti, P.F.; Botti, M.; Mazzieri, I.
A space-time discontinuous Galerkin method for coupled poroelasticity-elasticity problems

49/2023 Ieva, F.; Ronzulli, M.; Romo, J.; Paganoni, A.M.
A Spearman Dependence Matrix for Multivariate Functional Data

48/2023 Renzi, F.; Vergara, C.; Fedele, M.; Giambruno, V.; Quarteroni, A.; Puppini, G.; Luciani, G.B.
Accurate and Efficient 3D Reconstruction of Right Heart Shape and Motion from Multi-Series Cine-MRI

45/2023 Gironi, P.; Petraro, L.; Santoni, S.; Dede', L.; Colosimo, B.M.
A Computational Model of Cell Viability and Proliferation of Extrusion-based 3D Bioprinted Constructs During Tissue Maturation Process

44/2023 Fontana, N.; Savaré, L.; Rea, F.; Di Angelantonio, E.; Ieva, F.
Long-term adherence to polytherapy in heart failure patients: a novel approach emphasising the importance of secondary prevention

42/2023 Tonini, A.; Vergara, C.; Regazzoni, F.; Dedè, L.; Scrofani, R.; Cogliati, C.; Quarteroni, A.
A mathematical model to assess the effects of COVID-19 on the cardiocirculatory system

41/2023 Corti M.; Bonizzoni, F.; Antonietti, P.F.; Quarteroni, A.M.
Uncertainty Quantification for Fisher-Kolmogorov Equation on Graphs with Application to Patient-Specific Alzheimer Disease

40/2023 Ballini, E.; Chiappa, A.S.; Micheletti, S.
Reducing the Drag of a Bluff Body by Deep Reinforcement Learning

39/2023 Riccobelli, D.; Al-Terke, H. H.; Laaksonen, P.; Metrangolo, P.; Paananen, A.; Ras, R. H. A.; Ciarletta, P.; Vella, D.
Flattened and wrinkled encapsulated droplets: Shape-morphing induced by gravity and evaporation