# Functional Boxplot

**Article** *in* Journal of Computational and Graphical Statistics · October 2010

DOI: 10.2307/23110490

CITATIONS
41

READS
1,079

**2 authors:**

Ying Sun
King Abdullah University of Science and Technology
**103** PUBLICATIONS **1,006** CITATIONS
SEE PROFILE

Marc G. Genton
King Abdullah University of Science and Technology
**301** PUBLICATIONS **7,863** CITATIONS
SEE PROFILE

*Supplementary materials for this article are available online.*
*Please click the JCGS link at http://pubs.amstat.org.*

# Functional Boxplots

Ying SUN and Marc G. GENTON

This article proposes an informative exploratory tool, the functional boxplot, for visualizing functional data, as well as its generalization, the enhanced functional boxplot. Based on the center outward ordering induced by band depth for functional data, the descriptive statistics of a functional boxplot are: the envelope of the 50% central region, the median curve, and the maximum non-outlying envelope. In addition, outliers can be detected in a functional boxplot by the 1.5 times the 50% central region empirical rule, analogous to the rule for classical boxplots. The construction of a functional boxplot is illustrated on a series of sea surface temperatures related to the El Niño phenomenon and its outlier detection performance is explored by simulations. As applications, the functional boxplot and enhanced functional boxplot are demonstrated on children growth data and spatio-temporal U.S. precipitation data for nine climatic regions, respectively. This article has supplementary material online.

**Key Words:** Depth; Functional data; Growth data; Precipitation data; Space–time data; Visualization.

## 1. INTRODUCTION

Functional data analysis is an attractive approach to study complex data in statistics. In many statistical experiments, the observations are functions by nature, such as temporal curves or spatial surfaces, where the basic unit of information is the entire observed function rather than a string of numbers. Such functional data appear in many fields, including meteorology, biology, medicine, and engineering. Human growth curves, weather station temperatures, gene expression signals, medical images, and human speech are all real-life examples; see, for example, the work of Dryden and Mardia (1998), Fletcher et al. (2004), and Ramsay and Silverman (2005).

To analyze functional data, researchers often used mathematical models, among which Ramsay and Silverman (2005) provided various parametric methods while Ferraty and Vieu (2006) developed detailed nonparametric techniques. Quantile regression, as a popular model-based method, has been widely used, and many economic applications were discussed by Fitzenberger, Koenker, and Machado (2002). In contrast to model-based analy-

Ying Sun is Graduate Student (E-mail: *sunwards@stat.tamu.edu*) and Marc G. Genton is Professor (E-mail: *genton@stat.tamu.edu*), Department of Statistics, Texas A&M University, College Station, TX 77843-3143.


*Journal of Computational and Graphical Statistics, Volume 20, Number 2, Pages 316–334*
*DOI: 10.1198/jcgs.2011.09224*

sis, visualization methods often help to display the data, highlight their characteristics, and reveal interesting features. For functional data, Hyndman and Shang (2010) proposed two graphical methods with outlier detection capability: the functional bagplot and the functional highest density region boxplot, both of which are based on the first two robust principal component scores. They applied the bivariate bagplot (Rousseeuw, Ruts, and Tukey 1999) to the first two robust principal component scores, and then mapped the features of the bagplot into the functional space. In this article, we aim to develop visualization tools for functional data directly in the functional space rather than in the feature space that requires principal component analysis techniques.

It is well known that the boxplot is a graphical method for displaying five descriptive statistics: the median, the first and third quartiles, and the non-outlying minimum and maximum observations. A boxplot may also indicate which observations, if any, can be considered as outliers. First introduced by Tukey (1970) and Tukey (1977, pp. 39–43) in exploratory data analysis, boxplots have evolved into a straightforward but informative method in data interpretation. The first step to construct a boxplot is the data ordering. In the univariate setting, the ranking is simply from the smallest observation to the largest. However, multivariate ordering is much more complicated and has attracted considerable interest over the years. To generalize order statistics or ranks to the multivariate setting, different versions of data depth have been introduced to measure how deep (central) or outlying an observation is. Examples of data depth include the Mahalanobis depth (Mahalanobis 1936), the Tukey halfspace location depth (Tukey 1975), the Oja depth (Oja 1983), the simplicial depth (Liu 1990), the majority depth (Singh 1991), and the likelihood depth (Fraiman and Meloche 1999). Vardi and Zhang (2000) proposed an $L_1$-depth which can be extended to functional data. Febrero, Galeano, and González-Manteiga (2007, 2008) have reviewed a series of functional depths, such as the functional depth of Fraiman and Muniz (2001), the functional depth of Cuevas, Febrero, and Fraiman (2006), and the random projection functional depth of Cuevas, Febrero, and Fraiman (2007).

For functional data, López-Pintado and Romo (2009) recently introduced a notion of band depth (BD). It allows for ordering a sample of curves from the center outward and, thus, introduces a measure to define functional quantiles and the centrality or outlyingness of an observation. Having the ranks of curves, the functional boxplot is a natural extension of the classical boxplot and is an appealing visualization tool for functional data.

This article is organized as follows. Section 2 explains the definition of band depth for functional data and its modified version. Section 3 illustrates the construction of functional boxplots and enhanced functional boxplots, as well as the associated outlier detection rule. Simulation results on the performance of our outlier detection method are reported in Section 4. The visualization capabilities of the functional boxplots are demonstrated in Section 5 when applied to classical functional data and a space-time dataset. A discussion is provided in Section 6.

## 2. BAND DEPTH FOR FUNCTIONAL DATA

In functional data analysis, each observation is a real function $y_i(t)$, $i = 1, \ldots, n$, $t \in \mathcal{I}$, where $\mathcal{I}$ is an interval in $\mathbb{R}$. The band depth for functional data provides a method to order

all the sample curves. Indeed, we can compute the band depths of all the sample curves and order them according to decreasing depth values. Let $y_{[i]}(t)$ denote the sample curve associated with the $i$th largest band depth value. We view $y_{[1]}(t), \ldots, y_{[n]}(t)$ as order statistics, with $y_{[1]}(t)$ being the deepest (most central) curve or simply the median curve, and $y_{[n]}(t)$ being the most outlying curve. The implication is that a smaller rank is associated with a more central position with respect to the sample curves. The order statistics induced by a band depth start from the most central sample curve and move outward in all directions. Therefore, they are different from the usual order statistics which are simply ordered from the smallest sample value to the largest.

With this basic idea, López-Pintado and Romo (2009) introduced the band depth concept through a graph-based approach. The graph of a function $y(t)$ is the subset of the plane $G(y) = \{(t, y(t)) : t \in \mathcal{I}\}$. The band in $\mathbb{R}^2$ delimited by the curves $y_{i_1}, \ldots, y_{i_k}$ is $B(y_{i_1}, \ldots, y_{i_k}) = \{(t, x(t)) : t \in \mathcal{I}, \min_{r=1,\ldots,k} y_{i_r}(t) \le x(t) \le \max_{r=1,\ldots,k} y_{i_r}(t)\}$. Let $J$ be the number of curves determining a band, where $J$ is a fixed value with $2 \le J \le n$. If $Y_1(t), \ldots, Y_n(t)$ are independent copies of the stochastic process $Y(t)$ generating the observations $y_1(t), \ldots, y_n(t)$, the population version of the band depth for a given curve $y(t)$ with respect to the probability measure $P$ is defined as

$$\mathrm{BD}_J(y, P) = \sum_{j=2}^{J} \mathrm{BD}^{(j)}(y, P) = \sum_{j=2}^{J} P\{G(y) \subset B(Y_1, \ldots, Y_j)\},$$

where $B(Y_1, \ldots, Y_j)$ is a band delimited by $j$ random curves. The sample version of $\mathrm{BD}^{(j)}(y, P)$ is obtained by computing the fraction of the bands determined by $j$ different sample curves containing the whole graph of the curve $y(t)$. In other words, $\mathrm{BD}_n^{(j)}(y) = \binom{n}{j}^{-1} \sum_{1 \le i_1 < i_2 < \cdots < i_j \le n} I\{G(y) \subseteq B(y_{i_1}, \ldots, y_{i_j})\}$, where $I\{\cdot\}$ denotes the indicator function. The implication is that by computing the fraction of the bands containing the curve $y(t)$, the bigger the value of band depth, the more central position the curve has. Then, the sample band depth of a curve $y(t)$ is

