Rainbow plots, bagplots and boxplots for functional data

Rob J Hyndman and Han Lin Shang
Department of Econometrics and Business Statistics,
Monash University, Clayton, Australia

June 5, 2009

Abstract

We propose new tools for visualizing large amounts of functional data in the form of smooth curves. The proposed tools include functional versions of the bagplot and boxplot, and make use of the first two robust principal component scores, Tukey's data depth and highest density regions.

By-products of our graphical displays are outlier detection methods for functional data. We compare these new outlier detection methods with existing methods for detecting outliers in functional data, and show that our methods are better able to identify the outliers.

An R-package containing computer code and data sets is available in the online supplements.

Keywords: Highest density regions, Robust principal component analysis, Kernel density estimation, Outlier detection, Tukey's halfspace depth.

1 Introduction

Functional data are becoming increasingly common in a wide range of fields, and there is a need to develop new statistical tools for analyzing such data. In this paper, we are interested in visualizing data comprising smooth curves (e.g., Ramsay & Silverman, 2005; Locantore et al., 1999). Such functional data may be age-specific mortality or fertility rates (Hyndman & Ullah, 2007), term-structured yield curves (Kargin & Onatski, 2008), spectrometry data (Reiss & Ogden, 2007), or one of the many applications described by Ramsay & Silverman (2002). Ramsay & Silverman (2005) and Ferraty & Vieu (2006) provide detailed surveys of the many parametric and nonparametric techniques for analyzing functional data.

Visualization methods help in the discovery of characteristics that might not have been apparent using mathematical models and summary statistics; and yet this area of research has not received much attention in the functional data analysis literature to date. Most of the literature focuses on the modeling, clustering and forecasting of functional data, with visualization playing a minor role, at best. Notable exceptions are the phase-plane plot of Ramsay & Ramsey (2002) and the rug plot of Hyde et al. (2006), which highlight important distributional characteristics from the first and second derivatives of functional data. Another exception is the singular value decomposition plot of Zhang et al. (2007), which displays the changes in latent components as the sample size or dimensionality increases. We aim to contribute to the functional data analytic toolbox by proposing three new graphical methods: the rainbow plot, the functional bagplot and the functional highest density region (HDR) boxplot.

A side benefit of two of these new graphical methods is the identification of outliers, which may not be obvious from a plot of the original data. Outlying curves may either lie outside the range of the vast majority of the data (we call these "magnitude outliers"), or they may be within the range of the rest of the data but have a very different shape from other curves (we call these "shape outliers"), or they may exhibit a combination of these features. Any attempt to identify outlying curves should be able to deal with all types of outliers.

The presence of outliers has a serious effect on the modeling and forecasting of functional data. Statistical analysis which does not involve identifying outliers can often lead to inaccurate conclusions. Despite the obvious importance of this problem, we are aware of only two previous approaches to functional outlier detection. Hyndman & Ullah (2007) used a method based on robust principal component analysis, while Febrero et al. (2007, 2008) considered functional outlier detection using successive likelihood ratio tests and smoothed bootstrapping.

To motivate the discussion, consider Figure 1, which shows annual smoothed age-specific mortality curves for French males between 1899 and 2005. The data were taken from the Human Mortality Database (2008) and smoothed using penalized splines, as described by Hyndman & Ullah (2007). The mortality rates are the ratios of death counts to population exposure in the relevant year for the given age (based on one-year age groups). In this example, $y_i(x)$ denotes the logarithm of the mortality rates in year i for males of age x. Outliers clearly exist in the data due to wars and epidemics, and we seek to identify them.

