Functional Outlier Detection with a Local Spatial Depth

Carlo Sguera
Department of Statistics, Universidad Carlos III de Madrid
28903 Getafe (Madrid), Spain
(csguera@est-econ.uc3m.es)

Pedro Galeano
Department of Statistics, Universidad Carlos III de Madrid
28903 Getafe (Madrid), Spain
(pedro.galeano@uc3m.es)

and

Rosa E. Lillo
Department of Statistics, Universidad Carlos III de Madrid
28903 Getafe (Madrid), Spain
(rosaelvira.lillo@uc3m.es)

Abstract

This paper proposes methods to detect outliers in functional datasets and the task of identifying atypical curves is carried out using the recently proposed kernelized functional spatial depth (KFSD). KFSD is a local depth that can be used to order the curves of a sample from the most to the least central, and since outliers are usually among the least central curves, we present a probabilistic result which allows to select a threshold value for KFSD such that curves with depth values lower than the threshold are detected as outliers. Based on this result, we propose three new outlier detection procedures. The results of a simulation study show that our proposals generally outperform a battery of competitors. Finally, we consider a real application with environmental data consisting in levels of nitrogen oxides.

Keywords: Functional depths; Functional outlier detection; Kernelized functional spatial depth; Nitrogen oxides; Smoothed resampling.
1 INTRODUCTION

The accurate identification of outliers is an important aspect in any statistical data analysis. Nowadays there are well-established outlier detection techniques in the univariate and multivariate frameworks (for a complete review of the topic, see for example Barnett and Lewis [1994]). In recent years, new types of data have become available and tractable thanks to the evolution of computing resources, e.g., big multivariate datasets having more variables than observations (high-dimensional multivariate data) or samples composed of repeated measurements of the same observation taken over an ordered set of points that can be interpreted as realizations of stochastic processes (functional data). In this paper we focus on functional data, which are usually studied with the tools provided by functional data analysis (FDA). For overviews on FDA methods, see Ramsay and Silverman (2005), Ferraty and Vieu (2006), Horváth and Kokoszka (2012) or Cuevas (2014).

As in univariate or multivariate analysis, the detection of outliers is also fundamental in FDA. According to Febrero et al (2007, 2008), a functional outlier is a curve generated by a stochastic process with a different distribution than the one of normal curves. This definition covers many types of outliers, e.g., magnitude outliers, shape outliers and partial outliers, i.e., curves having atypical behaviors only in some segments of the domain. Shape and partial outliers are typically harder to detect than magnitude outliers (in the case of high magnitude, outliers can even be recognized by simply looking at a graph), and therefore entail more challenging outlier detection problems. In this paper we focus on samples contaminated by low magnitude, shape or partial outliers.

We propose to detect functional outliers using the notion of functional depth. A functional depth is a measure providing a \( P \)-based center-outward ordering criterion for observations of a functional space \( \mathbb{H} \), where \( P \) is a probability distribution on \( \mathbb{H} \). When a sample of curves is available, a functional depth orders the curves from the most to the least central according to their depth values and, if any outlier is in the sample, its depth is expected to be among the lowest values. Therefore, it is reasonable to build outlier detection methods that use functional depths.
In this paper we enlarge the number of available functional outlier detection procedures by presenting three new methods based on a specific depth, the kernelized functional spatial depth (KFSD, Sguera et al. [2014]). KFSD is a local-oriented depth, that is, a depth which orders curves looking at narrow neighborhoods and giving more weight to close than distant curves. Its approach is opposite to what global-oriented depths do. Indeed, any global depth makes depend the depth of a given curve on the whole rest of observations, with equal weights for all of them. This is the case of a global-oriented depth such as the functional spatial depth (FSD, Chakraborty and Chaudhuri [2014]), of which KFSD is its local version. A local depth such as KFSD may result useful to analyze functional samples having a structure deviating from unimodality or symmetry. Moreover, the local approach behind KFSD proved to be a good strategy in supervised classification problems with groups of curves not extremely clear-cut (see Sguera et al. [2014]). Alternatively, we illustrate that KFSD ranks well low magnitude, shape or partial outliers, that is, their corresponding KFSD values are in general lower than those of normal curves. Then, we propose different procedures to select a threshold for KFSD to distinguish between normal curves and outliers. These procedures employ smoothing resampling techniques and are based on a theoretical result which allows to obtain a probabilistic upper bound on a desired false alarm probability of detecting normal curves as outliers. Note that the probabilistic foundations of the proposed methods represent a novelty in FDA outlier detection problems.

We study the performances of our procedures in a simulation study and consider a real outlier detection problem with environmental data consisting in nitrogen oxides (NO\textsubscript{x}) emission daily levels measured every hour close to an industrial area in Poblenou (Barcelona). We compare our methods with some alternative outlier detection procedures: Febrero et al. (2008) proposed to label as outliers those curves with depth values lower than a certain threshold. As functional depths, they considered the Fraiman and Muniz depth (Fraiman and Muniz [2001]), the h-modal depth (Cuevas et al. [2006]) and the integrated dual depth (Cuevas and Fraiman [2009]). To determine the depth threshold, they proposed two different bootstrap procedures based on depth-based trimmed or weighted resampling, respectively; Sun and Genton (2011) introduced the functional boxplot, which is constructed using the ranking of curves provided by the modified band depth (Lópe-Pintado and Romo [2009]). The proposed functional
boxplot detects outliers using a rule that is similar to the one of the standard boxplot; Hyndman and Shang (2010) proposed to reduce the outlier detection problem from functional to multivariate data by means of functional principal component analysis (FPCA), and to use two alternative multivariate techniques on the scores to detect outliers, i.e., the bagplot and the high density region boxplot, respectively.

The remainder of the article is organized as follows. In Section 2 we recall the definition of KFSD. In Section 3 we consider the functional outlier detection problem. In Theorem 1 we present the result on which are based three new outlier detection methods which employ KFSD as depth function. In Section 4 we report the results of our simulation study, whereas in Section 5 we perform outlier detection on the NO\textsubscript{x} dataset. In Section 6 we draw some conclusions. Finally, in the Appendix we report a sketch of the proof of Theorem 1.

2 THE KERNELIZED FUNCTIONAL SPATIAL DEPTH

In functional spaces a depth measure has the purpose of measuring the degree of centrality of curves relative to the distribution of a functional random variable. Various functional depths have been proposed following two alternative approaches: a global approach, which implies that the depth of an observation depends equally on all the observations allowed by \( P \) on \( \mathbb{H} \), and a local approach, which instead makes depend the depth of an observation more on close than distant observations. Among the existing global-oriented depths there is the Fraiman and Muniz depth (FMD, Fraiman and Muniz 2001), the random Tukey depth (RTD, Cuesta-Albertos and Nieto-Reyes 2008), the integrated dual depth (IDD, Cuevas and Fraiman 2009), the modified band depth (MBD, López-Pintado and Romo 2009) or the functional spatial depth (FSD, Chakraborty and Chaudhuri 2014). Proposals of local-oriented depths are instead the h-modal depth (HMD, Cuevas et al. 2006) or the kernelized functional spatial depth (KFSD, Sguera et al. 2014).