$$\mathrm{BD}_{n,J}(y) = \sum_{j=2}^{J} \mathrm{BD}_n^{(j)}(y). \tag{2.1}$$

Instead of considering the indicator function, López-Pintado and Romo (2009) also proposed a more flexible definition, the modified band depth (MBD), by measuring the proportion of time that a curve $y(t)$ is in the band: $\mathrm{MBD}_n^{(j)}(y) = \binom{n}{j}^{-1} \sum_{1 \le i_1 < i_2 < \cdots < i_j \le n} \lambda_r\{A(y; y_{i_1}, \ldots, y_{i_j})\}$, where $A_j(y) \equiv A(y; y_{i_1}, \ldots, y_{i_j}) \equiv \{t \in \mathcal{I} : \min_{r=i_1,\ldots,i_j} y_r(t) \le y(t) \le \max_{r=i_1,\ldots,i_j} y_r(t)\}$ and $\lambda_r(y) = \lambda(A_j(y))/\lambda(\mathcal{I})$, if $\lambda$ is the Lebesgue measure on $\mathcal{I}$. If $y(t)$ is always inside the band, the modified band depth degenerates to the band depth in (2.1).

Because the modified band depth takes the proportion of times that a curve is in the band into account, it avoids having too many depth ties and is more convenient to obtain the most representative curves in terms of magnitude. The band depth is more dependent on the shape of curves often yielding ties, thus it can be used to obtain the most representative curves in terms of shape. Consequently, there are two types of outliers: magnitude outliers and shape outliers. In general, magnitude outliers are distant from the mean and shape outliers have a pattern different from the other curves.

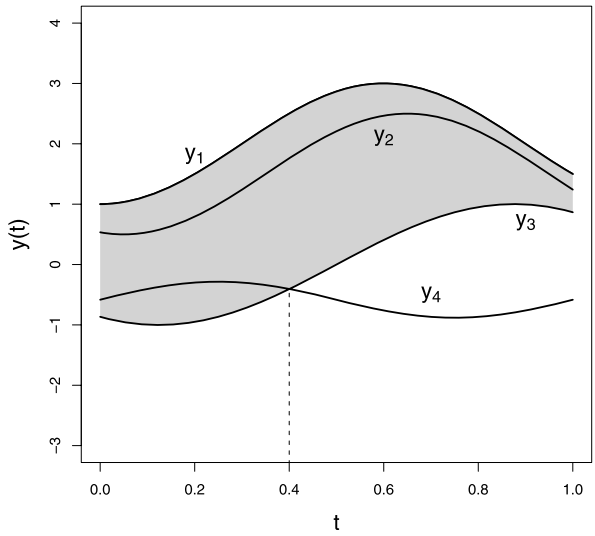


Figure 1. An example of BD and MBD computation: the gray area is the band delimited by $y_1(t)$ and $y_3(t)$. The curve $y_2(t)$ completely belongs to the band, but $y_4(t)$ only partly does.

A sample median function is a curve from the sample with largest depth value, defined by $\arg\max_{y\in\{y_1,\ldots,y_n\}} \mathrm{BD}_{n,J}(y)$. If there are ties, the median will be the average of the curves maximizing depth.

Although the number of curves determining a band, $j$, could be any integer between 2 and $J$, the order of curves induced by band depth is very stable in $J$. To avoid computational issues, we use $J = 2$, and for simplicity, we write $\mathrm{BD}_n^{(2)}$ as BD and $\mathrm{MBD}_n^{(2)}$ as MBD in the sequel.

Figure 1 provides a simple example with $n = 4$ curves on how to compute BD and MBD in practice. When $J = 2$, there are six possible bands delimited by two curves. For instance, the gray area in Figure 1 is the band delimited by $y_1(t)$ and $y_3(t)$. We can see that the curve $y_2(t)$ completely belongs to the band, but $y_4(t)$ only partly does. We define that a curve is contained in a band even if this curve is on the border of the band. Then $\mathrm{BD}(y_2) = 5/6 = 0.83$ since only the band delimited by $y_3(t)$ and $y_4(t)$ does not completely contain the curve $y_2(t)$ and $\mathrm{BD}(y_4) = 3/6 = 0.5$ as it is only completely contained in the bands delimited by itself and another curve. Similarly, we could compute $\mathrm{BD}(y_1) = 0.5$ and $\mathrm{BD}(y_3) = 0.5$. To compute MBD, note that the curve $y_2(t)$ is always contained in the five bands, hence $\mathrm{MBD}(y_2) = 0.83$, the same value as BD. In contrast, the curve $y_4(t)$ only belongs to the band in gray 40% of the time, thus $\mathrm{MBD}(y_4) = (3 + 0.4 + 0.4)/6 = 0.63$ by definition. For the other two curves, $\mathrm{MBD}(y_1) = 0.5$ and $\mathrm{MBD}(y_3) = 0.7$.

## 3. CONSTRUCTION OF FUNCTIONAL BOXPLOTS

In the classical boxplot, the box itself represents the middle 50% of the data. An interesting idea that can be extended to functional data is the concept of central region introduced by Liu, Parelius, and Singh (1999). The band delimited by the $\alpha$ proportion ($0 < \alpha < 1$) of

deepest curves from the sample is used to estimate the $\alpha$ central region. In particular, the sample 50% central region is

$$C_{0.5} = \Big\{(t, y(t)) : \min_{r=1,\ldots,\lceil n/2\rceil} y_{[r]}(t) \le y(t) \le \max_{r=1,\ldots,\lceil n/2\rceil} y_{[r]}(t)\Big\},$$

where $\lceil n/2\rceil$ is the smallest integer not less than $n/2$. The border of the 50% central region is defined as the envelope representing the box in a classical boxplot. Thus, this 50% central region is the analog to the "inter-quartile range" (IQR) and gives a useful indication of the spread of the central 50% of the curves. This is a robust range for interpretation because the 50% central region is not affected by outliers or extreme values, and gives a less biased visualization of the curves' spread. There is also a curve in the box that indicates the median $y_{[1]}(t)$, or the most central curve which has largest band depth value. The median curve is also a robust statistic to measure centrality.

The "whiskers" of the boxplot are the vertical lines of the plot extending from the box and indicating the maximum envelope of the dataset except the outliers. Thus, we need to identify the outliers first. Again, we extend the 1.5 times IQR empirical outlier criterion to the functional boxplot. The fences are obtained by inflating the envelope of the 50% central region by 1.5 times the range of the 50% central region. Any curves outside the fences are flagged as potential outliers. It is worth noting that when each curve is simply a point, the functional boxplot degenerates to a classical boxplot. We suggest the constant factor 1.5 as in a classical boxplot, but we leave to the user the possibility of modifying it.

Now that the pieces of the functional boxplot have been identified, we illustrate its construction on a dataset used by Hyndman and Shang (2010) to demonstrate their functional bagplot shown in Figure 3(c). The data consist of monthly sea surface temperatures (SST) measured in degrees Celsius over the east-central tropical Pacific Ocean and are shown in Figure 2. In this case, each curve represents one year of observed SST in degrees Celsius from January 1951 to December 2007. In our functional boxplot (Figure 3(a)), only the median curve and the flagged outliers are real observations. The border of the box in the middle denotes the envelope of the 50% central region and the minimum and maximum provide the range of non-outlying envelope. To show this difference, we use blue curves to denote envelopes, a black curve to represent the median curve, and red dashed curves to indicate outlier candidates. Thus, instead of having five summary statistics as in a classical boxplot, the functional boxplot has the envelope of the central 50% region, the median curve, and the maximum non-outlying envelope as descriptive statistics. As can be seen from Figure 3, (a) and (c), the two methods display the same median curve in this example, but slightly different outlier detection results. Our functional boxplot detects two outliers by using MBD: the years 1983 and 1997. In addition, the year 1982 from September to December and the year 1998 from January to June are viewed as being part of the maximum envelope. The information discovered by the functional boxplot that September 1982 to December 1983 and January 1997 to June 1998 are abnormal is in close agreement with the recent major El Niño events reported by Dioses, Dávalos, and Zuzunaga (2002). Similarly, the functional bagplot of Hyndman and Shang (2010) detects the years 1982–1983 and 1997–1998 as outliers. For functional data, such as these sea surface temperatures, there will be necessarily dependence in time. This is why the outliers come