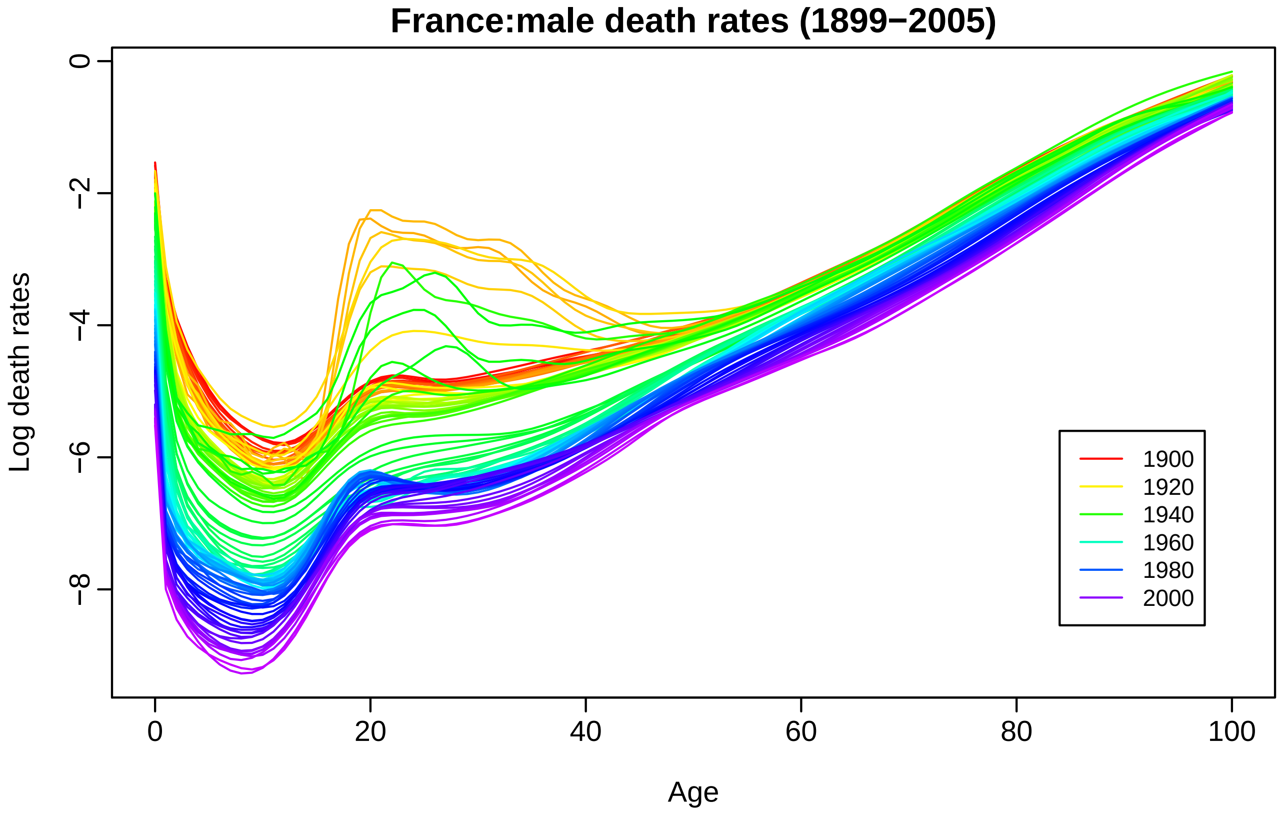


Figure 1: *French male age-specific mortality rates (1899–2005). The oldest curves are shown in red, with the most recent curves in violet. Curves are ordered chronologically according to the colors of the rainbow.*

Figure 1 is an example of a “rainbow plot” where the colors of the curves follow the order of a rainbow, with the oldest data in red and the most recent data in violet. This is one of the plots discussed in Section 3.

The rainbow plot is a simple plot of all the data, with the only added feature being a color palette based on an ordering of the data. Figure 1 shows time-ordering, but other possibilities are based on data depth, data density or other unique ranking procedures. In Section 2, we explore various ordering methods for functional data and show how the rainbow plot can be surprisingly illuminating with a careful choice of ordering.

With a large number of overlapping curves, it is difficult to identify where the “median curve” might lie, or where the bulk of the data fall. It may also be difficult to see outliers if they are obscured by other curves (for example, curves having a different shape from the rest of the data).

With univariate data, we commonly use boxplots to solve these issues. Therefore, we aim to define functional variations on boxplots which will show outlying curves, a “central” curve, and a region containing the “middle” 50% of curves.

Two of the ordering methods that will prove useful are based on a robust principal component algorithm applied to the first two principal component scores calculated from $\{y_i(x)\}$. These bivariate points can then be ordered by applying Tukey’s (1974) halfspace depth, for example. This idea immediately lends itself to a functional bagplot obtained by applying the bivariate bagplot (Rousseeuw et al., 1999) to the first two principal component scores. Recently, this idea has been used for clustering functional data by Sood et al. (2009).