In this paper we focus on KFSD. Before giving its definition, we recall the definition of the functional spatial depth (FSD, Chakraborty and Chaudhuri 2014). Let \( \mathbb{H} \) be an infinite-
dimensional Hilbert space, then for \( x \in \mathbb{H} \) and the functional random variable \( Y \in \mathbb{H} \), FSD of \( x \) relative to \( Y \) is given by

\[
FSD(x, Y) = 1 - \left\| \mathbb{E} \left[ \frac{x - Y}{\|x - Y\|} \right] \right\|
\]

where \( \| \cdot \| \) is the norm inherited from the usual inner product in \( \mathbb{H} \). For a \( n \)-size random sample of \( Y \), i.e., \( Y_n = \{y_1, \ldots, y_n\} \), the sample version of FSD has the following form:

\[
FSD(x, Y_n) = 1 - \frac{1}{n} \left\| \sum_{i=1}^{n} \frac{x - y_i}{\|x - y_i\|} \right\|.
\] (1)

As mentioned before, FSD is a global-oriented depth and Sguera et al. (2014) proposed a local version of FSD, i.e., KFSD. KFSD is obtained writing (1) in terms of inner products and then replacing the inner product function with a positive definite and stationary kernel function. This replacement exploits the relationship

\[
\kappa(x, y) = \langle \phi(x), \phi(y) \rangle, \quad x, y \in \mathbb{H},
\] (2)

where \( \kappa \) is the kernel \( \kappa : \mathbb{H} \times \mathbb{H} \to \mathbb{R} \), \( \phi \) is the embedding map \( \phi : \mathbb{H} \to \mathbb{F} \) and \( \mathbb{F} \) is a feature space. Indeed, a definition of KFSD in terms of \( \phi \) can be given, that is,

\[
KFSD(x, Y) = 1 - \left\| \mathbb{E} \left[ \frac{\phi(x) - \phi(Y)}{\|\phi(x) - \phi(Y)\|} \right] \right\|
\] (3)

and it can be interpreted as a recoded version of \( FSD(x, Y) \) since \( KFSD(x, Y) = FSD(\phi(x), \phi(Y)) \). The sample version of (3) is given by

\[
KFSD(x, Y_n) = 1 - \frac{1}{n} \left\| \sum_{i=1}^{n} \frac{\phi(x) - \phi(y_i)}{\|\phi(x) - \phi(y_i)\|} \right\|.
\]

Then, standard calculations that use (2) allow to provide an alternative expression of \( KFSD(x, Y_n) \), in this case in terms of \( \kappa \):
KFSD\((x, Y_n) = 1 - \\
\frac{1}{n} \left( \sum_{i,j=1; y_i \neq x; y_j \neq x}^{n} \frac{\kappa(x, y_i) + \kappa(y_i, y_j) - \kappa(x, y_i) - \kappa(x, y_j)}{\sqrt{\kappa(x, x) + 2\kappa(x, y_i)\sqrt{\kappa(x, x) + \kappa(y_j, y_j) - 2\kappa(x, y_j)}}} \right)^{1/2}, \quad (4)

Note that (4) only requires the choice of \(\kappa\), and not of \(\phi\), which can be left implicit. As in Sguera et al (2014), we use as \(\kappa\) the Gaussian kernel function given by

\[
\kappa(x, y) = \exp \left( -\frac{||x - y||^2}{\sigma^2} \right), \quad (5)
\]

where \(x, y \in \mathbb{H}\). In turn, (5) depends on the norm function inherited by the functional Hilbert space where data are assumed to lie, and on the bandwidth \(\sigma\). Regarding \(\sigma\), we initially consider 9 different \(\sigma\), each one equal to 9 different percentiles of the empirical distribution of \(\{||y_i - y_j||, y_i, y_j \in Y_n\}\). The first percentile is 10\%, and by increments of 10 we obtain the ninth percentile, i.e., 90\%. Note that the lower \(\sigma\), the more local the approach, and therefore the percentiles that we use cover different degrees of KFSD-based local approaches: strongly (e.g., 20\%), moderately (e.g., 50\%) and weakly (e.g., 80\%) local approaches. In Section 4 we present a method to select \(\sigma\) in outlier detection problems.

In general, since any functional depth measures the degree of centrality or extremality of a given curve relative to a distribution or a sample, outliers are expected to have low depth values. More in particular, in presence of low magnitude, shape or partial outliers, an approach based on the use of a local depth like KFSD may help in detecting outliers. To illustrate this fact, we present the following example: first, we generated 100 datasets of size 50 from a mixture of two stochastic processes, one for normal curves and one for high magnitude outliers, with the probability that a curve is an outlier equal to 0.05. Second, we generated another group of 100 datasets from a different mixture which produces shape outliers. In Figure 1 we report a contaminated dataset for each mixture.

Let \(n_{out,j}, j = 1, \ldots, 100\), be the number of outliers generated in the \(j\)th dataset. For each dataset and functional depth, it is desirable to assign the \(n_{out,j}\) lowest depth values to the
Fig. 1: Examples of contaminated datasets: high magnitude contamination (top) and shape contamination (bottom). The solid curves are normal curves and the dashed curves are outliers.

For both mixtures and each generated dataset, we recorded how many times the depth of an outlier is among the $n_{out,j}$ lowest values. As depth functions, we considered five global depths (FMD, RTD, IDD, MBD and FSD) and two local depths (HMD and KFSD). The results reported in Table 1 show that for all the functional depths the ranking of high magnitude outliers is an easier task than the ranking of shape outliers. However, while the ranking of high magnitude outliers is reasonably good in different cases, e.g., for the local KFSD (94.87%) and the global RTD (90.17%), the ranking of shape outliers is markedly better with local depths (86.72% for KFSD and 85.47% for HMD) than with the best global depths (with 39.06%, FSD is the best performing global depth). These results suggest that, selecting a proper threshold, KFSD can isolate well outliers.

Table 1: Percentages of times a depth assigns a value among the $n_{out,j}$ lowest ones to an outlier.

| Type of outliers          | Global depths | Local depths |
|---------------------------|---------------|--------------|
| High magnitude outliers   | 86.32 90.17  81.62 69.23 68.80 | 85.47 94.87 |
| Shape outliers            | 7.81 33.59  38.67 12.11 39.06 | 85.94 86.72 |
3 OUTLIER DETECTION FOR FUNCTIONAL DATA

The outlier detection problem can be described as follows: let $Y_n = \{y_1, \ldots, y_n\}$ be a sample generated from a mixture of two functional random variables in $\mathbb{H}$, one for normal curves and one for outliers, say $Y_{nor}$ and $Y_{out}$, respectively. Let $Y_{mix}$ be a mixture, i.e.,

$$Y_{mix} = \begin{cases} 
Y_{nor}, & \text{with probability } 1 - \alpha, \\
Y_{out}, & \text{with probability } \alpha,
\end{cases} \tag{6}$$

where $\alpha \in [0, 1]$ is the contamination probability (usually, a value rather close to 0). The curves composing $Y_n$ are all unlabeled, and the goal of the analysis is to decide whether each curve is a normal curve or an outlier.

KFSD is a functional extension of the kernelized spatial depth for multivariate data (KSD) proposed by Chen et al (2009), who also proposed a KSD-based outlier detector that we generalize to KFSD: for a given dataset $Y_n$ generated from $Y_{mix}$ and $t \in [0, 1]$, the KFSD-based outlier detector for $x \in \mathbb{H}$ is given by

$$g(x, Y_n) = \begin{cases} 
1, & \text{if } KFSD(x, Y_n) \leq t, \\
0, & \text{if } KFSD(x, Y_n) > t, \tag{7}
\end{cases}$$

where $t$ is a threshold which allows to discriminate between outliers (i.e., $g(x, Y_n) = 1$) and normal curves (i.e., $g(x, Y_n) = 0$), and it is a parameter that needs to be set.

