



Working Paper 14-14 (10)  
Statistics and Econometrics Series  
June 2014

Departamento de Estadística  
Universidad Carlos III de Madrid  
Calle Madrid, 126  
28903 Getafe (Spain)  
Fax (34) 91 624-98-49

## FUNCTIONAL OUTLIER DETECTION WITH A LOCAL SPATIAL DEPTH

Carlo Sguera, Pedro Galeano, Rosa Lillo

### Abstract

---

This paper proposes methods to detect outliers in functional datasets. We are interested in challenging scenarios where functional samples are contaminated by outliers that may be difficult to recognize. 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. Since outliers are usually among the least central curves, we introduce three new procedures that provide a threshold value for KFSD such that curves with depth values lower than the threshold are detected as outliers. 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

Departamento de Estadística, Universidad Carlos III de Madrid, C/ Madrid 126, 28903 Getafe (Madrid), e-mail: [csguera@est-econ.uc3m.es](mailto:csguera@est-econ.uc3m.es) (Carlo Sguera), [pgaleano@est-econ.uc3m.es](mailto:pgaleano@est-econ.uc3m.es) (Pedro Galeano), [lillo@est-econ.uc3m.es](mailto:lillo@est-econ.uc3m.es) (Rosa Lillo).

**Acknowledgements:** 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.

# 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](mailto:csguera@est-econ.uc3m.es))

**Pedro Galeano**

Department of Statistics, Universidad Carlos III de Madrid  
28903 Getafe (Madrid), Spain  
([pedro.galeano@uc3m.es](mailto:pedro.galeano@uc3m.es))

and

**Rosa E. Lillo**

Department of Statistics, Universidad Carlos III de Madrid  
28903 Getafe (Madrid), Spain  
([rosaelvira.lillo@uc3m.es](mailto:rosaelvira.lillo@uc3m.es))

## Abstract

This paper proposes methods to detect outliers in functional datasets. We are interested in challenging scenarios where functional samples are contaminated by outliers that may be difficult to recognize. 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. Since outliers are usually among the least central curves, we introduce three new procedures that provide a threshold value for KFSD such that curves with depth values lower than the threshold are detected as outliers. 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 such scenarios, and we refer to low magnitude, shape and partial outliers as “faint outliers” and to high magnitude outliers as “clear 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.

Indeed, methods of this nature already exist in the literature. For example, Febrero

et al (2008) proposed to label as outliers those curves with depth values lower than a certain threshold. As functional depths, they considered three alternatives, i.e., 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 alternative bootstrap procedures based on depth-based trimmed and weighted resampling, respectively. Also, Sun and Genton (2011) introduced the functional boxplot, which is constructed using the ranking of curves provided by the modified band depth (López-Pintado and Romo 2009). The proposed functional boxplot allows to detect outliers as well as the standard boxplot does. Note that the use of a functional depth is only one among the possible strategies for tackling the functional outlier detection problem. For example, 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.

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. In Sguera et al (2014), the local approach behind KFSD proved to be a good strategy in supervised classification problems with groups of curves not extremely clear-cut or under the presence of outliers. Similarly, in this paper we show that faint outliers are well detected by the methods based on KFSD that we propose.

We present a first result that allows to select a threshold for KFSD to detect outliers. This result is based on a probabilistic upper bound on a desired false alarm probability of detecting normal curves as outliers. However, its practical application requires the availability of two samples, circumstance rather uncommon in classical outlier detection problems. For this reason, we propose three different methods based on smoothed resampling techniques

that require instead a unique sample.

We study the performances of these resampling-based procedures in a simulation study where we consider faint outliers. Furthermore, we present a real outlier detection problem with environmental data, more precisely, nitrogen oxides ( $\text{NO}_x$ ) emission daily levels measured every hour close to an industrial area in Poblenou (Barcelona). Along both the simulation and real study, we compare our methods with the above-mentioned depth-based procedures proposed by Febrero et al (2008) and Sun and Genton (2011), using them with KFSD and six additional functional depths (the same as in Sguera et al 2014), as well as with the FPCA-based methods introduced by Hyndman and Shang (2010). The results that we observe support our proposals.

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 and present three new outlier detection methods based on KFSD. In Section 4 we report the results of our simulation study, whereas in Section 5 we perform outlier detection on the  $\text{NO}_x$  dataset. In Section 6 we draw some conclusions, and finally in the Appendix we report a sketch of the proof for a result given in the paper.

## 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).

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. Sguera et al (2014) have used depth-based methods in supervised functional classification problems, and it was observed that a local approach is preferable when the classes involved in the problem are not extremely different or distant. Here, we show that an approach based on the use of a local depth succeeds in detecting faint outliers such as low magnitude, shape or partial outliers. To illustrate this fact, we present the following example: first, we generated 10 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 10 datasets from a different mixture which produces faint shape outliers. In Figure 1 we report a contaminated dataset for each mixture.

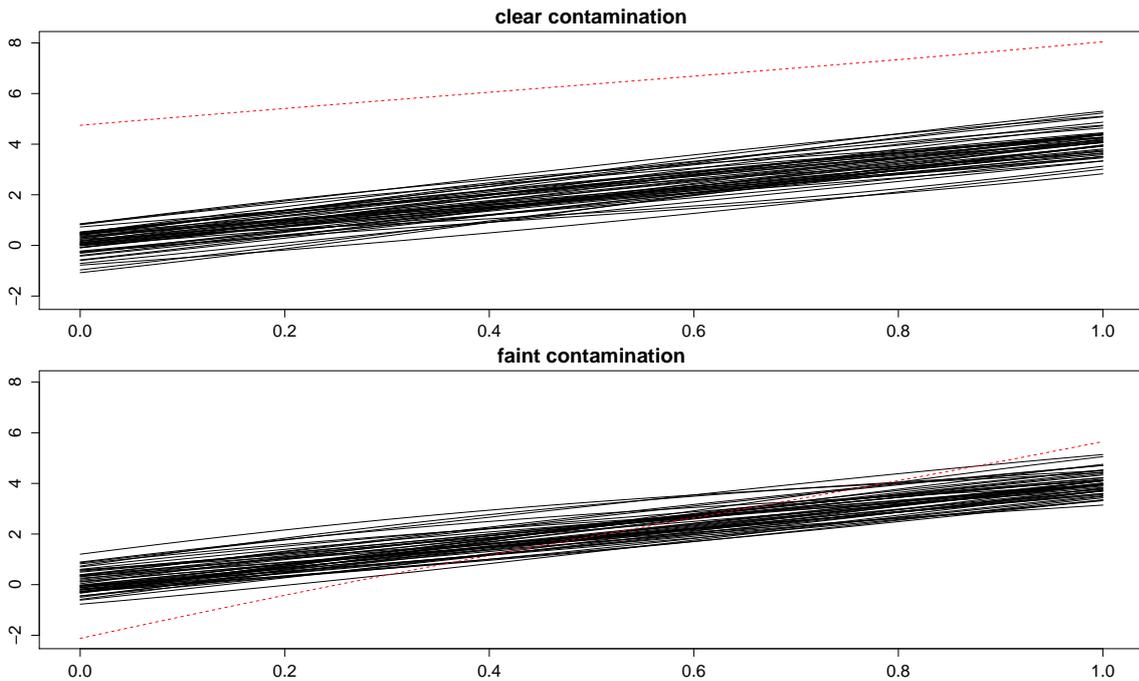


Figure 1: Examples of contaminated datasets: clear contamination (top) and faint contamination (bottom). The solid curves are normal curves and the dashed curves are outliers

Let  $n_{out,j}, j = 1, \dots, 10$ , 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  $n_{out,j}$  generated outliers. For both mixtures and each generated dataset, we registered how many times the depth of an outlier is indeed 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 clear outliers is an easier task than the ranking of faint outliers.