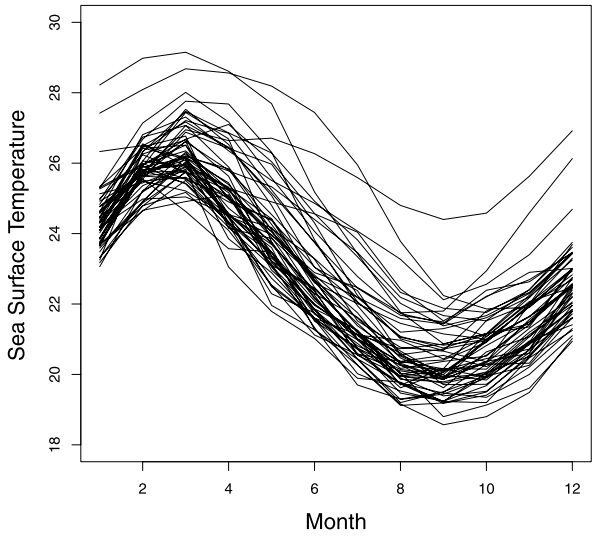


Figure 2. Data of monthly sea surface temperatures measured in degrees Celsius over the east-central tropical Pacific Ocean from 1951 to 2007.

in adjacent years. Considering that the dependence in time may affect outlier detection performance, we allow the constant factor 1.5 to be adjustable in practice.

By introducing the concept of central regions, the functional boxplot can be generalized to an enhanced functional boxplot shown in Figure 3(b). Besides the 50% central region, the 25% and 75% central regions are provided as well. We have implemented a function `fbplot` in `R` (R Development Core Team 2010) to produce functional boxplots and enhanced functional boxplots. It is available as supplemental material on the *JCGS* website.

One may think of using the most intuitive approach, the pointwise boxplots shown in Figure 3(d), which do not treat each curve as one observation. Obviously, such an approach has lost the information of the curves' shapes. In general, the central regions provided by pointwise boxplots are narrower than those given by the functional boxplot, thus many more points would be detected as outliers. By comparing these two types of boxplots, we see that the functional median could be equivalent to the medians in pointwise boxplots only if all the points on the functional median curve are the pointwise 50% quantiles simultaneously. This is rarely true for functional data, especially when curves are very irregular. Specifically, in the above sea surface temperatures example, outliers are detected for each month without taking the annual trend into account. One may connect those monthly outliers from the same year, but it is very difficult to visualize the whole outlying yearly curve and there are cases where only one or two monthly observations within one year are relatively extreme. Furthermore, using the connected pointwise medians (the middle black line in Figure 3(d)) as the most representative curve is not very sensible since it smooths out too many monthly features of a typical yearly temperature curve and is no longer a true curve of the sample.

It is important to note that the box, the whiskers, and the median can reveal useful information about a functional dataset by looking at their position, size, length, and even

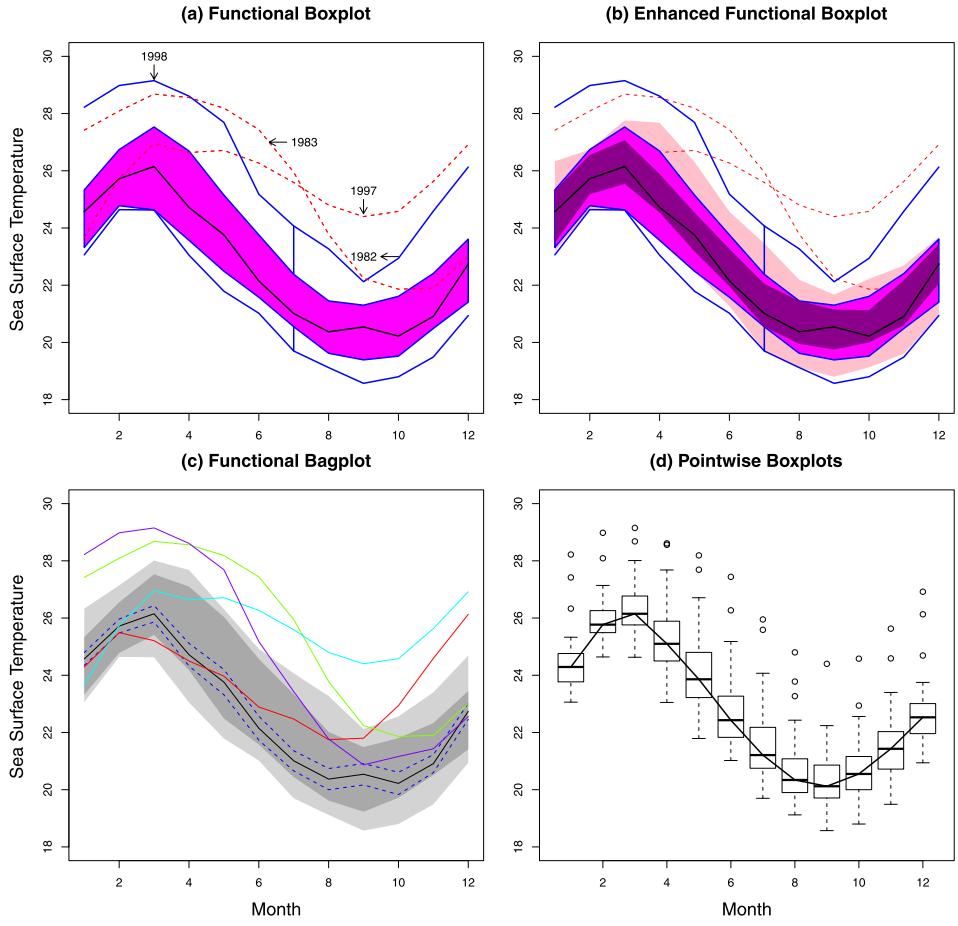


Figure 3. (a) The functional boxplot of SST with blue curves denoting envelopes, and a black curve representing the median curve. The red dashed curves are the outlier candidates detected by the 1.5 times the 50% central region rule. (b) The enhanced functional boxplot of SST with dark magenta denoting the 25% central region, magenta representing the 50% central region, and pink indicating the 75% central region. (c) The functional bagplot of SST. (d) The pointwise boxplots of SST with medians connected by a black line.

the shape of the box or the median curve. Moreover, the spacings between the different parts of the box help indicate the degree of skewness in the data and identify outliers.

## 4. SIMULATION STUDIES

Hyndman and Shang (2010) proposed the functional bagplot and the functional highest density region (HDR) boxplot, which both can detect outliers. The former obtains the outer region (the "fence") by inflating the inner region (the "bag") by a constant factor 2.58 and the latter needs to prespecify the coverage probability of the outlying region. We will focus on comparing our functional boxplot with their functional bagplot since the empirical outlier rule we have proposed obtains the outer region (the "fence") by inflating the inner region (the "envelope") by 1.5 times the range of the 50% central region. We prefer not to

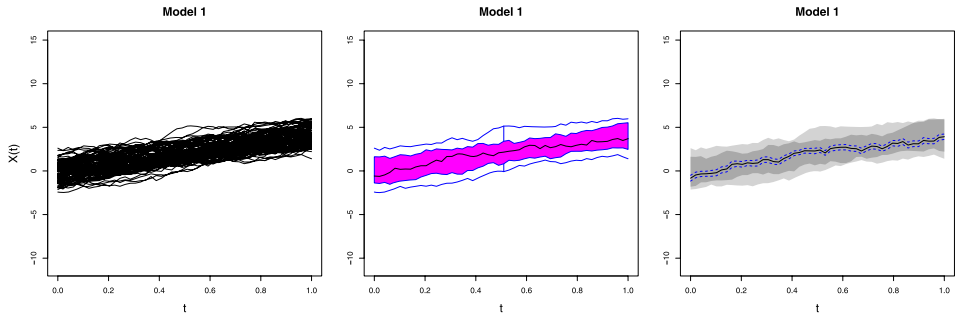


Figure 4. Left panel: curves generated from model 1. Middle panel: the corresponding functional boxplot. Right panel: the corresponding functional bagplot.

have to prespecify the coverage probability of the outlying region in case there is no outlier or the fraction of outliers is unknown.