For the multivariate case, Chen et al (2009) studied KSD-based outlier detection under different scenarios. One of them consists in an outlier detection problem where two samples are available, and for which they proposed to select the threshold $t$ by controlling the probability that normal observations are classified as outliers, i.e., the false alarm probability (FAP). They proved a result providing a KSD-based probabilistic upper bound on the FAP which depends on $t$. Then, the maximum value of $t$ such that the upper bound does not exceed a given desired FAP provides a threshold for KSD. We extend this result to KFSD:

**Theorem 1** Let $Y_{ny} = \{y_i, \ldots, y_{ny}\}$ and $Z_{nz} = \{z_i, \ldots, z_{nz}\}$ be i. i. d. samples generated from the unknown mixture of random variables $Y_{mix} \in \mathbb{H}$ described by (6), with $\alpha > 0$. Let
$g(\cdot, Y_{ny})$ be the outlier detector defined in [7]. Fix $\delta \in (0, 1)$ and suppose that $\alpha \leq r$ for some $r \in [0, 1]$. For a new random element $x$ generated from $Y_{nor}$, the following inequality holds with probability at least $1 - \delta$:

$$
\mathbb{E}_{x|Y_{ny}} [g(x, Y_{ny})] \leq \frac{1}{1 - r} \left[ \frac{1}{n_Z} \sum_{i=1}^{n_Z} g(z_i, Y_{ny}) + \sqrt{\frac{\ln 1/\delta}{2n_Z}} \right],
$$
where $\mathbb{E}_{x|Y_{ny}}$ refers to the expected value of $x$ for a given $Y_{ny}$.

The proof of Theorem 1 is presented in the Appendix. Recall that the FAP is the probability that a normal observation $x$ is classified as outlier. For the elements of Theorem 1, $\Pr_{x|Y_{ny}} (g(x, Y_{ny}) = 1)$ is the FAP. Moreover,

$$
\Pr_{x|Y_{ny}} (g(x, Y_{ny}) = 1) = \mathbb{E}_{x|Y_{ny}} [g(x, Y_{ny})].
$$

Therefore, the probabilistic upper bound of Theorem 1 applies also to the FAP.

It is worth noting that the application of Theorem 1 requires to observe two samples, circumstance rather uncommon in classical outlier detection problems, in which usually a single functional sample $Y_n$ generated from an unknown mixture of random variables $Y_{mix}$ is observed. For this reason, we suggest the following practical solution: set $Y_{ny} = Y_n$ and obtain $Z_{n_Z}$ by resampling with replacement from $Y_n$. Moreover, with our solution and for sufficiently large values of $n_Z$, the effect of $\delta$ on the probabilistic upper bound drastically reduces. Concerning $r$, that is the upper bound for the unknown contamination probability $\alpha$, a true range between 0 and 0.1 appears to be appropriate to cover most of the situations found in practice. Regarding the resampling procedure to obtain $Z_{n_Z}$, we consider three different schemes, all of them with replacement. Since we deal with potentially contaminated datasets, besides simple resampling, we also consider two robust KFSD-based resampling procedures inspired by the work of Febrero et al. (2008). The three resampling schemes that we consider are:

1. Simple resampling.
2. KFSD-based trimmed resampling: once $KFSD(y_i, Y_n), i = 1, \ldots, n$ are obtained, it is possible to identify the $[\alpha_T] \%$ least deepest curves, for a certain $0 < \alpha_T < 1$ usually
close to 0. These least deep curves are deleted from the sample, and simple resampling is carried out with the remaining curves.

3. KFSD-based weighted resampling: once \( KFSD(y_i, Y_n), i = 1, \ldots, n \) are obtained, weighted resampling is carried out with weights \( w_i = KFSD(y_i, Y_n) \).

All the above procedures generate samples with some repeated curves. However, in a preliminary stage of our study we observed that it is preferable to work with \( Z_{nZ} \) composed of non-repeated curves. To obtain such samples, we add a common smoothing step to the previous three resampling schemes.

To describe the smoothing step, first recall that each curve in \( Y_n \) is in practice observed at a discretized and finite set of domain points, and that the sets may differ from one curve to another. For this reason, the estimation of \( Y_n \) at a common set of \( m \) equidistant domain points may be required. Let \((y_i(s_1), \ldots, y_i(s_m))\) be the observed or estimated \( m \)-dimensional equidistant discretized version of \( y_i \); \( \Sigma_{Y_n} \) be the covariance matrix of the discretized form of \( Y_n \) and \( \gamma \) be a smoothing parameter. Consider a zero-mean Gaussian process whose discretized form has \( \gamma \Sigma_{Y_n} \) as covariance matrix. Let \((\zeta(s_1), \ldots, \zeta(s_m))\) be a discretized realization of the previous Gaussian process. Consider any of the previous three resampling procedures and assume that at the \( j \)th trial, \( j = 1, \ldots, n_{Z} \), the \( i \)th curve in \( Y_n \) has been sampled. Then, the discretized form of the \( j \)th curve in \( Z_{nZ} \) would be given by \((z_j(s_1), \ldots, z_j(s_m)) = (y_i(s_1) + \zeta(s_1), \ldots, y_i(s_m) + \zeta(s_m))\), or, in functional form, by \( z_j = y_i + \zeta \). Therefore, combining each resampling scheme with this smoothing step, we provide three different approximate ways to obtain \( Z_{nZ} \), and we refer to them as \( smo \), \( tri \) and \( wei \), respectively. Then, for fixed \( \delta, r \) and desired FAP, the threshold \( t \) for (7) is selected as the maximum value of \( t \) such that the right-hand side of (8) does not exceed the desired FAP. Let \( t^* \) be the selected threshold, which is then used in (7) to compute \( g(y_i, Y_n), i = 1, \ldots, n \). If \( g(y_i, Y_n) = 1 \), \( y_i \) is detected as outlier. To summarize, we provide three KFSD-based outlier detection procedures and we refer to them as \( KFSD_{smo} \), \( KFSD_{tri} \) and \( KFSD_{wei} \) depending on how \( Z_{nZ} \) is obtained (\( smo \), \( tri \) and \( wei \), respectively; recall that \( Y_{nV} = Y_n \)). As competitors of the proposed procedures, we consider the methods mentioned in Section \( \text{I} \) that we now describe.
Sun and Genton (2011) proposed a depth-based functional boxplot and an associated outlier detection rule based on the ranking of the sample curves that MBD provides. The ranking is used to define a sample central region, that is, the smallest band containing at least half of the deepest curves. The non-outlying region is defined inflating the central region by 1.5 times. Curves that do not belong completely to the non-outlying region are detected as outliers. The original functional boxplot is based on the use of MBD as depth, but clearly any functional depth can be used. Another contribution of this paper is the study of the performances of the outlier detection rule associated to the functional boxplot (from now on, FBP) when used together with the battery of functional depths mentioned in Section 2.

Febrero et al (2008) proposed two depth-based outlier detection procedures that select a threshold for FMD, HMD or IDD by means of two alternative robust smoothed bootstrap procedures whose single bootstrap samples are obtained using the above described tri and wei, respectively. At each bootstrap sample, the 1% percentile of empirical distribution of the depth values is obtained, say $p_{0.01}$. If $B$ is the number of bootstrap samples, $B$ values of $p_{0.01}$ are obtained. Each method selects as cutoff $c$ the median of the collection of $p_{0.01}$ and, using $c$ as threshold, a first outlier detection is performed. If some curves are detected as outliers, they are deleted from the sample, and the procedure is repeated until no more outliers are found (note that $c$ is computed only in the first iteration). We refer to these methods as $B_{tri}$ and $B_{wei}$, and also in this case we evaluate these procedures using all the functional depths mentioned in Section 2.

Finally, we also consider two procedures proposed by Hyndman and Shang (2010) that are not based on the use of a functional depth. Both are based on the first two robust functional principal components scores and on two different graphical representations of them. The first proposal is the outlier detection rule associated to the functional bagplot (from now on, FBG), which works as follows: obtain the bivariate robust scores and order them using the multivariate halfspace depth (Tukey 1975). Define an inner region by considering the smallest region containing at least the 50% of the deepest scores, and obtain a non-outlying region by inflating the inner region by 2.58 times. FBG detects as outliers those curves whose scores are outside the non-outlying region. Note that the scores-based regions and outliers
allow to draw a bivariate bagplot, which produces a functional bagplot once it is mapped onto the original functional space. The second proposal is related to a different graphical tool, the high density region boxplot (from now on, we refer to its associated outlier detection rule as FHD). In this case, once obtained the scores, perform a bivariate kernel density estimation. Define the \((1 - \beta)\)-high density region (HDR), \(\beta \in (0, 1)\), as the region of scores with coverage probability equal to \((1 - \beta)\). FHD detects as outliers those curves whose scores are outside the \((1 - \beta)\)-HDR. In this case, it is possible to draw a bivariate HDR boxplot which can be mapped onto a functional version, thus providing the functional HDR boxplot.