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

type of depths	global depths					local depths	
depths	FMD	RTD	IDD	MBD	FSD	HMD	KFSD
clear outliers	85.00	95.00	70.00	60.00	60.00	85.00	100.00
faint outliers	0.00	28.57	38.10	14.29	33.33	76.19	76.19

However, while the ranking of clear outliers is reasonably good in different cases, e.g., local KFSD (100%) or global RTD (95%), the ranking of faint outliers is markedly better with local depths, i.e., HMD and KFSD (both 76.19%). These results give an idea of the potential of local depths in ranking correctly faint outliers.

Next, since we employ KFSD to detect outliers, we recall its definition. First, let us introduce 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.,  $Y_n = \{y_1, \dots, y_n\}$ , the sample version of FSD has the following form:

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

As mentioned before, FSD is a global-oriented depth. 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

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

where  $\kappa$  is the kernel  $\kappa : \mathbb{H} \times \mathbb{H} \rightarrow \mathbb{R}$ ,  $\phi$  is the embedding map  $\phi : \mathbb{H} \rightarrow \mathbb{F}$  and  $\mathbb{F}$  is a feature space. Indeed, Sguera et al (2014) first defined the kernelized functional spatial depth (KFSD) in terms of  $\phi$ , that is,

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

which can be interpreted as a recoded version of  $FSD(x, Y)$  since  $KFSD(x, Y) = FSD(\phi(x), \phi(Y))$ , and whose sample version 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) allowed Sguera et al (2014) to provide an alternative expression of  $KFSD(x, Y_n)$ , in this case in terms of  $\kappa$ :

$$KFSD(x, Y_n) = 1 - \left( \frac{1}{n} \sum_{\substack{i, j=1; \\ y_i \neq x; y_j \neq x}}^n \frac{\kappa(x, x) + \kappa(y_i, y_j) - \kappa(x, y_i) - \kappa(x, y_j)}{\sqrt{\kappa(x, x) + \kappa(y_i, y_i) - 2\kappa(x, y_i)} \sqrt{\kappa(x, x) + \kappa(y_j, y_j) - 2\kappa(x, y_j)}} \right)^{1/2}, \quad (3)$$

Note that (3) 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

$$\kappa(x, y) = \exp \left( -\frac{\|x - y\|^2}{\sigma^2} \right), \quad (4)$$

where  $x, y \in \mathbb{H}$ . In turn, (4) 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.

### 3 OUTLIER DETECTION FOR FUNCTIONAL DATA

The outlier detection problem can be described as follows: let  $Y_n = \{y_1, \dots, y_n\}$  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.,

$$Y_{mix} = \begin{cases} Y_{nor}, & \text{with probability } 1 - \alpha, \\ Y_{out}, & \text{with probability } \alpha, \end{cases} \quad (5)$$

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  $t, b \in [0, 1]$ , the KFSD-based outlier detector for  $x \in \mathbb{H}$  is given by

$$g(x, Y_n) = \begin{cases} 1, & \text{if } KFSD(x, Y_n) \leq t, \\ \frac{t+b-KFSD(x, Y_n)}{b}, & \text{if } t < KFSD(x, Y_n) \leq t + b, \\ 0, & \text{if } KFSD(x, Y_n) > t + b, \end{cases} \quad (6)$$

where  $t$  is a threshold and  $b$  determines the transition rate of  $x$  from being an outlier (i.e.,  $g(x, Y_n) = 1$ ) to be a normal curve (i.e.,  $g(x, Y_n) = 0$ ). Clearly, (6) depends on the values of  $t$  and  $b$ . On the one hand, it is desirable a value of  $t$  capable of discriminating between  $x$  generated from  $Y_{nor}$  or  $Y_{out}$ . On the other hand, the role of  $b$  depends on the goal of the analysis. If the options “outlier” and “normal curve” are the only ones of interest,  $b$  should be set at 0. However, if there is interest in further analysis of “potential outliers”,  $b$  may be

allowed to be greater than 0. In our case, since the main goal is outlier detection and  $t$  is the key parameter to be set, we let  $b = 0$ .

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. Next, we extend this result to KFSD:

**Theorem 1** *Let  $Y_{n_Y} = \{y_i, \dots, y_{n_Y}\}$  and  $Z_{n_Z} = \{z_i, \dots, z_{n_Z}\}$  be two i. i. d. samples generated from the unknown mixture of random variables  $Y_{mix} \in \mathbb{H}$  described by (5), with  $\alpha > 0$ . Let  $g(\cdot, Y_{n_Y})$  be the outlier detector defined in (6). 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  $1 - \delta$ :*

$$\mathbb{E}_{x \sim Y_{nor}} [g(x, Y_{n_Y})] \leq \frac{1}{1-r} \left[ \frac{1}{n_Z} \sum_{i=1}^{n_Z} g(z_i, Y_{n_Y}) + \sqrt{\frac{\ln 1/\delta}{2n_Z}} \right], \quad (7)$$

where  $\mathbb{E}_{x \sim Y_{nor}}$  refers to the expected value with respect to  $x$  generated from  $Y_{nor}$ .

The proof of Theorem 1 is presented in the Appendix. Recall that the FAP has been defined as the probability that a normal observation  $x$  is classified as outlier. For the elements of Theorem 1,  $\Pr_{x \sim Y_{nor}} (g(x, Y_{n_Y}) = 1)$  is the FAP. If we set  $b = 0$ ,

$$\Pr_{x \sim Y_{nor}} (g(x, Y_{n_Y}) = 1) = \mathbb{E}_{x \sim Y_{nor}} [g(x, Y_{n_Y})].$$

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, but also a considerably large  $n_Z$ . To show the last point, recall that the right-hand side of (7) has to be controlled under the desired FAP and is in practice composed by two addends, with the second equal to  $\frac{1}{1-r} \sqrt{\frac{\ln 1/\delta}{2n_Z}}$ . For some normal values such as  $r = 0.05$ ,  $\delta = 0.05$  and  $n_Z = 50$ , we would

have  $\frac{1}{1-r} \sqrt{\frac{\ln 1/\delta}{2n_Z}} = 0.18$ , which is greater than a normal desired FAP such as 0.10, and it shows that the use of Theorem 1 may be compromised under some common situations.