To further compare our functional boxplot with the principal component (PC) based functional bagplot and assess their performance for outlier detection, we have generated curves from different models introducing either magnitude outliers or shape outliers. The model structures are similar to those of López-Pintado and Romo (2009), but with different parameter values. Some of these models were already considered by Fraiman and Muniz (2001).

Model 1 is a basic one without contamination shown in the left panel of Figure 4. Model 2, model 3, and model 4 have magnitude outliers while model 5 has shape contamination as shown in the left panels of Figure 5. Model details are described as follows:

1. Model 1 is $X_i(t) = g(t) + e_i(t), 1 \le i \le n$, with mean $g(t) = 4t$, $t \in [0, 1]$, and where $e_i(t)$ is a stochastic Gaussian process with zero mean and covariance function $\gamma(s, t) = \exp\{-|t - s|\}$.

2. Model 2 includes a symmetric contamination: $Y_i(t) = X_i(t) + c_i\sigma_i K$, where $c_i$ is 1 with probability $q$ and 0 with probability $1 - q$, $K$ is a contamination size constant, and $\sigma_i$ is a sequence of random variables independent of $c_i$ taking values 1 and $-1$ with probability $1/2$.

3. Model 3 is partially contaminated: $Y_i(t) = X_i(t) + c_i\sigma_i K$, if $t \ge T_i$, and $Y_i(t) = X_i(t)$, if $t < T_i$, where $T_i$ is a random number generated from a uniform distribution on $[0, 1]$.

4. Model 4 is contaminated by peaks: $Y_i(t) = X_i(t) + c_i\sigma_i K$, if $T_i \le t \le T_i + \ell$, and $Y_i(t) = X_i(t)$ otherwise, where $T_i$ is a random number from a uniform distribution in $[0, 1 - \ell]$.

5. Model 5 considers shape contamination with different parameters in the covariance function $\gamma(s, t) = k \exp\{-c|t - s|^{\mu}\}$. The basic model 1, $X_i(t) = g(t) + e_{1i}(t)$, has parameter values $k = 1$, $c = 1$, and $\mu = 1$ for the covariance function of $e_{1i}(t)$. To generate irregular curves, let $Y_i(t) = g(t) + e_{2i}(t)$, where $e_{2i}(t)$ is a Gaussian process with zero mean and covariance function parameters $k = 8$, $c = 1$, and $\mu = 0.2$. The

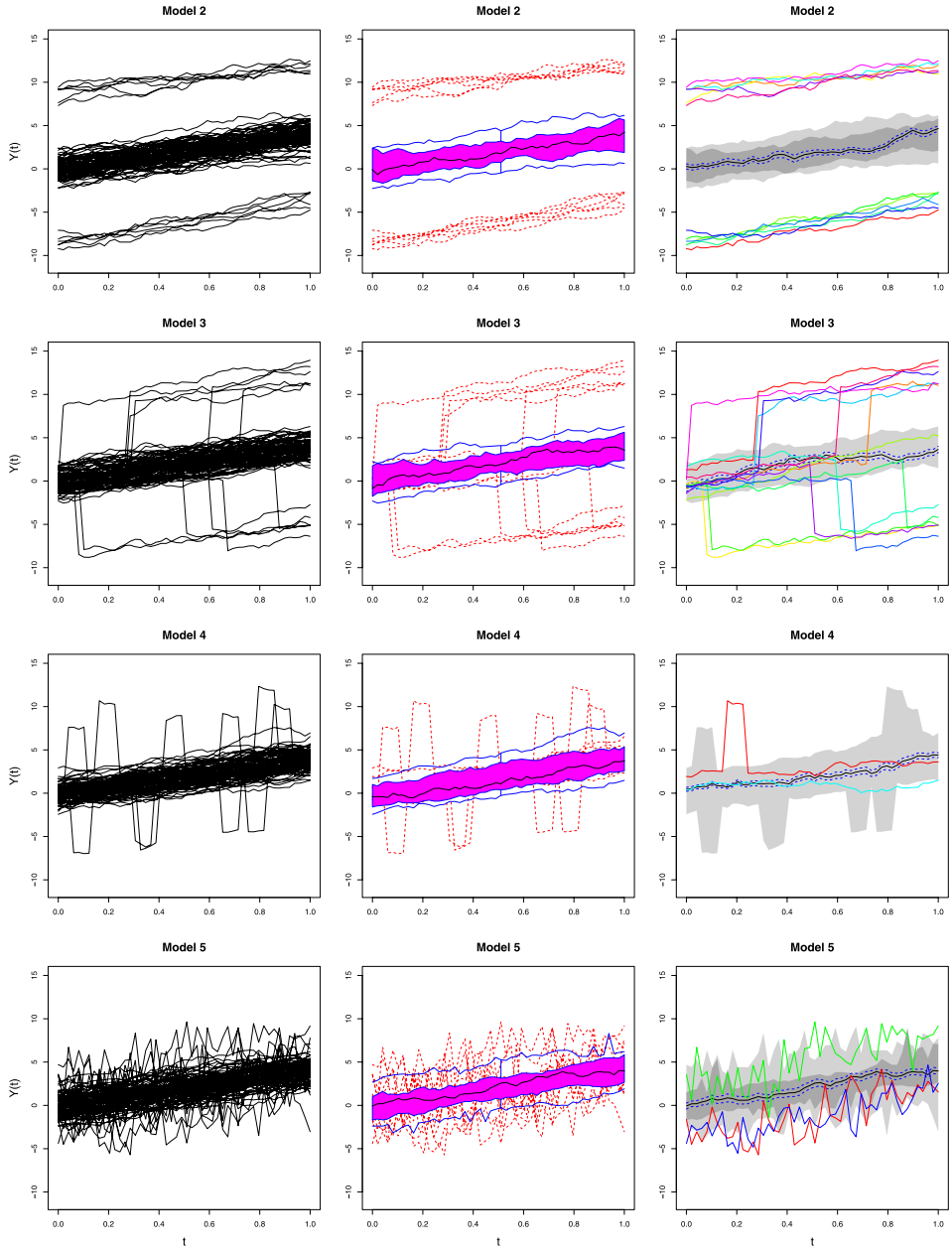


Figure 5. Left panels: curves generated from each contaminated model. Middle panels: the corresponding functional boxplots. Right panels: the corresponding functional bagplots.

contaminated model is given by $Z_i(t) = (1 - c_i)X_i(t) + c_i Y_i(t)$, $1 \le i \le n$, where $c_i$ is 1 with probability $q$ and 0 with probability $1 - q$.

In the simulation studies, we generate $n = 100$ curves with parameters $q = 0.1$, $K = 8$, $\ell = 3/49$, and compute depth values by MBD, the more flexible version of band depth.

Table 1. The percentage $\hat{p}_0$, the mean, and standard deviation of the percentage $\hat{p}_f$ for the functional boxplot and functional bagplot with 1000 replications, 100 curves for model 1.

| Method | $\hat{p}_0$ | Mean($\hat{p}_f$) | SD($\hat{p}_f$) |
|---|---|---|---|
| Functional boxplot | 93.2 | 0.07 | 0.27 |
| Functional bagplot | 24.4 | 2.42 | 6.24 |

Figures 4 and 5 show the difference of outlier detection between our band depth based functional boxplots and the functional bagplots of Hyndman and Shang (2010) based on the first two PC scores. For this particular generated dataset, both methods work equally well on the first three models and the first two PCs of the robust covariance matrices explain 87.0%, 85.0%, and 89.3% of the total variation, respectively. However, the PC based functional bagplot only detects one outlier in model 4, and in model 5 it misses most of the outliers and falsely detects one non-outlying curve. For these two models, the first two PCs explain only 78.3% and 77.5% of the total variation, respectively, which are smaller than those in models 1 to 3. Thus, using only the first two PCs is sometimes a potential drawback of the functional bagplot.

To assess the variability of the outlier detection methods, we are interested in the distribution of two quantities: $p_c$, the percentage of correctly detected outliers (number of correctly detected outliers divided by the total number of outlying curves), and $p_f$, the percentage of falsely detected outliers (number of falsely detected outliers divided by the total number of non-outlying curves).