4 SIMULATION STUDY

After introducing \(\text{KFSD}_\text{smo}, \text{KFSD}_\text{tri}\) and \(\text{KFSD}_\text{wei}\), their competitors (FBP, \(B_{\text{tri}}\), \(B_{\text{wei}}\), FBG and FHD), as well as seven different functional depths (FMD, HMD, RTD, IDD, MBD, FSD and KFSD), in this section we carry out a simulation study to evaluate the performances of the different methods. For FBP, \(B_{\text{tri}}\) and \(B_{\text{wei}}\), we use the notation procedure+depth: for example, FBP+FMD refers to the method obtained by using FBP together with FMD.

To perform our simulation study, we consider six models: all of them generate curves according to the mixture of random variables \(Y_{\text{mix}}\) described by (6). The first three mixture models (MM1, MM2 and MM3) share \(Y_{\text{nor}}\), with curves generated by

\[
y(s) = 4s + \epsilon(s),
\]

where \(s \in [0, 1]\) and \(\epsilon(s)\) is a zero-mean Gaussian component with covariance function given by

\[
\mathbb{E} (\epsilon(s), \epsilon(s')) = 0.25 \exp \left( -(s - s')^2 \right), \quad s, s' \in [0, 1].
\]

Also the remaining three mixture models (MM4, MM5 and MM6) share \(Y_{\text{nor}}\), but, in this case, the curves are generated by

\[
y(s) = u_1 \sin s + u_2 \cos s,
\]

(10)
where \( s \in [0, 2\pi] \) and \( u_1 \) and \( u_2 \) are observations from a continuous uniform random variable between 0.05 and 0.15.

MM1, MM2 and MM3 differ in their \( Y_{out} \) components. Under MM1, the outliers are generated by

\[
y(s) = 8s - 2 + \epsilon(s),
\]

which produces outliers of both shape and low magnitude nature. Under MM2, the outliers are generated by adding to (9) an observation from a \( N(0, 1) \), and as result outliers are more irregular than normal curves. Finally, under MM3, the outliers are generated by

\[
y(s) = 4 \exp(s) + \epsilon(s),
\]

which produces curves that are normal in the first part of the domain, but that become exponentially outlying.

Similarly, MM4, MM5 and MM6 differ in their \( Y_{out} \) components. Under MM4, the outliers are generated replacing \( u_2 \) with \( u_3 \) in (10), where \( u_3 \) is an observation from a continuous uniform random variable between 0.15 and 0.17. This change produces partial low magnitude outliers in the first and middle part of the domain of the curves. Under MM5, the outliers are generated by adding to (10) an observation from a \( N(0, (\frac{0.11}{2})^2) \), and they turn out to be more irregular curves. Finally, under MM6, the outliers are generated by

\[
y(s) = u_1 \sin s + \exp\left(\frac{0.69s}{2\pi}\right) u_4 \cos s, \tag{11}
\]

where \( u_4 \) is an observation from a continuous uniform random variable between 0.1 and 0.15. As MM3, MM6 allows outliers that are normal in the first part of the domain and become outlying with an exponential pattern. In Figure 2 we report a simulated dataset with at least one outlier for each mixture model.

The details of the simulation study are the following: for each mixture model, we generated 100 datasets, each one composed of 50 curves. As mentioned above, for each single samples Theorem 1 cannot be directly applied, and therefore KFSD_{smo}, KFSD_{tri} and KFSD_{we} represent practical alternatives. Two values of the contamination probability \( \alpha \)
Figure 2: Examples of contaminated functional datasets generated by MM1, MM2, MM3, MM4, MM5 and MM6. Solid curves are normal curves and dashed curves are outliers.

were considered: 0.02 and 0.05. All curves were generated using a discretized and finite set of 51 equidistant points in the domain of each mixture model (\([0, 1]\) for MM1, MM2 and MM3; \([0, 2\pi]\) for MM4, MM5 and MM6) and the discretized versions of the functional depths were used.

In relation with the methods and the functional depths that we consider in the study, their specifications are described next:

1. FBP when used with FMD, HMD, RTD, IDD, MBD, FSD and KFSD: regarding FBP, as reported in Section 3, the central region is built considering the 50% deepest curves and the non-outlying region by inflating by 1.5 times the central region. Regarding the depths, for HMD, we follow the recommendations in Febrero et al. (2008), that is, \(H\) is the \(L^2\) space, \(\kappa(x, y) = \frac{2}{\sqrt{2\pi}} \exp\left(-\frac{\|x-y\|^2}{2h^2}\right)\) and \(h\) is equal to the 15\% percentile of the empirical distribution of \(\{\|y_i - y_j\|, y_i, y_j \in Y_n\}\). For RTD and IDD, we work with 50 projections in random Gaussian directions. For MBD, we consider bands defined by two curves. For FSD and KFSD, we assume that the curves lie in the \(L^2\)
space. Moreover, in KFSD we set $\sigma$ equal to a moderately local percentile (50%) of the empirical distribution of $\{\|y_i - y_j\|, y_i, y_j \in Y_n\}$.

2. $B_{tri}$ and $B_{wei}$ when used with FMD, HMD, RTD, IDD, MBD, FSD and KFSD: $\gamma = 0.05$, $B = 200$, $\alpha_T = \alpha$. Regarding the depths, we use the specifications reported for FBP.

3. FBG: as reported in Section 3, the central region is built considering the 50% deepest bivariate robust functional principal component scores and the non-outlying region by inflating by 2.58 times the central region.

4. FHD: $\beta = \alpha$.

5. $KFSD_{smo}$, $KFSD_{tri}$ and $KFSD_{wei}$: $n_Y = n = 50$ (since $Y_{ny} = Y_n$), $\gamma = 0.05$, $\alpha_T = \alpha$ (only for $KFSD_{tri}$), $n_Z = 6n$, $\delta = 0.05$, $r = \alpha$, desired FAP = 0.10. Moreover, as introduced in Section 2 for these methods we consider 9 percentiles to set $\sigma$ in KFSD.

The way in which we propose to choose the most suitable percentile for outlier detection is presented below.

In supervised classification, the availability of training curves with known class memberships makes possible the definition of some natural procedures to set $\sigma$ for KFSD, such as cross-validation. However, in an outlier detection problem, it is common to have no information whether curves are normal or outliers. Therefore, training procedures are not immediately available.

We propose to overcome this drawback by obtaining a “training sample of peripheral curves”, and then choosing the percentile that ranks better these peripheral curves as final percentile for KFSD in $KFSD_{smo}$, $KFSD_{tri}$ and $KFSD_{wei}$. We now describe this procedure, which is based on $J$ replications. Let $Y_n$ be the functional dataset on which outlier detection has to be done and let $Y_{(n)} = \{y_{(1)}, \ldots, y_{(n)}\}$ be the depth-based ordered version of $Y_n$, where $y_{(1)}$ and $y_{(n)}$ are the curves with minimum and maximum depth, respectively. The steps to obtain a set of peripheral curves are the following:

I. Let $\{p_1, \ldots, p_K\}$ be the set of percentiles in use (in our case, as explained in Section 2, $p_k = (10k)\%$, $k \in \{1, \ldots, K = 9\}$), and choose randomly a percentile from the set. For the $j$th replication, $j \in \{1, \ldots, J\}$, denote the selected percentile as $p^j$. We use $J = 20$
in the rest of the paper.