Similarly, a functional HDR boxplot is defined by computing a bivariate kernel density estimate (Scott, 1992) on the first two principal component scores, and then applying the bivariate HDR boxplot of Hyndman (1996). The HDR boxplot has the advantage of being able to display multimodality if it is present in the data. The bagplot and boxplot are introduced in Section 3.

In Figure 1, some of the mortality curves shown in yellow and green indicate sudden increases in mortality rates between the ages of 20 and 40 for a number of years. These are due to the dramatic changes in mortality patterns resulting from the first and second World Wars, as well as the Spanish flu which occurred in 1918 and 1919.

Outliers in the functional data are identified as outliers in the bivariate score space. A related idea has previously been employed by Jones & Rice (1992) for solving the graphical obscurity problem associated with large collections of curves. In Section 4 the outlier detection methods which are made possible by our functional bagplot and functional boxplot are compared with several existing functional outlier detection methods and multivariate outlier detection methods applied to the discretized functions.

Section 5 provides a summary of our main results, and some thoughts on how the plots might be extended. In this paper, we do not address the issue of preprocessing data. For a noise data set or an unequally spaced data set, it is advisable to preprocess data using techniques such as curve registration methods (Ramsay & Silverman, 2005, Chapter 10) or nonparametric smoothing methods (Eubank, 1999), in order to extract the most informative aspects of the data.

2 Ordering functional data

All of our graphical methods involve some kind of ordering of the functional data. Figure 1 showed the data in time-order, but for many data sets we will want an ordering based on the values of the data themselves. In this section, we review some possible ordering methods, and consider how they can be used in conjunction with the rainbow plot.

Each of the ordering methods uses a form of data depth (Tukey, 1974) or data density (Hyndman, 1996), which provides a way to measure the "depth" or "density" of a given observation with respect to the set of observations or their underlying distribution. Often, the contours of a depth function or a density function are used to reveal the shape and structure of multivariate data.

2.1 Functional depth method

Febrero et al. (2007) proposed an outlier detection method that used the notion of functional depth, defined as

$$o_i = \int D(y_i(x))\, dx, \tag{1}$$

where $D(y_i(x))$ is a univariate depth measure for a specific value of x. Using this definition, we define the ordering of the curves by an increasing order of $\{o_i\}$, so the first curve has the lowest functional depth and the last curve has the greatest functional depth.

However, as the functional depth is calculated by integrating the univariate depth, this method may not detect curves that have an unusual shape but lie within the range of the majority of curves (López-Pintado & Romo, 2006). Consequently, this definition of functional depth is not adequate for many functional data sets. We thus propose a measure of functional depth which is based on the first two principal component scores instead.

2.2 Bivariate score depth

Let $\{\phi_k(x)\}$ represent the principal components, and $\{z_{i,k}\}$ the principal component scores from a functional principal component decomposition. Much of the information inherent in the original data $\{y_i(x)\}$ is captured in the first few principal components and scores (Jones & Rice, 1992; Sood et al., 2009). Therefore we will take the first two score vectors $(z_{1,1}, \dots, z_{n,1})$, and $(z_{1,2}, \dots, z_{n,2})$

and consider methods of bivariate depth that could be applied to these vectors. We shall refer to the bivariate point $(z_{i,1}, z_{i,2})$ as $\boldsymbol{z}_i$.

Because principal component decomposition is not resistant to outliers, we apply Croux & Ruiz-Gazen's (2005) robust principal component algorithm, which uses a form of projection pursuit. This algorithm was designed for multivariate rather than functional data, but we can also apply it to discretized curves $\{y_i(x)\}$. The virtue of this approach is that it can still be applied even when the number of variables is significantly greater than the number of observations, which is the case with finely discretized curves.

The bivariate scores can be ordered using Tukey's halfspace location depth (Tukey, 1974), denoted by $d(\boldsymbol{\theta}, \boldsymbol{Z})$ for some point $\boldsymbol{\theta} \in R^2$ relative to the bivariate data cloud $\boldsymbol{Z} = \{\boldsymbol{z}_i; i = 1, \dots, n\}$. Tukey's depth is defined as the smallest number of data points contained in a closed half-plane containing $\boldsymbol{\theta}$ on its boundary. Then the observations can be ordered as the distances $o_i = d(\boldsymbol{z}_i, \boldsymbol{Z})$ in an increasing order. The first curve by this ordering can be considered as a "median" curve, while the last curve can be considered as the outermost curve in a sample of curves.