We propose three solutions to overcome these limitations. Assume to observe a functional sample  $Y_n$  generated from an unknown mixture of random variables  $Y_{mix}$ . The goal is to identify which curves in  $Y_n$  are outliers, but in this situation there are not available two samples and Theorem 1 cannot be applied. We propose to use  $Y_n$  as  $Y_{n_Y}$ , and to obtain  $Z_{n_Z}$  by resampling with replacement from  $Y_n$ . In this way, we also solve the problematic issue related to the second addend of the right-hand side of (7) because it is possible to set  $n_Z$  as large as needed. 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). Then, the three resampling schemes that we consider are:

1. Simple resampling.
2. KFSD-based trimmed resampling: once obtained  $KFSD(y_i, Y_n), i = 1, \dots, n$ , it is possible to identify the  $\lceil \alpha_T \rceil\%$  of 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 obtained  $KFSD(y_i, Y_n), i = 1, \dots, n$ , 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  $Z_{n_Z}$  composed of curves different among them. 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  $(y_i(s_1), \dots, y_i(s_m))$  be the observed or eventually 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  $(\zeta(s_1), \dots, \zeta(s_m))$  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,  $j = 1, \dots, n_Z$ , the  $i$ th curve in  $Y_n$  has been sampled. Then, the discretized form of the  $j$ th curve in  $Z_{n_Z}$  would be given by  $(z_j(s_1), \dots, z_j(s_m)) = (y_i(s_1) + \zeta(s_1), \dots, y_i(s_m) + \zeta(s_m))$ , 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  $Z_{n_Z}$ , and we refer to them as *smo*, *tri* and *wei*, respectively. Then, for fixed  $\delta$ ,  $r$  and desired FAP, the threshold  $t$  for (6) is selected as the maximum value of  $t$  such that the right-hand side of (7) does not exceed the desired FAP. Let  $t^*$  be the selected threshold, which is then used in (6) with  $b = 0$  to compute  $g(y_i, Y_n)$ ,  $i = 1, \dots, n$ . 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  $\text{KFSD}_{smo}$ ,  $\text{KFSD}_{tri}$  and  $\text{KFSD}_{wei}$  depending on how  $Z_{n_Z}$  is obtained (*smo*, *tri* and *wei*, respectively; recall that  $Y_{n_Y} = Y_n$ ). As competitors of the proposed procedures, we consider the methods mentioned in Section 1 that we next 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  $KFSD_{smo}$ ,  $KFSD_{tri}$  and  $KFSD_{wei}$ , their competitors (FBP,  $B_{tri}$ ,  $B_{wei}$ , 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,  $B_{tri}$  and  $B_{wei}$ , 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  $Y_{mix}$  described by (5). The first three mixture models (MM1, MM2 and MM3) share  $Y_{nor}$ , with curves generated by

$$y(s) = 4s + \epsilon(s), \quad (8)$$

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

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

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

$$y(s) = u_1 \sin s + u_2 \cos s, \quad (9)$$

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 faint outliers of both shape and low magnitude nature. Under MM2, the outliers are generated by adding to (8) 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 (9), 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 (9) an observation from a  $N(0, (\frac{0.1}{2})^2)$ , and they turn out to be more irregular curves. Finally, under MM6, the outliers are generated by

$$y(s) = u_1 \sin s + \exp\left(\frac{0.69s}{2\pi}\right) u_4 \cos s, \quad (10)$$

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.

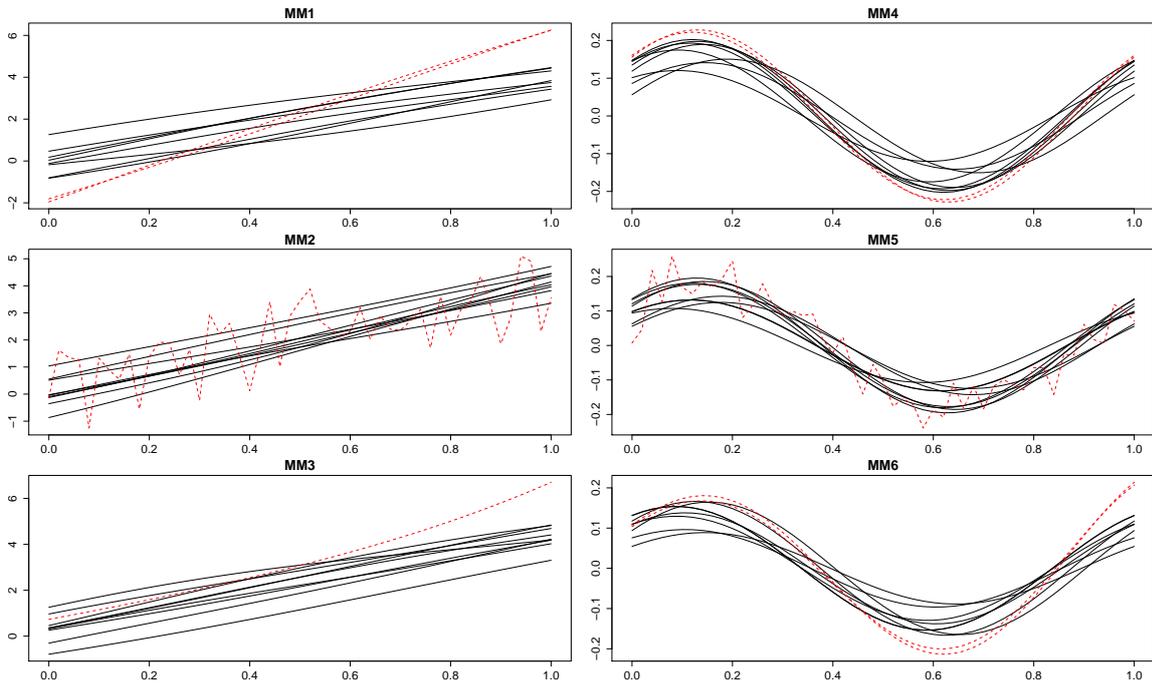


Figure 2: Examples of contaminated functional datasets generated by MM1, MM2, MM3, MM4, MM5 and MM6. Solid curves are normal curves and dashed curves are outliers

The details of the simulation study are the following: for each mixture model, we generated 100 datasets, each one composed of 50 curves. Two values of the contamination probability  $\alpha$  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,  $\mathbb{H}$  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:  $\gamma = 0.05$ ,  $B = 100$ ,  $\alpha_T = \alpha$ . 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  $Y_{n_Y} = Y_n$ ),  $\gamma = 0.05$ ,  $\alpha_T = \alpha$  (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  $KFSD_{wei}$ . Next, we describe the procedure, which is based on  $J$  replications. Let  $Y_n$  be the functional dataset on which outlier detection has to be done and let  $Y_{(n)} = \{y_{(1)}, \dots, y_{(n)}\}$  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  $\{p_1, \dots, p_K\}$  be the set of percentiles in use (in our case, as explained in Section 2,  $p_k = (10k)\%$ ,  $k \in \{1, \dots, K = 9\}$ ), and choose randomly a percentile from the set. For the  $j$ th replication,  $j \in \{1, \dots, J\}$ , denote the selected percentile as  $p^j$ . We use  $J = 20$  in the rest of the paper.
- II. Using  $p^j$ , compute  $KFSD_{p^j}(y_i, Y_n)$ ,  $i = 1, \dots, n$ , where the notation  $KFSD_{p^j}(\cdot, \cdot)$  is used to describe what percentile is used. For the  $j$ th replication, denote the KFSD-based ordered curves as  $y_{(1),j}, \dots, y_{(n),j}$ .
- III. Take  $y_{(1),j}, \dots, y_{(l_j),j}$ , 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  $y_{(1),j}^*, \dots, y_{(l_j),j}^*$ .
- IV. Repeat  $J$  times steps I.-III. to obtain a collection of  $L = \sum_{j=1}^J l_j$  peripheral curves, say  $Y_L$  (for an example, see Figure 3).