II. Using \( p^j \), compute \( KFSDF_{p^j}(y_i, Y_n) \), \( i = 1, \ldots, n \), where the notation \( KFSDF_{p^j}(\cdot, \cdot) \) is used to describe what percentile is used. For the \( j \)th replication, denote the KFSD-based ordered curves as \( y_{(1),j}, \ldots, y_{(n),j} \).

III. Take \( y_{(1),j}, \ldots, y_{(l_j),j} \), where \( l_j \sim Bin(n, \frac{1}{n}) \). Apply the smoothing step described in Section 3 to these curves. For the smoothing step, we use \( \Sigma_{Y_n} \) and \( \gamma = 0.05 \). For the \( j \)th replication, denote the peripheral and smoothed curves as \( y_{(1),j}^\ast, \ldots, y_{(l_j),j}^\ast \).

IV. Repeat \( J \) times steps II-III to obtain a collection of \( L = \sum_{j=1}^J l_j \) peripheral curves, say \( Y_L \) (for an example, see Figure 3).

![Figure 3: Example of a training sample of peripheral curves for a contaminated dataset generated by MM1 with \( \alpha = 0.05 \). The solid and shaded curves are the original curves (both normal and outliers). The dashed curves are the peripheral curves to use as training sample.](image)

Next, \( Y_L \) acts as training sample according to the following steps: for each \( y_{(i),j}^\ast \in Y_L \), \( i \leq l_j \), and \( p_k \in \{p_1, \ldots, p_K\} \), compute \( KFSDF_{p_k}(y_{(i),j}^\ast, Y_{-(i),j}^\ast) \), where \( Y_{-(i),j} = Y_n \setminus \{y_{(i),j}\} \). At the end, a \( L \times K \) matrix is obtained, say \( D_{LK} = \{d_{lk}\}_{l=1, \ldots, L, k=1, \ldots, K} \), whose \( k \)th column is composed of the KFSD values of the \( L \) training peripheral curves when the \( k \)th percentile is employed in KFSD. Next, let \( r_{lk} \) be the rank of \( d_{lk} \) in the vector \( (KFSDF_{p_1}(y_1, Y_n), \ldots, KFSDF_{p_K}(y_n, Y_n)) \).
$d_{lk}$, e.g., $r_{lk}$ is equal to 1 or $n+1$ if $d_{lk}$ is the minimum or the maximum value in the vector, respectively. Let $R_{lk}$ be the result of this transformation of $D_{lk}$, and sum the elements of each column, obtaining a $K$-dimensional vector, say $\mathbf{R}_K$. Since the goal is to assign ranks as low as possible to the peripheral curves, choose the percentile associated to the minimum value of $\mathbf{R}_K$. When a tie is observed, we break it randomly.

The comparison among methods is performed in terms of both correct and false outlier detection percentages, which are reported in Tables 2-7. To ease the reading of the tables, for each model and $\alpha$, we report in bold the 5 best methods in terms of correct outlier detection percentage (c). For each model, if a method is among the 5 best ones for both contamination probabilities $\alpha$, we report its label in bold.

Table 2: MM1, $\alpha = \{0.02, 0.05\}$. Correct (c) and false (f) outlier detection percentages of FBP, B_{tri}, B_{wei}, FBG, FHD, KFSD_{smo}, KFSD_{tri} and KFSD_{wei}.

| Method     | FBP+FMD | FBP+HMD | FBP+RTD | FBP+IDD | FBP+MBD | FBP+FSD | FBP+KFSF |
|------------|---------|---------|---------|---------|---------|---------|----------|
| $\alpha = 0.02$ | 100.00  | 71.12   | 13.21   | 17.92   | 2.83    | 0.00    | 61.32    |
| $\alpha = 0.05$ | 100.00  | 71.12   | 13.21   | 17.92   | 2.83    | 0.00    | 61.32    |

Table 3: MM2, $\alpha = \{0.02, 0.05\}$. Correct (c) and false (f) outlier detection percentages of FBP, B_{tri}, B_{wei}, FBG, FHD, KFSD_{smo}, KFSD_{tri} and KFSD_{wei}.

| Method     | FBP+FMD | FBP+HMD | FBP+RTD | FBP+IDD | FBP+MBD | FBP+FSD | FBP+KFSF |
|------------|---------|---------|---------|---------|---------|---------|----------|
| $\alpha = 0.02$ | 100.00  | 71.12   | 13.21   | 17.92   | 2.83    | 0.00    | 61.32    |
| $\alpha = 0.05$ | 100.00  | 71.12   | 13.21   | 17.92   | 2.83    | 0.00    | 61.32    |

The results in Tables 2-7 show that:

1. KFSD_{tri} and KFSD_{wei} are always among the 5 best methods. KFSD_{smo} is among the 5 best methods 10 times over 12, but when its performance is not among the 5 best, it is

---

1In presence of tie, the method with lower false outlier detection percentage (f) is preferred.
Table 4: MM3, $\alpha = \{0.02, 0.05\}$. Correct (c) and false (f) outlier detection percentages of FBP, $B_{\text{tri}}$, $B_{\text{wei}}$, FBG, FHD, KFSD$_{\text{smo}}$, KFSD$_{\text{tri}}$ and KFSD$_{\text{wei}}$.

|                | $\alpha = 0.02$ | $\alpha = 0.05$ |
|----------------|-----------------|-----------------|
|                | c   | f   | c   | f   |
| FBP+$FMD$      | 55.69| 0.92| 49.19| 0.97|
| FBP+$HMD$      | 89.22| 0.57| 85.89| 0.63|
| FBP+$RTD$      | 86.27| 0.45| 76.61| 0.34|
| FBP+$IDD$      | 79.41| 0.51| 70.56| 0.38|
| FBP+$MBD$      | 74.31| 0.88| 59.27| 0.84|
| FBP+$FSD$      | 79.41| 0.51| 73.79| 0.42|
| FBP+$KFSD$     | 89.22| 0.57| 83.06| 0.59|

$B_{\text{smo}}$ 50.00 0.00 2.88 0.00 29.29 0.14 62.92 0.14

Table 5: MM4, $\alpha = \{0.02, 0.05\}$. Correct (c) and false (f) outlier detection percentages of FBP, $B_{\text{tri}}$, $B_{\text{wei}}$, FBG, FHD, KFSD$_{\text{smo}}$, KFSD$_{\text{tri}}$ and KFSD$_{\text{wei}}$.

|                | $\alpha = 0.02$ | $\alpha = 0.05$ |
|----------------|-----------------|-----------------|
|                | c   | f   | c   | f   |
| FBP+$FMD$      | 1.02| 0.00| 0.00| 0.00|
| FBP+$HMD$      | 6.12| 0.00| 1.60| 0.02|
| FBP+$RTD$      | 0.00| 0.00| 0.00| 0.00|
| FBP+$IDD$      | 0.00| 0.00| 0.00| 0.00|
| FBP+$MBD$      | 0.00| 0.00| 0.00| 0.00|
| FBP+$FSD$      | 0.00| 0.00| 0.00| 0.00|
| FBP+$KFSD$     | 2.04| 0.00| 0.80| 0.00|

$B_{\text{smo}}$ 100.00 8.20 0.16 47.70 0.14

Table 6: MM5, $\alpha = \{0.02, 0.05\}$. Correct (c) and false (f) outlier detection percentages of FBP, $B_{\text{tri}}$, $B_{\text{wei}}$, FBG, FHD, KFSD$_{\text{smo}}$, KFSD$_{\text{tri}}$ and KFSD$_{\text{wei}}$.