For model 1, the basic model without outliers, we estimate the percentage, $p_0$, that each of the two methods detects no outliers, and obtain the distribution of the percentage $\hat{p}_f$ with 1000 replications and 100 curves. The percentage, the mean, and standard deviation of $\hat{p}_f$ are shown in Table 1. For models 2 to 5, we obtain the distribution of the two percentages $\hat{p}_c$ and $\hat{p}_f$ with 1000 replications and 100 curves. The means and standard deviations are shown in Table 2. A good performance is defined as high correct detection percentages $p_0$ and $p_c$, but a low false detection percentage $p_f$. As can be seen, overall the functional boxplot method works better than the functional bagplot except for model 3, where, however, the two methods are not significantly different considering the variation. Focusing on the models 1, 4, and 5, the better performance of the functional boxplot method is obvious

Table 2. The mean and standard deviation (in parentheses) of the percentages $\hat{p}_c$ and $\hat{p}_f$ for the functional boxplots and functional bagplots with 1000 replications, 100 curves for models 2 to 5.

| | Model 2 | Model 3 | Model 4 | Model 5 |
|---|---|---|---|---|
| $\hat{p}_c$ | | | | |
| Functional boxplot | 99.1 (3.1) | 83.7 (13.9) | 55.0 (18.4) | 78.6 (15.3) |
| Functional bagplot | 99.5 (5.5) | 88.4 (10.8) | 18.6 (15.7) | 32.7 (17.0) |
| $\hat{p}_f$ | | | | |
| Functional boxplot | 0.03 (0.19) | 0.03 (0.20) | 0.05 (0.27) | 0.03 (0.18) |
| Functional bagplot | 1.81 (5.80) | 1.51 (4.59) | 1.82 (5.51) | 1.66 (5.13) |

and significant. The simulation results show that the functional bagplot method is more likely to either miss a true outlier or falsely detect a non-outlying curve because it only depends on the first two principal components.

Notice that the peaks only appear during short intervals in model 4. By definition, BD would give small depth values for this type of outlying curves but MBD may not. Thus, an alternative would be to compute depth values by BD and to break the possible ties by their MBD values. In this way, simulation results show that the mean of $\hat{p}_c$ could be increased to 95%.

As another simulation study with harmonic signals, we simulated $n = 100$ curves of the form $Y_i(x) = (1 - c_i)\{a_{1i}\sin(t) + a_{1i}\cos(t)\} + c_i\{b_{1i}\sin(t) + b_{2i}\cos(t)\}$, where $0 < t < 2\pi$, $c_i$ is 1 with probability 0.1 and 0 with probability 0.9. The coefficients $a_{1i}$ and $a_{2i}$ follow independent uniform distributions on $[0, 0.05]$, and $b_{1i}$ and $b_{2i}$ also follow independent uniform distributions but on $[0.1, 0.15]$. This model (model 6) is similar to the third example studied by Hyndman and Shang (2010), but we introduce outliers randomly with probability 0.1.

For one particular generated dataset, the original curves and the corresponding functional boxplot and functional bagplot are shown in the top panels of Figure 6. Since the functional highest density region (HDR) boxplot needs to prespecify $\alpha$, the coverage probability of the outlying region, the corresponding HDR boxplots for $\alpha = 0.05, 0.1, 0.2$ are shown in the bottom panels of Figure 6. For this dataset, the functional boxplot correctly detects all the outliers, but the functional bagplot fails to detect any. The three HDR box-

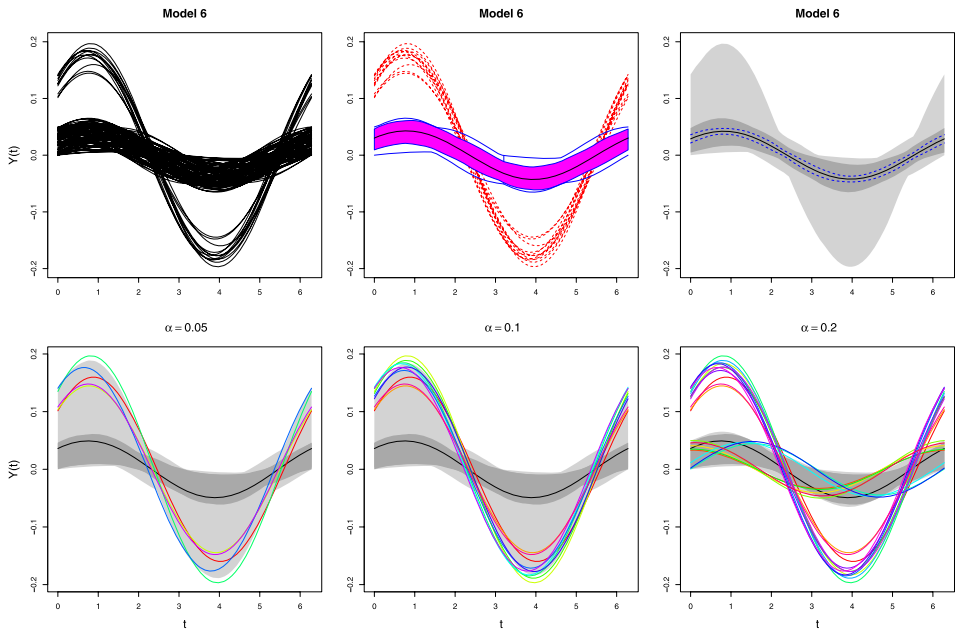


Figure 6. Top panels: the original curves generated from model 6, the corresponding functional boxplot, and the functional bagplot. Bottom panels: the corresponding functional HDR boxplots for $\alpha = 0.05, 0.1, 0.2$, respectively.

Table 3. The mean and standard deviation (in parentheses) of the percentages $\hat{p}_c$ and $\hat{p}_f$ for the functional boxplot, the functional bagplot, and the HDR boxplots with 1000 replications, 100 curves for model 6.

| | | | Functional HDR boxplots | | |
|---|---|---|---|---|---|
| Method | Functional boxplot | Functional bagplot | $\alpha = 0.05$ | $\alpha = 0.1$ | $\alpha = 0.2$ |
| $\hat{p}_c$ | 100 (0.2) | 72.8 (42.4) | 54.7 (17.7) | 90.7 (12.9) | 100 (0.6) |
| $\hat{p}_f$ | 0 (0) | 0.48 (4.50) | 0.07 (0.32) | 1.41 (1.75) | 11.1 (3.0) |

plots clearly show that the outlier detection performance highly depends on the prespecified $\alpha$. When $\alpha$ increases, more outliers are detected but non-outlying curves are also more likely to be flagged as potential outliers at the same time.

Similarly, we obtain the distribution of the two percentages $\hat{p}_c$ and $\hat{p}_f$ for model 6 with 1000 replications and 100 curves. The means and standard deviations are reported in Table 3. The simulation results show that the functional boxplot also works better than the functional bagplot for model 6 and also better than the functional HDR boxplot even with the correctly prespecified outlier probability. For the functional HDR boxplots, the means of $\hat{p}_c$ and $\hat{p}_f$ both increase as $\alpha$ increases. Hence, the outlier detection performance depends on the choice of $\alpha$.

Any outlier detection method should take care of both magnitude and shape outliers. However, to detect shape outliers not far from the median curve with lower density is not an easy task. The functional boxplot would be a good outlier detection method when outliers are either far away from the median in magnitude (models 2 to 4), or outlying in terms of shape but with some outlyingness in magnitude as well (models 5 and 6). However, it may miss outliers which are completely outlying in shape without showing any feature of magnitude outliers. This is where a density approach such as a functional highest density region boxplot can be useful, albeit the percentage of potential outliers must be known and the first two PC scores must explain most of the variation. To illustrate this situation, we let the parameters $a_{1i}$ and $a_{2i}$ in model 6 follow independent uniform distributions on $[0, 0.1]$, and $b_{1i}$ and $b_{2i}$ follow independent uniform distributions on $[0.1, 0.12]$. In this model (model 7), the parameters have the same values as the third example in the article by Hyndman and Shang (2010), which make the outliers not very outlying in magnitude. The only difference is that we still simulate 100 curves and introduce outliers randomly with a probability of 0.1.