Next,  $Y_L$  acts as training sample according to the following steps: for each  $y_{(i),j}^* \in Y_L$ , ( $i \leq l_j$ ), and  $p_k \in \{p_1, \dots, p_K\}$ , compute  $KFSD_{p_k}(y_{(i),j}^*, Y_{-(i),j})$ , where  $Y_{-(i),j} = Y_n \setminus \{y_{(i),j}\}$ . At the end, a  $L \times K$  matrix is obtained, say  $D_{LK} = \{d_{lk}\}_{\substack{l=1, \dots, L \\ k=1, \dots, K}}$ , 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  $\{KFSD_{p_k}(y_1, Y_n), \dots, KFSD_{p_k}(y_n, Y_n), d_{lk}\}$ , e.g.  $r_{lk}$  is equal to 1 ( $n + 1$ ) if  $d_{lk}$  is the minimum (maximum) value in the vector. 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 lower 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

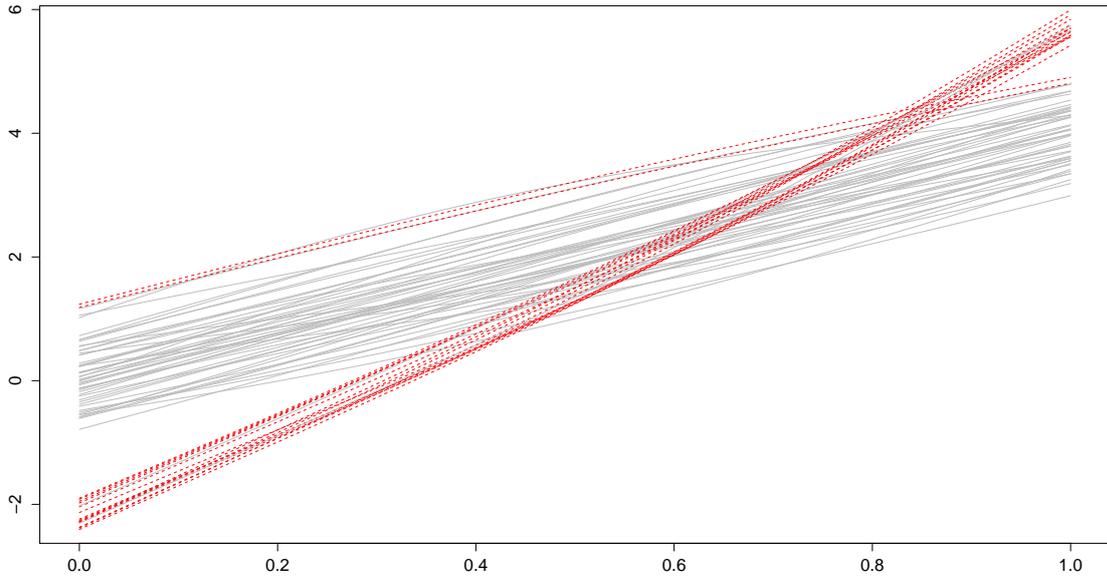


Figure 3: Example of a training sample of peripheral curves for a contaminated dataset generated by MM1 with  $\alpha = 0.05$ . The solid and shaded curves are the original curves (both normal and outliers). The dashed curves are the peripheral curves to use as training sample

detection percentage (c).<sup>1</sup> For each model, if a method is among the 5 best ones for both contamination probabilities  $\alpha$ , we report its label in bold.

The results in Tables 2-7 show that:

1.  $\text{KFSD}_{tri}$  and  $\text{KFSD}_{wei}$  are always among the 5 best methods.  $\text{KFSD}_{smo}$  is among the 5 best methods 10 times over 12, but when its performance is not among the 5 best, it is neither extremely far from the fifth method (MM2,  $\alpha = 0.02$ : 95.18% against 96.39%; MM3,  $\alpha = 0.02$ : 73.79% against 78.63%). The rest of the methods are among the 5 best procedures at most 5 times over 12 (FBP+HMD).
2. Regarding MM5 and MM6, our procedures are clearly the best options in terms of correct detection (c), and in the following order:  $\text{KFSD}_{wei}$ ,  $\text{KFSD}_{tri}$  and  $\text{KFSD}_{smo}$ . 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  $\text{KFSD}_{smo}$ ,  $\text{KFSD}_{tri}$  and  $\text{KFSD}_{wei}$ , i.e., 91.58%, 93.68% and 96.84%, respectively, against 67.37% of the fourth best method

---

<sup>1</sup>In presence of tie, we look at the false outlier detection percentage (f), preferring the method with lower f.

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  $KFSD_{wei}$

	$\alpha = 0.02$		$\alpha = 0.05$	
	c	f	c	f
FBP+FMD	44.34	1.23	43.86	0.73
FBP+HMD	<b>74.53</b>	0.94	72.81	0.61
FBP+RTD	61.32	0.57	63.16	0.31
FBP+IDD	55.66	0.61	61.84	0.34
FBP+MBD	49.06	1.33	50.44	0.69
FBP+FSD	62.26	0.67	61.84	0.40
FBP+KFSD	66.04	0.86	<b>74.12</b>	0.44
$B_{tri}$ +FMD	0.00	0.92	0.00	1.80
$B_{tri}$ +HMD	72.64	1.43	62.28	1.51
$B_{tri}$ +RTD	8.49	0.37	14.47	0.40
$B_{tri}$ +IDD	12.26	0.39	17.11	0.65
$B_{tri}$ +MBD	0.00	0.67	0.00	1.51
$B_{tri}$ +FSD	1.89	0.84	5.70	1.22
$B_{tri}$ +KFSD	70.75	1.57	57.89	1.49
$B_{wei}$ +FMD	0.00	1.23	0.00	1.53
$B_{wei}$ +HMD	71.70	1.16	46.49	0.57
$B_{wei}$ +RTD	10.38	1.25	7.46	1.17
$B_{wei}$ +IDD	14.15	2.29	14.04	2.62
$B_{wei}$ +MBD	0.00	1.14	0.00	1.30
$B_{wei}$ +FSD	1.89	1.33	3.07	1.17
$B_{wei}$ +KFSD	66.04	0.94	57.02	0.52
<b>FBG</b>	<b>100.00</b>	2.27	<b>97.81</b>	2.37
FHD	48.11	1.00	73.68	2.77
$KFSD_{smo}$	<b>89.62</b>	4.50	<b>85.09</b>	2.58
$KFSD_{tri}$	<b>89.62</b>	4.92	<b>92.11</b>	4.40
$KFSD_{wei}$	<b>97.17</b>	9.44	<b>96.93</b>	6.54