|                | $\alpha = 0.02$ | $\alpha = 0.05$ |
|----------------|-----------------|-----------------|
|                | c   | f   | c   | f   |
| FBP+$FMD$      | 2.11| 2.11| 0.11| 0.11|
| FBP+$HMD$      | 48.39| 0.00| 2.75| 0.00|
| FBP+$RTD$      | 3.90| 2.61| 0.12| 0.12|
| FBP+$IDD$      | 5.08| 4.61| 0.14| 0.14|
| FBP+$MBD$      | 3.90| 2.61| 0.12| 0.12|
| FBP+$FSD$      | 3.90| 2.61| 0.12| 0.12|
| FBP+$KFSD$     | 9.48| 5.08| 6.80| 3.09|

$B_{\text{smo}}$ 97.76 2.16 50.00 1.24

Table 7: MM6, $\alpha = \{0.02, 0.05\}$. Correct (c) and false (f) outlier detection percentages of FBP, $B_{\text{tri}}$, $B_{\text{wei}}$, FBG, FHD, KFSD$_{\text{smo}}$, KFSD$_{\text{tri}}$ and KFSD$_{\text{wei}}$.

|                | $\alpha = 0.02$ | $\alpha = 0.05$ |
|----------------|-----------------|-----------------|
|                | c   | f   | c   | f   |
| FBP+$FMD$      | 48.42| 0.00| 4.61| 0.00|
| FBP+$HMD$      | 60.00| 0.18| 62.92| 0.00|
| FBP+$RTD$      | 55.79| 0.00| 54.68| 0.00|
| FBP+$IDD$      | 46.32| 0.00| 40.07| 0.00|
| FBP+$MBD$      | 48.42| 0.00| 45.69| 0.00|
| FBP+$FSD$      | 52.63| 0.00| 52.43| 0.00|
| FBP+$KFSD$     | 57.89| 0.00| 56.93| 0.00|

$B_{\text{smo}}$ 71.58 2.21 45.69 3.02

Table 8: MM7, $\alpha = \{0.02, 0.05\}$. Correct (c) and false (f) outlier detection percentages of FBP, $B_{\text{tri}}$, $B_{\text{wei}}$, FBG, FHD, KFSD$_{\text{smo}}$, KFSD$_{\text{tri}}$ and KFSD$_{\text{wei}}$.

|                | $\alpha = 0.02$ | $\alpha = 0.05$ |
|----------------|-----------------|-----------------|
|                | c   | f   | c   | f   |
| FBP+$FMD$      | 29.41| 0.16| 29.41| 0.16|
| FBP+$HMD$      | 71.58| 0.24| 71.58| 0.24|
| FBP+$RTD$      | 36.00| 0.05| 36.00| 0.05|
| FBP+$FSD$      | 33.71| 0.32| 33.71| 0.32|
| FBP+$KFSD$     | 97.06| 8.96| 97.06| 8.96|

KFSD$_{\text{smo}}$ 98.99 0.18 94.00 0.44

KFSD$_{\text{tri}}$ 98.99 0.18 94.00 0.44

KFSD$_{\text{smo}}$ 100.00 4.61 98.40 2.11

KFSD$_{\text{tri}}$ 98.99 0.18 94.00 0.44

KFSD$_{\text{wei}}$ 98.99 0.18 94.00 0.44

18
neither extremely far from the fifth method (MM2, $\alpha = 0.05$: 95.18% against 96.79%; MM3, $\alpha = 0.05$: 73.79% against 78.63%). The rest of the methods are among the 5 best procedures at most 4 times over 12 (FBP+HMD and $B_{tri}$+HMD).

2. Regarding MM5 and MM6, our procedures are clearly the best options in terms of correct detection (c), and in the following order: KFSD$_{wei}$, KFSD$_{tri}$ and KFSD$_{smo}$. In general, this pattern is observed overall the simulation study. Note that for MM6 and $\alpha = 0.02$ we observe the best relative performances of KFSD$_{smo}$, KFSD$_{tri}$ and KFSD$_{wei}$, i.e., 91.58%, 93.68% and 96.84%, respectively, against 71.58% of the fourth best method (B$_{wei}$+KFSD), that is, we observe at least 20% differences, and approaching 30% if KFSD$_{wei}$ and B$_{wei}$+KFSD are compared.

3. About MM3, KFSD$_{wei}$ is clearly the best method in terms of correct detection, however at the price of having a greater false detection (f). This is in general the main weak point of KFSD$_{smo}$, KFSD$_{tri}$ and KFSD$_{wei}$. As for correct detection, we observe a overall pattern in our methods in false detection, but in an opposite way, indicating therefore a trade-off between c and f. Relative high false detection percentages are however something expected in KFSD$_{smo}$, KFSD$_{tri}$ and KFSD$_{wei}$ since these methods are based on the definition of a desired false alarm probability, which is equal to 10% in this study. Concerning MM2, we observe similar results to MM3, but in this case the performances of the best competitors for KFSD$_{wei}$, i.e., KFSD$_{smo}$, KFSD$_{tri}$, FBP-based methods and $B_{tri}$ when used with local depths, are closer to the results of KFSD$_{wei}$.

4. Finally, there are only 3 cases in which a competitor outperforms our methods: for MM1 and both $\alpha$ the best method is FBG, whereas for MM3 and $\alpha = 0.05$ the best method is B$_{wei}$+IDD. However, both FBG and B$_{wei}$+IDD do not show behaviors as stable as KFSD$_{smo}$, KFSD$_{tri}$ and KFSD$_{wei}$ do. Indeed, they show poor performances under other scenarios, e.g., MM2, MM5 or MM6.

In Figure 4 we report a series of boxplots summarizing which percentiles have been selected in the training steps for KFSD$_{smo}$, KFSD$_{tri}$ and KFSD$_{wei}$, and the following general remarks can be made. First, MM6 is the mixture model for which lower percentiles have been selected, and it is also a scenario in which our methods considerably outperform their
competitors. The need for a more local approach for MM6-data may explain the two observed facts about this mixture model. Second, lower and more local percentiles have been chosen for mixture models with nonlinear mean functions (MM4, MM5 and MM6) than for mixture models with linear mean functions (MM1, MM2 and MM3). Finally, the percentiles selected by means of the proposed training procedure seem to vary among datasets. However, except for MM3 and $\alpha = 0.02$, at least for half of the datasets a percentile not greater than the median has been chosen, which implies at most a moderately local approach.

5 REAL DATA STUDY: NITROGEN OXIDES (NO$_x$) DATA

Besides simulated data, we consider a real dataset which consists in nitrogen oxides (NO$_x$) emission level daily curves measured every hour close to an industrial area in Poblenou (Barcelona) and is available in the R package fda.usc (Febrero and Oviedo de la Fuente 2012). Outlier detection on this dataset was first performed by Febrero et al. (2008) where
these authors proposed $B_{\text{tri}}$ and $B_{\text{wei}}$. We carry on their study considering more methods and depths. According to Febrero et al. (2008), NO$_x$ are one of the most important pollutants, and it is important to identify outlying trajectories because these curves may both compromise any statistical analysis and be of special interest for further analysis. Precisely, the NO$_x$ levels were measured in $\mu g/m^3$ every hour of every day for the period 23/02/2005-26/06/2005, but only for 115 days of the period are available the 24 measurements. The curves of these 115 days compose the final NO$_x$ dataset. However, since the NO$_x$ dataset clearly includes working as well as nonworking day curves, following Febrero et al. (2008), it is more appropriate to consider two different datasets, that is, a sample of 76 working day curves (from now on, W) and another of 39 nonworking day curves (from now on, NW). Both W and NW are showed in Figure 5.

![Figure 5: NO$_x$ data: working (top) and non working (bottom) day curves.](image)

At first glance, each dataset may contain outliers, especially partial outliers, and therefore a local depth approach by means of KFSD$_{\text{smo}}$, KFSD$_{\text{tri}}$, and KFSD$_{\text{wei}}$ appears to be a good strategy to detect outliers. Besides them, we do outlier detection with all the methods used in Section 4. For all the procedures we use the same specifications as in Section 4 and we assume $\alpha = 0.05$. For each method, we report the labels of the curves detected as outliers.
in Table 8 and we highlight these curves in Figure 6.