For one particular generated dataset, the original curves and the corresponding functional boxplot and functional bagplot are shown in the top panels of Figure 7, and the corresponding HDR boxplots for $\alpha = 0.05, 0.1, 0.2$ are shown in the bottom panels of Figure 7. For this dataset, the functional boxplot and functional bagplot fail to detect any of the outliers because the outlying curves are not sufficiently distant from the median. All three HDR boxplots detect some of the outliers but also flag other curves as potential outliers. As in model 6, when $\alpha$ increases, more and more outliers are detected. With 1000 replications and 100 curves, we obtain the distribution of the two percentages $\hat{p}_c$ and $\hat{p}_f$ for model 7. The means and standard deviations are reported in Table 4. It is shown that the functional boxplot fails to detect the outliers that are not far from the median, and the

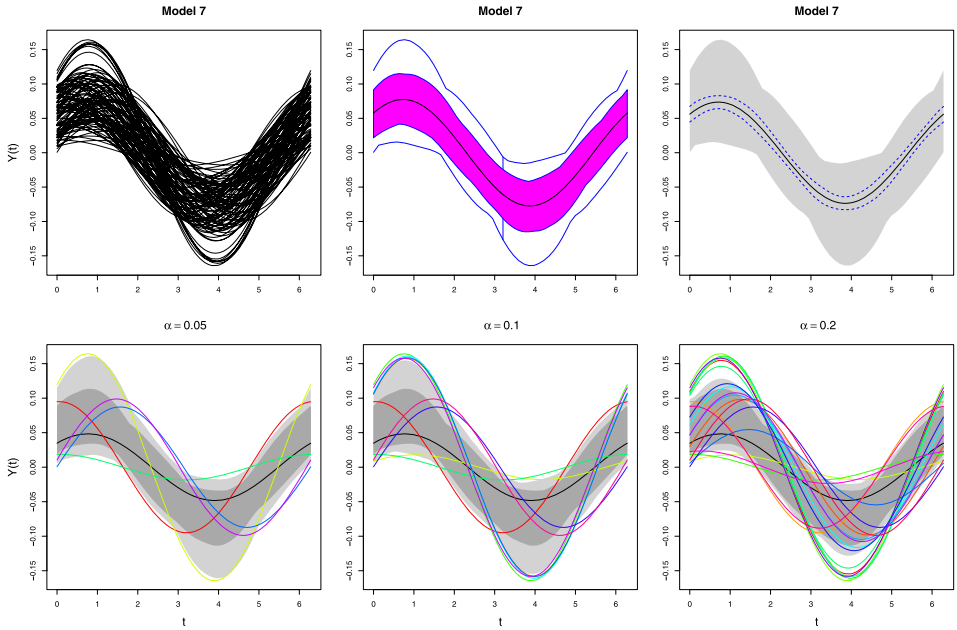


Figure 7. Top panels: the original curves generated from model 7, the corresponding functional boxplot, and the functional bagplot. Bottom panels: the corresponding functional HDR boxplots for $\alpha = 0.05, 0.1, 0.2$, respectively.

functional bagplot also fails most of the time. In contrast, the HDR boxplots can identify more such outliers but the correct detection rate is not high. For instance, the mean of $\hat{p}_c$ is only 17.5% even with the correctly prespecified $\alpha = 0.1$. A larger $\alpha$ could increase the correct detection rate; however, the false detection rate increases as well.

## 5. APPLICATIONS

### 5.1 CHILDREN GROWTH DATA

A strong point of the functional boxplot is its ability to display differences between populations without making any assumptions on the underlying statistical distribution. We start by applying the functional boxplot to the children growth data of Ramsay and Silverman (2005). The heights of 54 girls and 39 boys were measured at 31 unequally spaced

Table 4. The mean and standard deviation (in parentheses) of the percentages $\hat{p}_c$ and $\hat{p}_f$ for the functional boxplot, the functional bagplot, and the HDR boxplots with 1000 replications, 100 curves for model 7.

| | | | Functional HDR boxplots | | |
|---|---|---|---|---|---|
| Method | Functional boxplot | Functional bagplot | $\alpha = 0.05$ | $\alpha = 0.1$ | $\alpha = 0.2$ |
| $\hat{p}_c$ | 0 (0) | 1.73 (12.31) | 8.25 (19.39) | 17.5 (29.9) | 33.0 (38.5) |
| $\hat{p}_f$ | 0 (0) | 0.66 (5.26) | 5.07 (1.15) | 9.95 (2.08) | 19.7 (3.3) |

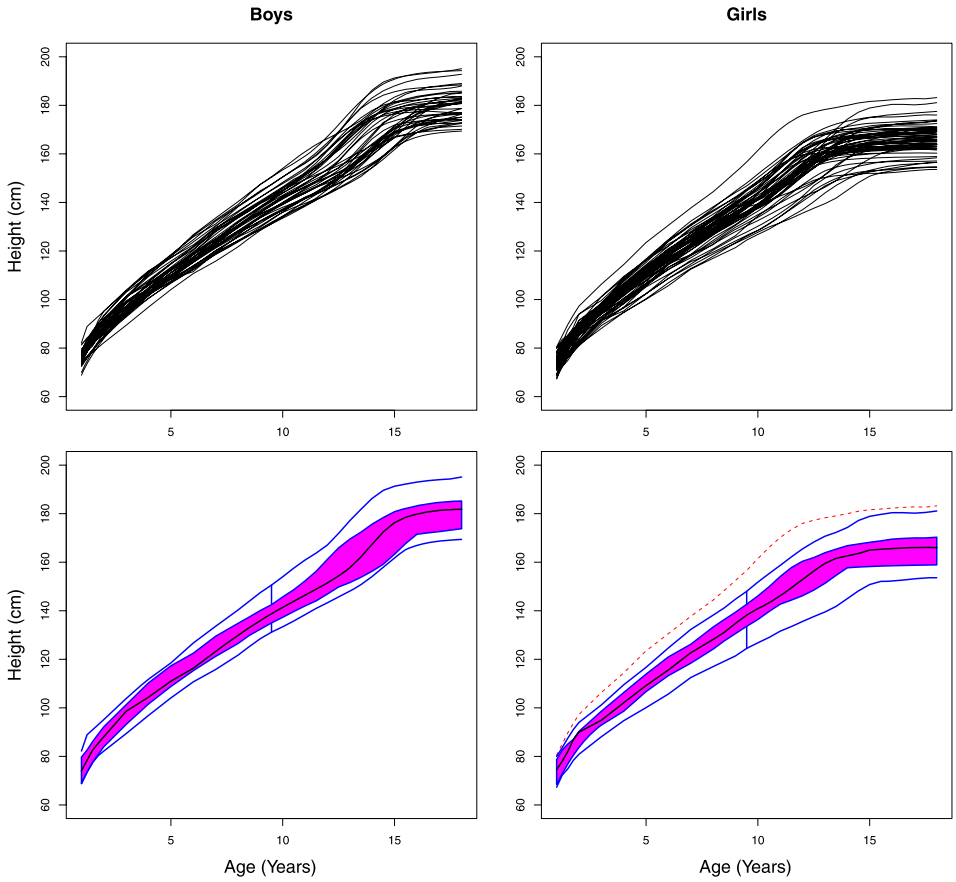


Figure 8. Top panels: the heights of 39 boys and 54 girls at 31 unequally spaced ages. Bottom panels: the corresponding functional boxplots of the children growth data using MBD.

ages from 1 year to 18 years. Within each population, the growth curves are monotonic and similar to a shifted version of each other. Thus we use the MBD because it is more suitable for magnitude outliers as we have discussed in Section 2.

Comparing the original curves to the functional boxplots in Figure 8, we see that the latter are very informative to compare the boys and girls data. The four blue curves and the black curve are the analog to the five summary statistics in a classical boxplot as we explained in the previous section. The median curves can be interpreted as the most representative observed patterns of children growth with age. In the functional boxplot for girls, we notice that there is one detected outlier candidate (red dashed curve), and girls reach lower height values at the end of the growth curves. Also, the shape of the boxes and the median curves depict that boys grew faster than girls between age 13 and 15 years. This information is difficult to obtain by simply looking at the original curves. In addition, one girl is detected as an outlier candidate whose growth curve is always higher in magnitude than the rest. In terms of the shape, this girl grew a little faster at her early age and stopped growing earlier but then still ended up taller than the others.