Table 4: MM3,  $\alpha = \{0.02, 0.05\}$ . Correct (c) and false (f) outlier detection percentages of FBP,  $B_{tri}$ ,  $B_{wei}$ , FBG, FHD,  $KFSD_{smo}$ ,  $KFSD_{tri}$  and  $KFSD_{wei}$

	$\alpha = 0.02$		$\alpha = 0.05$	
	c	f	c	f
FBP+FMD	65.69	0.92	49.19	0.97
<b>FBP+HMD</b>	<b>89.22</b>	0.57	<b>85.89</b>	0.63
FBP+RTD	86.27	0.45	76.61	0.34
FBP+IDD	79.41	0.51	70.56	0.38
FBP+MBD	74.51	0.88	59.27	0.84
FBP+FSD	79.41	0.51	73.79	0.42
<b>FBP+KFSD</b>	<b>89.22</b>	0.57	<b>83.06</b>	0.59
$B_{tri}$ +FMD	2.94	0.96	4.84	1.24
$B_{tri}$ +HMD	59.80	1.61	55.65	1.64
$B_{tri}$ +RTD	5.88	0.33	4.03	0.40
$B_{tri}$ +IDD	34.31	0.49	23.79	0.76
$B_{tri}$ +MBD	0.98	1.12	3.63	1.49
$B_{tri}$ +FSD	14.71	1.06	17.74	1.41
$B_{tri}$ +KFSD	59.80	1.65	47.98	1.39
$B_{wei}$ +FMD	2.94	1.10	5.24	0.84
$B_{wei}$ +HMD	59.80	1.25	37.90	0.80
$B_{wei}$ +RTD	19.61	0.92	12.90	0.78
$B_{wei}$ +IDD	29.41	2.67	20.97	2.67
$B_{wei}$ +MBD	0.98	1.31	3.23	1.26
$B_{wei}$ +FSD	16.67	1.10	11.29	0.90
$B_{wei}$ +KFSD	55.88	1.12	41.13	0.72
<b>FBG</b>	<b>86.27</b>	2.65	<b>78.63</b>	1.73
FHD	49.02	1.02	65.73	2.88
$KFSD_{smo}$	<b>89.22</b>	3.90	73.79	2.95
$KFSD_{tri}$	<b>90.20</b>	4.63	<b>83.47</b>	4.71
$KFSD_{wei}$	<b>97.06</b>	8.96	<b>90.32</b>	6.50

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  $KFSD_{wei}$

	$\alpha = 0.02$		$\alpha = 0.05$	
	c	f	c	f
FBP+FMD	99.09	1.08	<b>96.39</b>	0.84
FBP+HMD	96.36	0.96	<b>96.39</b>	0.88
FBP+RTD	<b>99.09</b>	0.61	94.78	0.25
FBP+IDD	<b>99.09</b>	0.70	95.18	0.38
FBP+MBD	99.09	1.06	<b>96.39</b>	0.82
FBP+FSD	99.09	0.57	94.78	0.36
FBP+KFSD	98.18	0.63	93.98	0.36
$B_{tri}$ +FMD	0.00	0.96	0.00	2.00
$B_{tri}$ +HMD	94.55	1.60	95.18	1.73
$B_{tri}$ +RTD	5.45	0.37	7.63	0.93
$B_{tri}$ +IDD	6.36	0.45	10.04	0.97
$B_{tri}$ +MBD	0.00	1.08	0.40	2.10
$B_{tri}$ +FSD	4.55	1.06	6.02	1.64
$B_{tri}$ +KFSD	99.09	1.60	96.39	1.56
$B_{wei}$ +FMD	0.00	1.41	0.00	1.39
$B_{wei}$ +HMD	94.55	0.94	83.53	0.32
$B_{wei}$ +RTD	7.27	1.51	8.43	1.89
$B_{wei}$ +IDD	8.18	2.49	8.84	2.86
$B_{wei}$ +MBD	0.00	1.29	0.40	1.54
$B_{wei}$ +FSD	6.36	1.43	4.82	1.41
$B_{wei}$ +KFSD	92.73	0.72	81.53	0.51
<b>FBG</b>	8.18	3.07	4.42	2.95
FHD	7.27	1.88	12.45	5.66
$KFSD_{smo}$	<b>100.00</b>	3.91	95.18	2.76
$KFSD_{tri}$	<b>100.00</b>	5.19	<b>97.99</b>	4.84
$KFSD_{wei}$	<b>100.00</b>	9.20	<b>99.60</b>	6.48

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

	$\alpha = 0.02$		$\alpha = 0.05$	
	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
$B_{tri}$ +FMD	64.29	0.18	46.80	0.15
$B_{tri}$ +HMD	43.88	0.06	20.40	0.21
$B_{tri}$ +RTD	27.55	1.08	14.80	0.80
$B_{tri}$ +IDD	67.35	0.59	<b>47.60</b>	0.46
$B_{tri}$ +MBD	66.33	0.14	43.20	0.06
$B_{tri}$ +FSD	<b>68.37</b>	0.12	46.80	0.13
$B_{tri}$ +KFSD	57.14	0.24	27.20	0.11
$B_{wei}$ +FMD	51.02	0.12	22.40	0.02
$B_{wei}$ +HMD	40.82	0.04	12.00	0.00
$B_{wei}$ +RTD	24.49	0.18	16.00	0.04
<b><math>B_{wei}</math>+IDD</b>	<b>90.82</b>	2.26	<b>73.60</b>	1.47
$B_{wei}$ +MBD	56.12	0.08	26.40	0.00
$B_{wei}$ +FSD	61.22	0.08	28.00	0.00
$B_{wei}$ +KFSD	56.12	0.12	20.40	0.00
<b>FBG</b>	9.18	0.53	6.80	1.09
FHD	51.02	1.02	37.60	4.34
$KFSD_{smo}$	<b>87.76</b>	2.16	<b>50.00</b>	1.24
$KFSD_{tri}$	<b>91.84</b>	3.00	<b>64.80</b>	2.91
$KFSD_{wei}$	<b>95.92</b>	5.08	<b>62.00</b>	3.35