Table 8: NO\textsubscript{x} data, Working and Nonworking datasets. Curves detected as outliers by FBP, B\textsubscript{tri}, B\textsubscript{wei}, FBG, FHD, KFSD\textsubscript{smo}, KFSD\textsubscript{tri} and KFSD\textsubscript{wei}.

|          | working days | nonworking w days |
|----------|--------------|-------------------|
| FBP+FMD  | -            | -                 |
| FBP+HMD  | 12, 16, 37   | 5, 7, 20, 21      |
| FBP+RTD  | 37           | 20                |
| FBP+IDD  | -            | 5, 7, 20          |
| FBP+MBD  | -            | -                 |
| FBP+FSD  | 37           | -                 |
| FBP+KFSD | 12, 16, 37   | 5, 7, 20, 21      |
| B\textsubscript{tri}+FMD | 16, 37 | 7 |
| B\textsubscript{tri}+HMD | 14, 16, 37 | 7, 20 |
| B\textsubscript{tri}+RTD | 16 | 7, 20 |
| B\textsubscript{tri}+IDD | 16, 37 | 7, 20 |
| B\textsubscript{tri}+MBD | 16, 37 | 7 |
| B\textsubscript{tri}+FSD | 14, 16, 37 | - |
| B\textsubscript{tri}+KFSD | 12, 14, 16, 37 | 7, 20 |
| B\textsubscript{wei}+FMD | 16 | 7, 20 |
| B\textsubscript{wei}+HMD | 16, 37 | 7, 20 |
| B\textsubscript{wei}+RTD | 16 | - |
| B\textsubscript{wei}+IDD | 16, 37 | 20 |
| B\textsubscript{wei}+MBD | 16 | 7 |
| B\textsubscript{wei}+FSD | 16, 37 | - |
| B\textsubscript{wei}+KFSD | 16, 37 | 7, 20 |
| FBG      | 16, 37       | -                 |
| FHD      | 12, 14, 16, 37 | 7, 20 |
| KFSD\textsubscript{smo} | 14, 16, 37 | 7, 20, 21 |
| KFSD\textsubscript{tri} | 12, 14, 16, 37 | 7, 20, 21 |
| KFSD\textsubscript{wei} | 11, 12, 13, 14, 15, 16, 37, 38 | 7, 20, 21 |

Figure 6: NO\textsubscript{x} dataset, curves detected as outliers in Table 8 working (top) and nonworking (bottom) days.
Concerning W, most of the methods detect as outlier day 37, which apparently shows a partial outlying behavior before noon and at the end of the day. Another day detected as outlier by many methods is day 16, whose curve has the highest morning peak. In addition to curves 16 and 37, KFSD\textsubscript{smo} detects as outlier curve 14, as other nine methods do, recognizing a seemingly outlying pattern in early hours of the day. Additionally, KFSD\textsubscript{tri} includes among the outliers also day 12, which may be atypical because of its behavior in early afternoon. Finally, KFSD\textsubscript{wei} detects as outliers the greatest number of curves. This last result may appear exaggerated, but all the curves that are outliers according to KFSD\textsubscript{wei} seem to have some partial deviations from the majority of curves. For example, day 13, whose curve is considered normal by the rest of the procedures, shows a peak at end of the day. Similar peaks can be observed also in other curves detected as outliers by other methods (e.g., days 16 and 37), which means that it may be occurring a masking effect to day 13’s detriment, and only KFSD\textsubscript{wei} points out this possibly outlying feature of the curve. Regarding the training step for KFSD to set $\sigma$, it gives as result the 70% percentile. Observing the first graph of Figure 5, it can be noticed that some curves have a likely outlying behavior, and this may be the reason why a weakly local approach for KFSD may be adequate enough.

In the case of NW, some methods detect no curves as outliers (e.g., all the FSD-based methods), exclusively three FBP-based methods flag day 5 as outlier, whereas days 7, 20 and 21 are detected as outliers by our methods as well as others. Days 7 and 20, which have two peaks, at the beginning and end of the day, are also flagged by other twelve and eight methods, respectively, while day 21, which shows a single peak in the first hours of the day, is considered atypical by only two other methods, which happen to be local (FBP+HMD and FBP+KFSD). This last result may be connected with what has been observed at the KFSD training step for selecting the percentile, i.e., the selection of the 30% percentile. Therefore, KFSD\textsubscript{smo}, KFSD\textsubscript{tri} and KFSD\textsubscript{wei} work with a strongly local percentile, and their results partially resemble the ones of the previously mentioned local techniques.
6 CONCLUSIONS

This paper proposes to tackle outlier detection in functional samples using the kernelized functional spatial depth as a tool. In Theorem 1 we presented a probabilistic result allowing to set a KFSD-threshold to identify outliers, but in practice it is necessary to observe two samples to apply Theorem 1. To overcome this practical limitation, we proposed KFSD$_{smo}$, KFSD$_{tri}$ and KFSD$_{wei}$ which are methods that can be applied when a unique functional sample is available and are based on both a probabilistic approach and smoothed resampling techniques.

We also proposed a new procedure to set the bandwidth $\sigma$ of KFSD that is based on obtaining training samples by means of smoothed resampling techniques. The general idea behind this procedure can be applied to other functional depths or methods with parameters that need to be set.

We investigated the performances of KFSD$_{smo}$, KFSD$_{tri}$ and KFSD$_{wei}$ by means of a simulation study. We focused on challenging scenarios with low magnitude, shape and partial outliers instead of high magnitude outliers. The results support our proposals. Along the simulation study, KFSD$_{wei}$, KFSD$_{tri}$ and KFSD$_{smo}$ attained the largest correct detection performances in most of the analyzed setups, but in some cases they paid a price in terms of false detection. However, KFSD$_{wei}$, KFSD$_{smo}$ and KFSD$_{tri}$ work with a given desired false alarm probability, and therefore higher false detection percentages than the competitors are due to the inherent structure of the methods. Concerning the remaining methods, there are competitors that in few scenarios outperformed our methods. However, in these few cases the differences are not great, especially for KFSD$_{wei}$ and KFSD$_{tri}$, and more important, these competitors do not show stability across scenarios in their results. Finally, we also considered a real application the NO$_x$ emission daily curves.

To conclude, two possible future research lines are: first, since KFSD is a depth whose local approach is in part based on the choice of the kernel function, it would be interesting to explore how the choice of different kernels affects the behavior of KFSD; second, since outlier detection can be seen as a special case of cluster analysis (it is a cluster problem with maximum two clusters, and one of them with size much smaller than the other, even 0),
a natural step ahead in our research may be the definition of KFSD-based cluster analysis procedures.

**ACKNOWLEDGMENTS**

This research was partially supported by Spanish Ministry of Science and Innovation grant ECO2011-25706 and by Spanish Ministry of Economy and Competition grant ECO2012-38442.