### 5.2 SPATIO-TEMPORAL PRECIPITATION DATA

Another feature of the functional boxplot is its ability to summarize information from complex data, such as space-time datasets. To illustrate this aspect, we use the observed annual total precipitation data for the coterminous United States from 1895 to 1997, provided by the Institute for Mathematics Applied to Geosciences (*http://www.image.ucar.edu/Data/US.monthly.met/*). There are 11,918 stations reporting precipitation at some time in this period. The observations are functional data since we have one time series with $p = 103$ yearly precipitation observations, or one curve, at each spatial location. Before we apply the functional boxplot to this complex dataset, we first need to perform smoothing to estimate each mean precipitation curve because the records of precipitation at each weather station are so variable. The original data were smoothed by a spline smoothing approach in a nonparametric regression model $y_j = f(x_j) + \varepsilon_j$, where $\varepsilon_j$ iid $\sim N(0, \sigma^2)$, $j = 1, \ldots, p$. Spline smoothing uses all unique data points $x_1, \ldots, x_p$ as knots in the formulation of the cubic spline. Then the cubic spline estimator is obtained by minimizing $\sum_{j=1}^{p} \{y_j - f(x_j)\}^2 + \lambda_i \int \ddot{f}(x)^2 \, dx$, where $\ddot{f}(x)$ is the second derivative of $f(x)$ and $\lambda_i$ is the smoothing parameter of the $i$th curve. The smoothing parameters were estimated from the data by generalized cross-validation.

Using functional boxplots to summarize and compare the annual precipitation for different climatic regions is an interesting application. Nine climatic regions for precipitation in the United States are defined by the National Climatic Data Center (NCDC) and shown in Figure 9. The number of stations is large for each region: the minimum number is 823 for the East North Central region and the maximum number is 2084 for the South region. Blue dots denote stations with normal precipitation and red plus signs present potential outlying stations with respect to their respective climatic region detected by enhanced functional boxplots.

The nine enhanced functional boxplots based on MBD in Figure 10 reveal information about the different annual precipitation characteristics for different climatic regions. For

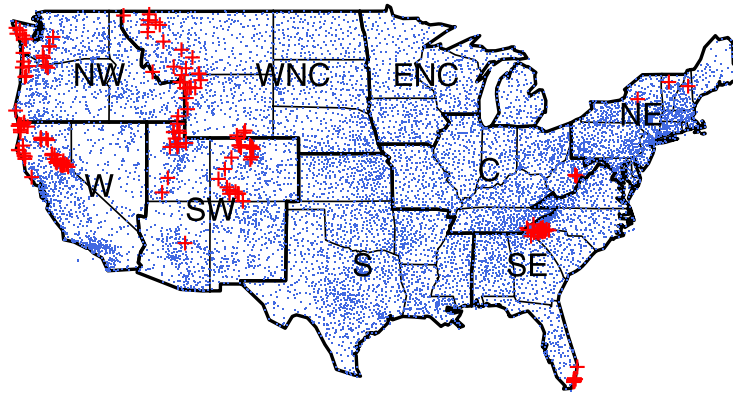


Figure 9. U.S. climatic regions for precipitation from NCDC with abbreviations for North East, East North Central, Central, South East, West North Central, South, South West, North West, and West. Blue dots denote stations with normal precipitation and red plus signs present potential outlying stations with respect to their respective climatic region detected by enhanced functional boxplots.

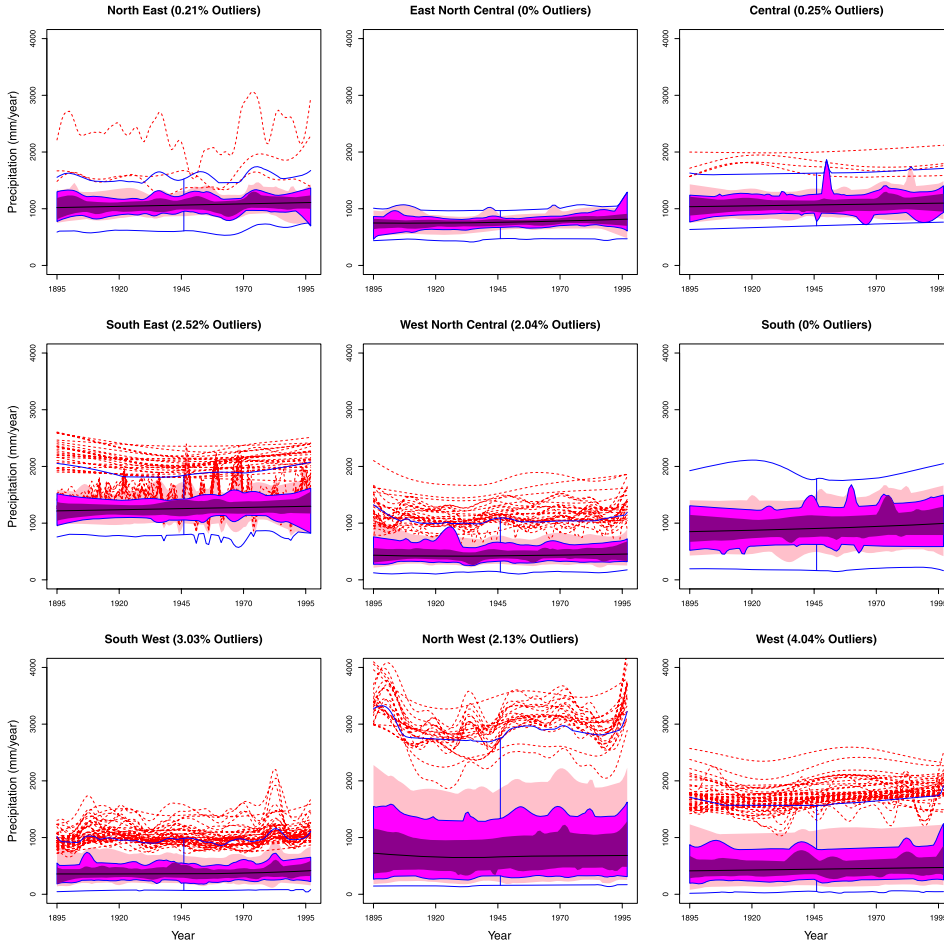


Figure 10. Enhanced functional boxplots of observed yearly precipitation over the nine climatic regions for the coterminous United States from 1895 to 1997 using MBD. Dark magenta, magenta, and pink denote the 25%, 50%, and 75% central regions, respectively, and the outlier rule is 1.5 times the 50% central region. The percentage of outliers in each climatic region is provided.

each region, the global spatial outliers denoted by red dashed curves correspond to the red plus signs on the U.S. map in Figure 9.

There are mainly four areas of potential outliers within the United States shown in Figure 9. Two of them are located along the Rocky Mountains in the West and the Appalachian Mountains in the East with different patterns from the other locations in the corresponding climatic regions. In addition, certain amounts of potential outliers appear along the west coast with higher precipitation which can be clearly seen in the enhanced functional boxplot of North West in Figure 10. By identifying the locations of the potential outliers in the enhanced functional boxplot of South East, we notice that the annual precipitation at the southmost tip in Florida shows an oscillatory pattern. It varies greatly from year to year when hurricanes and droughts have occurred. In Florida, wet springs and summers make up the wet season, and relatively dry winters and autumns form the dry season. If we go

back to look at the original monthly precipitation, it matches the wet and dry seasons at normal locations. However, the outlying locations usually have drier springs, but wet season from July to November even though it is during the drought. And during wet years, most of the precipitation is contributed by the period from July to November which is the hurricane season in Florida. Therefore, the high points of the oscillation in the enhanced functional boxplot capture the effects of hurricanes. If we use a logarithmic scale, it would yield fewer potential outliers. However, it is common that an observation could be an outlier in one scale but not in another. As the classical boxplot also suffers from the same problem, we prefer not to do any transformation in general.

As we have noticed, for spatio-temporal data, we do not have independent curves like in the children growth data example. These precipitation curves are spatially correlated, but the dependence between the curves should only affect the variance of the band depth estimator, not its unbiasedness. The percentage of potential outliers might be different because of the spatial correlation.