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

	$\alpha = 0.02$		$\alpha = 0.05$	
	c	f	c	f
FBP+FMD	55.56	0.00	54.00	0.00
FBP+HMD	66.67	0.00	68.40	0.04
FBP+RTD	57.58	0.00	54.40	0.00
FBP+IDD	52.53	0.00	56.00	0.00
FBP+MBD	55.56	0.00	55.20	0.00
FBP+FSD	55.56	0.00	55.60	0.00
FBP+KFSD	60.61	0.00	59.20	0.00
$B_{tri}$ +FMD	3.03	0.16	2.80	0.36
$B_{tri}$ +HMD	<b>96.97</b>	0.16	<b>89.20</b>	0.17
$B_{tri}$ +RTD	12.12	1.31	18.40	1.37
$B_{tri}$ +IDD	22.22	0.84	29.20	0.63
$B_{tri}$ +MBD	3.03	0.18	3.20	0.32
$B_{tri}$ +FSD	29.29	0.18	29.20	0.29
$B_{tri}$ +KFSD	90.91	0.27	<b>91.20</b>	0.19
$B_{wei}$ +FMD	3.03	0.22	2.40	0.19
$B_{wei}$ +HMD	<b>93.94</b>	0.02	71.20	0.00
$B_{wei}$ +RTD	16.16	0.41	20.00	0.38
$B_{wei}$ +IDD	23.23	3.20	21.60	2.74
$B_{wei}$ +MBD	4.04	0.24	3.60	0.23
$B_{wei}$ +FSD	26.26	0.12	21.60	0.08
$B_{wei}$ +KFSD	88.89	0.12	68.00	0.04
FBG	0.00	1.02	0.40	0.04
FHD	4.04	1.96	12.80	5.64
$KFSD_{smo}$	<b>98.99</b>	1.82	<b>94.00</b>	0.44
$KFSD_{tri}$	<b>98.99</b>	2.61	<b>98.00</b>	2.11
$KFSD_{wei}$	<b>100.00</b>	4.61	<b>98.40</b>	2.11

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

	$\alpha = 0.02$		$\alpha = 0.05$	
	c	f	c	f
FBP+FMD	48.42	0.00	44.19	0.00
FBP+HMD	60.00	0.18	<b>62.92</b>	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
$B_{tri}$ +FMD	30.53	0.16	35.21	0.32
$B_{tri}$ +HMD	<b>67.37</b>	0.24	50.94	0.15
$B_{tri}$ +RTD	22.11	1.06	17.23	0.61
$B_{tri}$ +IDD	32.63	0.57	20.97	0.51
$B_{tri}$ +MBD	28.42	0.24	31.46	0.36
$B_{tri}$ +FSD	50.53	0.20	44.94	0.21
$B_{tri}$ +KFSD	66.32	0.22	48.31	0.13
$B_{wei}$ +FMD	25.26	0.22	18.35	0.06
$B_{wei}$ +HMD	<b>67.37</b>	0.12	38.95	0.00
$B_{wei}$ +RTD	41.05	0.31	34.46	0.19
$B_{wei}$ +IDD	33.68	2.34	23.22	1.75
$B_{wei}$ +MBD	23.16	0.18	17.98	0.15
$B_{wei}$ +FSD	43.16	0.14	29.59	0.11
$B_{wei}$ +KFSD	64.21	0.14	43.45	0.00
FBG	17.89	0.02	14.98	0.06
FHD	52.63	1.02	<b>61.80</b>	2.85
$KFSD_{smo}$	<b>91.58</b>	2.08	<b>71.16</b>	0.95
$KFSD_{tri}$	<b>93.68</b>	2.69	<b>82.02</b>	2.49
$KFSD_{wei}$	<b>96.84</b>	4.69	<b>83.15</b>	2.75

( $B_{wei}$ +HMD), that is, we observe differences greater than 20%, and approaching 30% if  $KFSD_{wei}$  and  $B_{wei}$ +HMD are compared.

- 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  $B_{tri}$  and  $B_{wei}$  when used with local depths, are closer to the results of  $KFSD_{wei}$ .
- 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  $KFSD_{wei}$ .

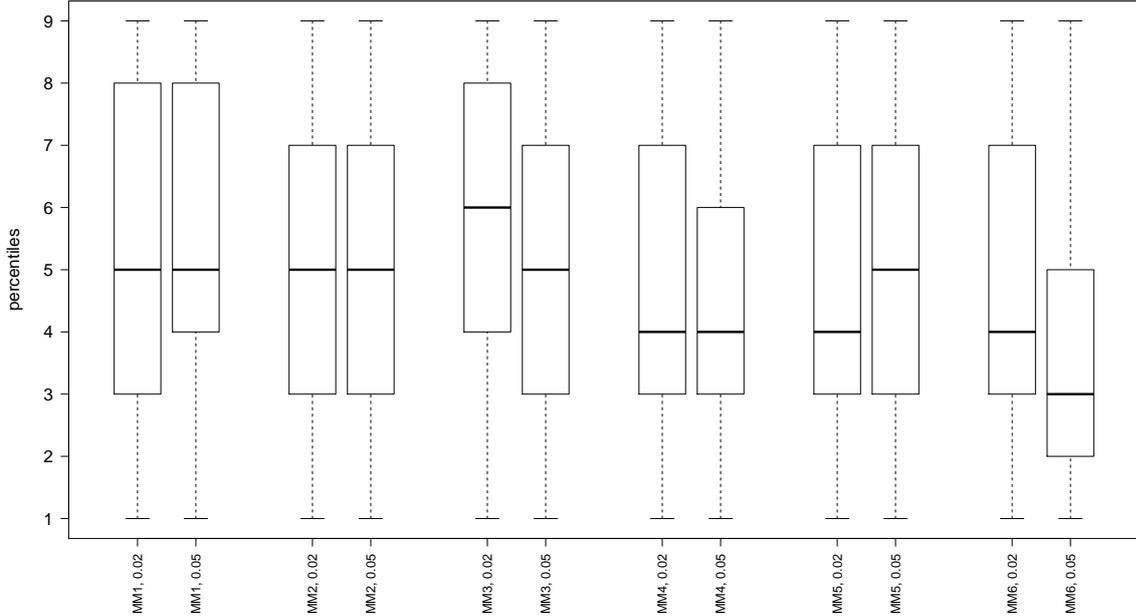


Figure 4: Boxplots of the percentiles selected in the training steps of the simulation study for  $KFSD_{smo}$ ,  $KFSD_{tri}$  and  $KFSD_{wei}$

Observing Figure 4, the following general remarks can be done. 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<sub>x</sub>) DATA

Besides simulated data, we consider a real dataset which consists in nitrogen oxides (NO<sub>x</sub>) emission level daily curves measured every hour close to an industrial area in Poblenou (Barcelona). The dataset is available in the R package `fda.usc` (Febrero and Oviedo de la Fuente 2012) and outlier detection on these data was first performed by Febrero et al (2008) in the paper where  $B_{tri}$  and  $B_{wei}$  were presented. We enhance their study by considering more methods and depths.

According to Febrero et al (2008), NO<sub>x</sub> 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.

More in details, the NO<sub>x</sub> 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 was possible to measure the NO<sub>x</sub> at every hour. These 115 curves compose the final NO<sub>x</sub> dataset. However, since the NO<sub>x</sub> 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: at first glance, it seems that each dataset may contain outliers, especially partial outliers.