**A Appendix: Proof of Theorem 1**

As explained in Section 3, Theorem 1 is a functional extension of a result derived by Chen et al (2009) for KSD, and since they are closely related, next we report a sketch of the proof of Theorem 1. The proof for KSD is mostly based on an inequality known as McDiarmid’s inequality (McDiarmid 1989), which also applies to general probability spaces, and therefore to functional Hilbert spaces. We report this inequality in the next lemma:

**Lemma 1** *(McDiarmid 1989 [1.2])* Let $\Omega_1, \ldots, \Omega_n$ be probability spaces. Let $\Omega = \prod_{j=1}^n \Omega_j$ and let $X : \Omega \to \mathbb{R}$ be a random variable. For any $j \in \{1, \ldots, n\}$, let $(\omega_1, \ldots, \omega_j, \ldots, \omega_n)$ and $(\omega_1, \ldots, \hat{\omega}_j, \ldots, \omega_n)$ be two elements of $\Omega$ that differ only in their $j$th coordinates. Assume that $X$ is uniformly difference-bounded by $\{c_j\}$, that is, for any $j \in \{1, \ldots, n\}$,

$$|X(\omega_1, \ldots, \omega_j, \ldots, \omega_n) - X(\omega_1, \ldots, \hat{\omega}_j, \ldots, \omega_n)| \leq c_j. \quad (12)$$

Then, if $\mathbb{E}[X]$ exists, for any $\tau > 0$

$$\Pr(X - \mathbb{E}[X] \geq \tau) \leq \exp \left( \frac{-2\tau^2}{\sum_{j=1}^n c_j^2} \right).$$

In order to apply Lemma 1 to our problem, define

$$X(z_1, \ldots, z_{nz}) = -\frac{1}{nz} \sum_{i=1}^{nz} g(z_i, Y_{ny_i}|Y_{ny_i}), \quad (13)$$
whose expected value is given by

\[
E[X] = E_{z_i|Y_{ny}} \left[ -\frac{1}{n_Z} \sum_{i=1}^{n_Z} g(z_i, Y_{ny}|Y_{ny}) \right] = -E_{z_i|Y_{ny}} \left[ g(z_1, Y_{ny}|Y_{ny}) \right]. \tag{14}
\]

Now, for any \( j \in \{1, \ldots, n_Z\} \) and \( \hat{z}_j \in H \), the following inequality holds

\[
|X(z_1, \ldots, z_j, \ldots, z_{n_Z}) - X(z_1, \ldots, \hat{z}_j, \ldots, z_{n_Z})| \leq \frac{1}{n_Z},
\]

and it provides assumption (12) of Lemma 1. Therefore, for any \( \tau > 0 \)

\[
\Pr \left( E_{z_1|Y_{ny}} \left[ g(z_1, Y_{ny}|Y_{ny}) \right] - \frac{1}{n_Z} \sum_{i=1}^{n_Z} g(z_i, Y_{ny}|Y_{ny}) \geq \tau \right) \leq \exp \left( -2n_Z\tau^2 \right),
\]

and by the law of total probability

\[
\mathbb{E} \left[ \Pr \left( E_{z_1|Y_{ny}} \left[ g(z_1, Y_{ny}|Y_{ny}) \right] - \frac{1}{n_Z} \sum_{i=1}^{n_Z} g(z_i, Y_{ny}|Y_{ny}) \geq \tau \right) \right] = \Pr \left( E_{z_1|Y_{ny}} \left[ g(z_1, Y_{ny}) \right] - \frac{1}{n_Z} \sum_{i=1}^{n_Z} g(z_i, Y_{ny}) \geq \tau \right) \leq \exp \left( -2n_Z\tau^2 \right)
\]

Next, setting \( \delta = \exp \left( -2n_Z\tau^2 \right) \), and solving for \( \tau \), the following result is obtained:

\[
\tau = \sqrt{\frac{\ln 1/\delta}{2n_Z}}.
\]

Therefore,

\[
\Pr \left( E_{z_1|Y_{ny}} \left[ g(z_1, Y_{ny}) \right] \leq \frac{1}{n_Z} \sum_{i=1}^{n_Z} g(z_i, Y_{ny}) + \sqrt{\frac{\ln 1/\delta}{2n_Z}} \right) \geq 1 - \delta. \tag{15}
\]

However, Theorem 1 provides a probabilistic upper bound for \( E_x|Y_{ny} [g(x, Y_{ny})] \). First, recall that \( z_1 \sim Y_{mix} \) and note that

\[
E_{(z_1 \sim Y_{mix})|Y_{ny}} \left[ g(z_1, Y_{ny}) \right] = (1 - \alpha)E_{(z_1 \sim Y_{nor})|Y_{ny}} \left[ g(z_1, Y_{ny}) \right] + \alpha E_{(z_1 \sim Y_{out})|Y_{ny}} \left[ g(z_1, Y_{ny}) \right].
\]

Then, since \( E_{(z_1 \sim Y_{nor})|Y_{ny}} \left[ g(z_1, Y_{ny}) \right] = E_x|Y_{ny} [g(x, Y_{ny})], \) for \( \alpha > 0, \)
\[ \mathbb{E}_{x|Y_{n_y}}[g(x, Y_{n_y})] \leq \frac{1}{1 - \alpha} \mathbb{E}_{(z_1 \sim Y_{n_z})|Y_{n_y}}[g(z_1, Y_{n_y})]. \] (16)

Consequently, combining (15) and (16), and for \( r \geq \alpha \), we obtain

\[
\Pr \left( \mathbb{E}_{x|Y_{n_y}}[g(x, Y_{n_y})] \leq \frac{1}{1 - r} \left[ \frac{1}{n_z} \sum_{i=1}^{n_z} g(z_i, Y_{n_y}) + \sqrt{\frac{\ln 1/\delta}{2n_z}} \right] \right) \geq 1 - \delta,
\]

which completes the proof.

References

Barnett V, Lewis T (1994) Outliers in Statistical Data, vol 3. Wiley New York

Chakraborty A, Chaudhuri P (2014) On data depth in infinite dimensional spaces. Annals of the Institute of Statistical Mathematics 66:303–324

Chen Y, Dang X, Peng H, Bart HL (2009) Outlier detection with the kernelized spatial depth function. IEEE Transactions on Pattern Analysis and Machine Intelligence 31:288–305

Cuesta-Albertos JA, Nieto-Reyes A (2008) The random Tukey depth. Computational Statistics and Data Analysis 52:4979–4988

Cuevas A (2014) A partial overview of the theory of statistics with functional data. Journal of Statistical Planning and Inference 147:1–23

Cuevas A, Fraiman R (2009) On depth measures and dual statistics. A methodology for dealing with general data. Journal of Multivariate Analysis 100:753–766

Cuevas A, Febrero M, Fraiman R (2006) On the use of the bootstrap for estimating functions with functional data. Computational Statistics and Data Analysis 51:1063–1074

Febrero M, Oviedo de la Fuente M (2012) Statistical computing in functional data analysis: the R package fda.usc. Journal of Statistical Software 51:1–28
Febrero M, Galeano P, González-Manteiga W (2007) A functional analysis of nox levels: location and scale estimation and outlier detection. Computational Statistics 22:411–427

Febrero M, Galeano P, González-Manteiga W (2008) Outlier detection in functional data by depth measures, with application to identify abnormal nox levels. Environmetrics 19:331–345

Ferraty F, Vieu P (2006) Nonparametric Functional Data Analysis : Theory and Practice. Springer, New York

Fraiman R, Muniz G (2001) Trimmed means for functional data. Test 10:419–440

Horváth L, Kokoszka P (2012) Inference for Functional Data With Applications. Springer, New York

Hyndman RJ, Shang HL (2010) Rainbow plots, bagplots, and boxplots for functional data. Journal of Computational and Graphical Statistics 19:29–45

López-Pintado S, Romo J (2009) On the concept of depth for functional data. Journal of the American Statistical Association 104:718–734

McDiarmid C (1989) On the method of bounded differences. In: Survey in Combinatorics, Cambridge University Press, Cambridge, pp 148–188

Ramsay JO, Silverman BW (2005) Functional Data Analysis. Springer, New York

Sguera C, Galeano P, Lillo R (2014) Spatial depth-based classification for functional data. TEST 23:725–750

Sun Y, Genton MG (2011) Functional boxplots. Journal of Computational and Graphical Statistics 20:316–334

Tukey JW (1975) Mathematics and the picturing of data. In: Proceedings of the International Congress of Mathematicians, vol 2, pp 523–531