## 6. DISCUSSION

This article presented the functional boxplot as an informative exploratory tool for visualizing functional data, as well as its generalization, the enhanced functional boxplot. These functional boxplots were applied to sea surface temperatures, children growth, and spatio-temporal precipitation datasets. With this new technique, outliers can be detected based on the 1.5 times the 50% central region empirical rule. Our approach is distinct from others in treating each curve as an observation rather than summarizing datasets pointwisely. The descriptive statistics in a functional boxplot are rank-based, hence they may lead to building robust statistical models to capture the features of complex datasets.

For spatio-temporal data, we have viewed the information as a temporal curve at each spatial location. An alternative would be to treat the dataset as a spatial surface at each time. In that case, we could define a volume-based surface band depth for a surface $S$ by counting the proportion of surface bands determined by $J$ different surfaces ($2 \leq J \leq n$) in $\mathbb{R}^3$, containing $S$. This would lead to a three-dimensional surface boxplot with similar characteristics as the functional boxplots defined in this article. An illustrative surface boxplot is shown in Figure 11. Similarly, the fences are obtained by the 1.5 times the 50% central region rule. Any surfaces outside the fences are flagged as outlier candidates. The surface boxplot is a natural extension of the functional boxplot to $\mathbb{R}^3$. However, to obtain a three-dimensional functional bagplot, one would definitely need robust principal component analysis techniques to an array rather than a matrix (Hyndman and Shang 2010).

## SUPPLEMENTARY MATERIALS

**Supplements:** *R-code for functional boxplots*: R-code for the command `fbplot` described in the article (fbplot.R) and help file (fbplot.html), with the code for applications (application.R). *Simulation code*: Simulation code for the seven models described in the article (simulation.R). *Children growth data*: Heights of 39 boys and 54 girls at 31

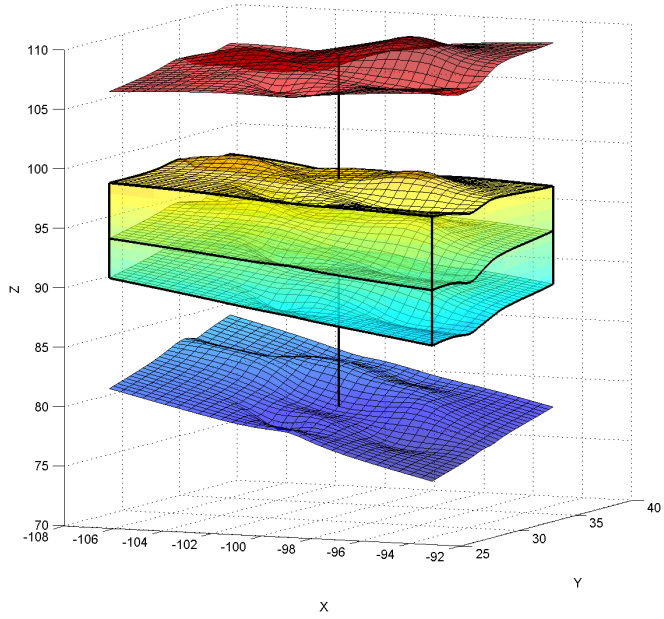


Figure 11. The surface boxplot with the box in the middle representing the 50% central region in $\mathbb{R}^3$, the middle surface inside the box denoting the median surface, and the upper and lower surfaces indicating the maximum non-outlying envelope.

unequally spaced ages described in the article (hgtgirls.dat and hgtboys.dat). *Sea surface temperatures data*: The data of monthly sea surface temperatures (SST) measured in degrees Celsius over the east-central tropical Pacific Ocean described in the article (sst.dat). All files can be found in a single zip file. (fbplot.zip)

## ACKNOWLEDGMENTS

This research was partially supported by NSF grants CMG ATM-0620624, DMS-1007504, and award no. KUS-C1-016-04, made by King Abdullah University of Science and Technology (KAUST). The authors thank the editor, an associate editor, and three anonymous referees for their valuable comments.

*[Received December 2009. Revised September 2010.]*

## REFERENCES

Cuevas, A., Febrero, M., and Fraiman, R. (2006), "On the Use of the Bootstrap for Estimating Functions With Functional Data," *Computational Statistics and Data Analysis*, 51 (2), 1063–1074. [317]

——— (2007), "Robust Estimation and Classification for Functional Data via Projection-Based Depth Notions," *Computational Statistics*, 22, 481–496. [317]

Dioses, T., Dávalos, R., and Zuzunaga, J. (2002), "El Niño 1982–1983 and 1997–1998: Effects on Peruvian Jack Mackerel and Peruvian Chub Mackerel," *Investigaciones Marinas*, 30 (1), 185–187. [320]

Dryden, I. L., and Mardia, K. V. (1998), *Statistical Shape Analysis. Wiley Series in Probability and Statistics: Probability and Statistics*, Chichester: Wiley. [316]

Febrero, M., Galeano, P., and González-Manteiga, W. (2007), "Functional Analysis of NOx Levels: Location and Scale Estimation and Outlier Detection," *Computational Statistics*, 22 (3), 411–427. [317]

——— (2008), "Outlier Detection in Functional Data by Depth Measures, With Application to Identify Abnormal NOx Levels," *Environmetrics*, 19 (4), 331–345. [317]

Ferraty, F., and Vieu, P. (2006), *Nonparametric Functional Data Analysis: Theory and Practice*, New York: Springer. [316]

Fitzenberger, B., Koenker, R., and Machado, J. A. F. (2002), *Economic Applications of Quantile Regression*, New York: Springer Verlag. [316]

Fletcher, P., Lu, C., Pizer, S., and Joshi, S. (2004), "Principal Geodesic Analysis for the Study of Nonlinear Statistics of Shape," *IEEE Transactions on Medical Imaging*, 23 (8), 995–1005. [316]

Fraiman, R., and Meloche, J. (1999), "Multivariate $L$-Estimation," *Test*, 8, 255–317. [317]

Fraiman, R., and Muniz, C. (2001), "Trimmed Means for Functional Data," *Test*, 10 (2), 419–440. [317,323]

Hyndman, R. J., and Shang, H. L. (2010), "Rainbow Plots, Bagplots, and Boxplots for Functional Data," *Journal of Computational and Graphical Statistics*, 19, 29–45. [317,320,322,325-327,332]

Liu, R. (1990), "On a Notion of Data Depth Based on Random Simplices," *The Annals of Statistics*, 18, 405–414. [317]

Liu, R. Y., Parelius, J. M., and Singh, K. (1999), "Multivariate Analysis by Data Depth: Descriptive Statistics, Graphics and Inference," *The Annals of Statistics*, 27, 783–858. [319]

López-Pintado, S., and Romo, J. (2009), "On the Concept of Depth for Functional Data," *Journal of the American Statistical Association*, 104, 718–734. [317,318,323]

Mahalanobis, P. C. (1936), "On the Generalized Distance in Statistics," *Proceedings of National Academy of Science of India*, 12, 49–55. [317]

Oja, H. (1983), "Descriptive Statistics for Multivariate Distributions," *Statistics & Probability Letters*, 1, 327–332. [317]

R Development Core Team (2010), *R: A Language and Environment for Statistical Computing*, Vienna, Austria: R Foundation for Statistical Computing. ISBN 3-900051-07-0, available at *http://www.R-project.org*. [321]

Ramsay, J. O., and Silverman, B. W. (2005), *Functional Data Analysis* (2nd ed.), New York: Springer Verlag. [316,328]

Rousseeuw, P., Ruts, I., and Tukey, J. W. (1999), "The Bagplot: A Bivariate Boxplot," *The American Statistician*, 53 (4), 382–387. [317]

Singh, K. (1991), "A Notion of Majority Depth," unpublished manuscript, Rutgers University. [317]

Tukey, J. W. (1970), *Exploratory Data Analysis (Limited Preliminary Edition)*, Vol. 1, Reading, MA: Addison-Wesley, Chapter 5. [317]

——— (1975), "Mathematics and the Picturing of Data," in *Proceedings of the International Congress of Mathematicians, August 21–29, 1974*, Vol. 2, ed. R. D. James, Vancouver: Canadian Mathematical Society, pp. 523–531. [317]

——— (1977), *Exploratory Data Analysis*, Reading, MA: Addison-Wesley. [317]

Vardi, Y., and Zhang, C. H. (2000), "The Multivariate $L_1$-Median and Associated Data Depth," *Proceedings of the National Academy of Sciences of the United States of America*, 97, 1423–1426. [317]