Because of the possible presence of faint outliers, a local depth approach by means of  $KFSD_{smo}$ ,  $KFSD_{tri}$  and  $KFSD_{wei}$  may 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 in Figure 6 we highlight these curves.

For what concerns 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 is the one with the highest morning peak. In addition to curves 16 and 37,  $KFSD_{smo}$  detects as outlier curve 14, as other nine methods do, recognizing a seeming outlying pattern in early hours of the

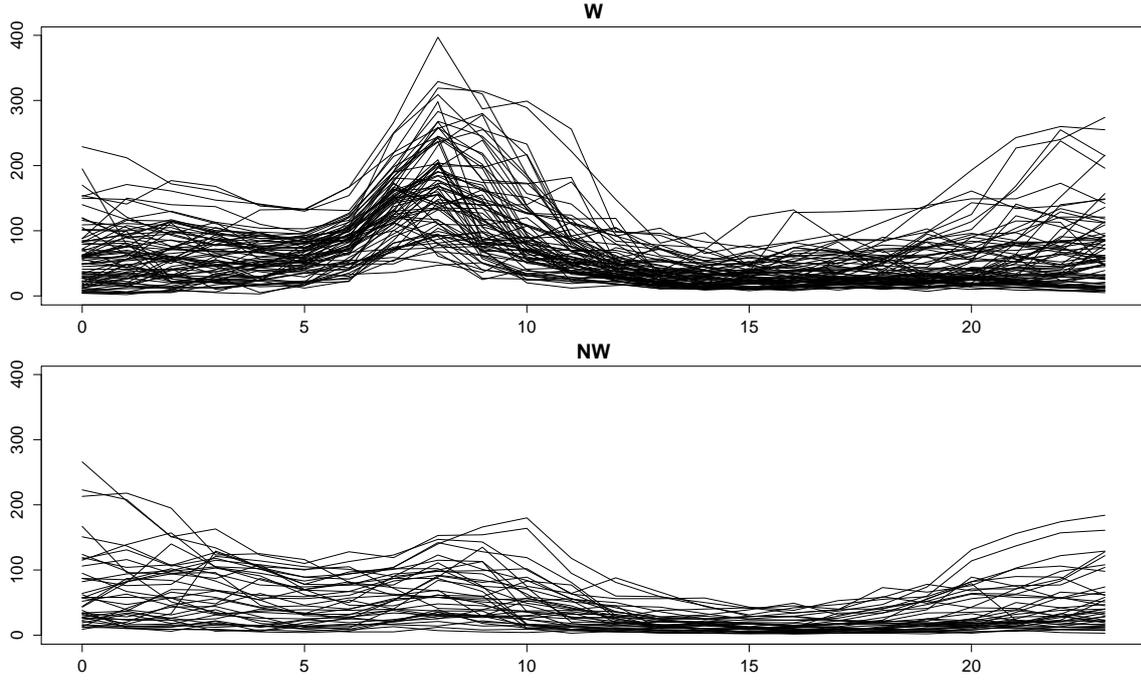


Figure 5:  $\text{NO}_x$  data: working (top) and non working (bottom) day curves.

Table 8:  $\text{NO}_x$  data, W and NW datasets. Curves detected as outliers by FBP,  $B_{tri}$ ,  $B_{wei}$ , FBG, FHD,  $\text{KFSD}_{smo}$ ,  $\text{KFSD}_{tri}$  and  $\text{KFSD}_{wei}$

	w days	non w days
	detected outliers	
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
$B_{tri}$ +FMD	16, 37	7
$B_{tri}$ +HMD	14, 16, 37	7, 20
$B_{tri}$ +RTD	14, 16, 37	-
$B_{tri}$ +IDD	11, 14, 16, 37	-
$B_{tri}$ +MBD	16, 37	7
$B_{tri}$ +FSD	14, 16, 37	-
$B_{tri}$ +KFSD	12, 14, 16, 37	7, 20
$B_{wei}$ +FMD	16	7
$B_{wei}$ +HMD	16, 37	7, 20
$B_{wei}$ +RTD	14, 16, 37, 38	-
$B_{wei}$ +IDD	16, 37	-
$B_{wei}$ +MBD	16	7
$B_{wei}$ +FSD	16, 37	-
$B_{wei}$ +KFSD	16, 37	7, 20
FBG	16, 37	-
FHD	12, 14, 16, 37	7, 20
$\text{KFSD}_{smo}$	14, 16, 37	7, 20, 21
$\text{KFSD}_{tri}$	12, 14, 16, 37	7, 20, 21
$\text{KFSD}_{wei}$	11, 12, 13, 14, 15, 16, 37, 38	7, 20, 21

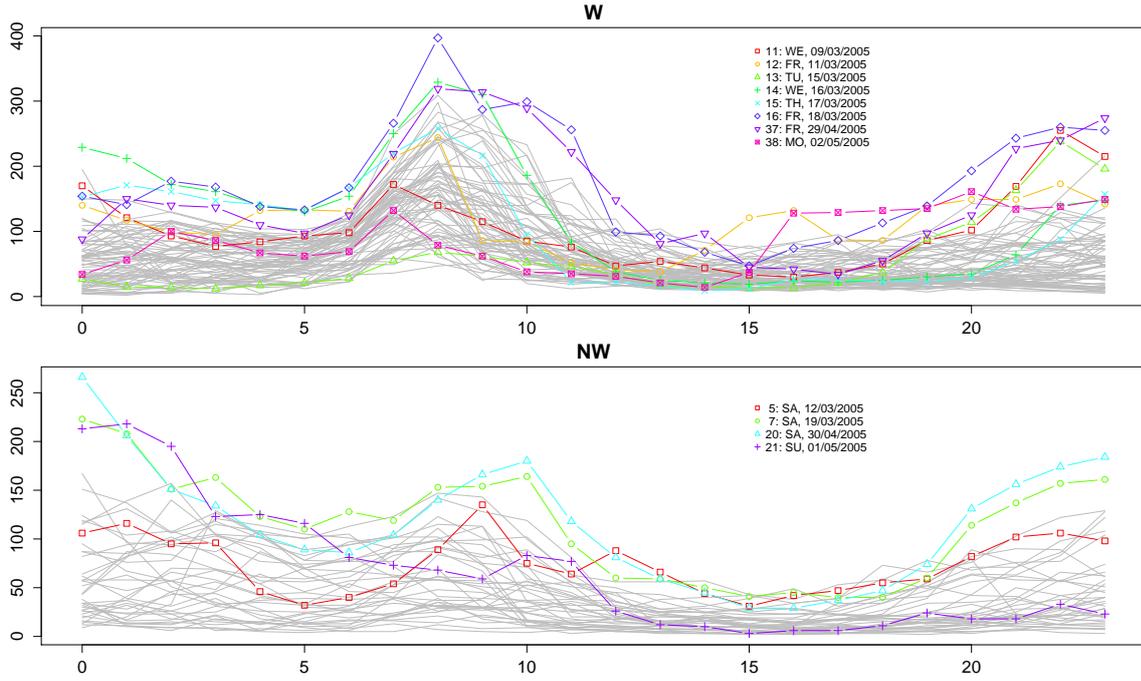


Figure 6:  $\text{NO}_x$  dataset, curves detected as outliers in Table 8: working (top) and non working (bottom) days

day. Additionally,  $\text{KFSD}_{tri}$  includes among the outliers also day 12, which may be atypical because of its behavior in early afternoon. Finally,  $\text{KFSD}_{wei}$  detects as outliers the greatest number of curves. This last result may appear exaggerated, but all the curves that are outliers according to  $\text{KFSD}_{wei}$  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  $\text{KFSD}_{wei}$  points out this possibly outlying feature of the curve. Regarding the training step for  $\text{KFSD}$  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  $\text{KFSD}$  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, among others, our methods. 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,  $\text{KFSD}_{smo}$ ,  $\text{KFSD}_{tri}$  and  $\text{KFSD}_{wei}$  work with a strongly local percentile, and their results partially resemble the ones of the previously mentioned local techniques.

## 6 CONCLUSIONS

This paper proposes three methods to detect outliers in functional samples based on the kernelized functional spatial depth (Sguera et al 2014). We presented a way to set a KFSD-threshold to identify outliers in Theorem 1. In practice, it is necessary to observe two samples to apply Theorem 1, and one sample must have a considerably large size. To overcome this practical limitation, we proposed  $\text{KFSD}_{smo}$ ,  $\text{KFSD}_{tri}$  and  $\text{KFSD}_{wei}$ : these methods are based on smoothed resampling techniques and, more important, they can be applied when a unique functional sample is available, no matter its size.

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  $\text{KFSD}_{smo}$ ,  $\text{KFSD}_{tri}$  and  $\text{KFSD}_{wei}$  by means of a simulation study. We focused on challenging scenarios with low magnitude, shape and partial outliers (faint outliers) instead of high magnitude outliers (clear outliers). The results support our proposals. Along the simulation study,  $\text{KFSD}_{wei}$ ,  $\text{KFSD}_{tri}$  and  $\text{KFSD}_{smo}$  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,  $\text{KFSD}_{wei}$ ,  $\text{KFSD}_{smo}$  and  $\text{KFSD}_{tri}$  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  $\text{KFSD}_{wei}$  and  $\text{KFSD}_{tri}$ , and more important, these competitors do not show stability across scenarios in

their results. Finally, we also considered a real application the  $\text{NO}_x$  emission daily curves.

To conclude, we would like to mention two possible future research lines. First, KFSD is a depth whose local approach is in part based on the choice of the kernel function. Therefore, it would be interesting to explore how the choice of different kernels affects the behavior of KFSD. Second, outlier detection can be seen as a special case of cluster analysis since 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

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  $\Omega_1, \dots, \Omega_n$  be probability spaces. Let  $\Omega = \prod_{j=1}^n \Omega_j$  and let  $X : \Omega \rightarrow \mathbb{R}$  be a random variable. For any  $j \in \{1, \dots, n\}$ , let  $(\omega_1, \dots, \omega_j, \dots, \omega_n)$  and  $(\omega_1, \dots, \hat{\omega}_j, \dots, \omega_n)$  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  $j \in \{1, \dots, n\}$ ,*

$$|X(\omega_1, \dots, \omega_j, \dots, \omega_n) - X(\omega_1, \dots, \hat{\omega}_j, \dots, \omega_n)| \leq c_j. \quad (11)$$

*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

$$X(z_1, \dots, z_{n_Z}) = -\frac{1}{n_Z} \sum_{i=1}^{n_Z} g(z_i, Y_{n_Y} | Y_{n_Y}), \quad (12)$$

whose expected value is given by

$$\mathbb{E}[X] = \mathbb{E}\left[-\frac{1}{n_Z} \sum_{i=1}^{n_Z} g(z_i, Y_{n_Y} | Y_{n_Y})\right] = -\mathbb{E}_{z_1 \sim Y_{mix}} [g(z_1, Y_{n_Y} | Y_{n_Y})]. \quad (13)$$

Now, for any  $j \in \{1, \dots, n_Z\}$  and  $\hat{z}_j \in \mathbb{H}$ , the following inequality holds

$$|X(z_1, \dots, z_j, \dots, z_{n_Z}) - X(z_1, \dots, \hat{z}_j, \dots, z_{n_Z})| \leq \frac{1}{n_Z},$$

and it provides assumption (11) of Lemma 1. Therefore, for any  $\tau > 0$

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

and by the law of total probability

$$\begin{aligned} & \mathbb{E}\left[\Pr\left(\mathbb{E}_{z_1 \sim Y_{mix}} [g(z_1, Y_{n_Y} | Y_{n_Y})] - \frac{1}{n_Z} \sum_{i=1}^{n_Z} g(z_i, Y_{n_Y} | Y_{n_Y}) \geq \tau\right)\right] \\ &= \Pr\left(\mathbb{E}_{z_1 \sim Y_{mix}} [g(z_1, Y_{n_Y})] - \frac{1}{n_Z} \sum_{i=1}^{n_Z} g(z_i, Y_{n_Y}) \geq \tau\right) \leq \exp(-2n_Z\tau^2) \end{aligned}$$

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

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

Therefore,

$$\Pr\left(\mathbb{E}_{z_1 \sim Y_{mix}} [g(z_1, Y_{n_Y})] \leq \frac{1}{n_Z} \sum_{i=1}^{n_Z} g(z_i, Y_{n_Y}) + \sqrt{\frac{\ln 1/\delta}{2n_Z}}\right) \geq 1 - \delta. \quad (14)$$

However, Theorem 1 provides a probabilistic upper bound for  $\mathbb{E}_{x \sim Y_{nor}} [g(x, Y_{n_Y})]$ . First, note that

$$\mathbb{E}_{x \sim Y_{mix}} [g(x, Y_{n_Y})] = (1 - \alpha) \mathbb{E}_{x \sim Y_{nor}} [g(x, Y_{n_Y})] + \alpha \mathbb{E}_{x \sim Y_{out}} [g(x, Y_{n_Y})],$$

and then, for  $\alpha > 0$ ,

$$\mathbb{E}_{x \sim Y_{nor}} [g(x, Y_{n_Y})] \leq \frac{1}{1 - \alpha} \mathbb{E}_{x \sim Y_{mix}} [g(x, Y_{n_Y})] = \frac{1}{1 - \alpha} \mathbb{E}_{z_1 \sim Y_{mix}} [g(z_1, Y_{n_Y})]. \quad (15)$$

Consequently, combining (14) and (15), we obtain

$$\Pr \left( \mathbb{E}_{x \sim Y_{nor}} [g(x, Y_{n_Y})] \leq \frac{1}{1 - r} \left[ \frac{1}{n_Z} \sum_{i=1}^{n_Z} g(z_i, Y_{n_Y}) + \sqrt{\frac{\ln 1/\delta}{2n_Z}} \right] \right) \geq 1 - \delta,$$

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* (forthcoming), DOI : 10.1007/s11749-014-0379-1
- 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