2.3 Data density

The third way to order the points is by the value of a bivariate kernel density estimate (Scott, 1992) at each observation. Let $o_i = \hat{f}(\boldsymbol{z}_i)$, where $\hat{f}(\boldsymbol{z})$ is a bivariate kernel density estimate calculated from all of the bivariate principal component scores. Then the functional data are ordered by values of $\{o_i\}$ in a decreasing order. So the curve with the highest density is the first observation, and the last curve has the lowest density value. Thus, the first curve may be considered the "modal curve" while the last curve may be considered the most unusual curve.

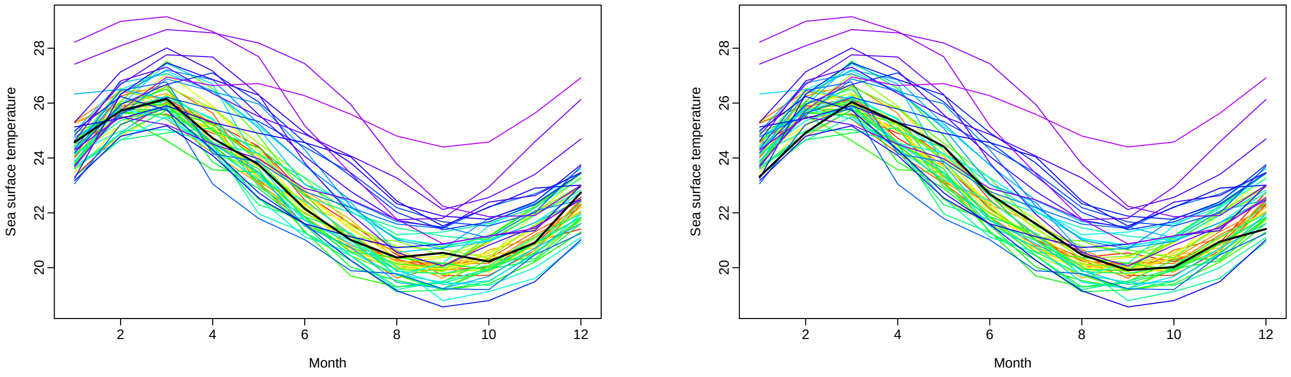
Note that the last curve by this ordering may not take values very different from the others and its bivariate scores may not be on the edge of the scatterplot of $(z_{1,i}, z_{2,i})$. It is possible to have a point which is on the interior of this scatterplot, but which has no other points nearby, and will hence have a low density value.

3 Three functional graphical tools

3.1 Rainbow plots

For data that are not naturally ordered by time, or some other index, the rainbow plot can still be used by constructing an ordering index such as the data depth or data density indexes defined above. Then the colors are chosen in rainbow order according to the ordering of $\{o_i\}$.

To demonstrate this, we consider a time series of average monthly sea surface temperatures from January 1951 to December 2007, available online at `http://www.cpc.noaa.gov/data/indices/sstoi.indices`. These temperatures are measured by moored buoys in the "Niño region", defined by the coordinates $0-10°$ South and $90-80°$ West. There is no time trend in these data, so a rainbow plot with time ordering is not particularly informative. Rainbow plots using depth and density order indexes are shown in Figure 2. The colors reflect the ordering and follow the order of the rainbow, with the curves closest to the center of the data set shown in red, while the most outlying curves are shown in violet. We plot the curves in order of depth or density, so the red curves are mostly obscured, but the violet outlier curves are seen clearly, even if they overlap the majority of the data.



(a) *Rainbow plot with depth ordering.*

(b) *Rainbow plot with density ordering.*

Figure 2: *Rainbow plots using different order indexes. The black lines show the median curve in (a) and the modal curve in (b).*

3.2 Functional bagplot

The functional bagplot is based on the bivariate bagplot of Rousseeuw et al. (1999), applied to the first two principal component scores. It uses Tukey's (1974) halfspace location depths. The depth region D_k is the set of all $\boldsymbol{\theta}$, with $d(\boldsymbol{\theta}, \boldsymbol{z}) \geq k$. Since the depth regions form a series of convex hulls, we have $D_{k_1} \subset D_{k_2}$ for $k_2 > k_1$. The Tukey bivariate depth median is defined as the value of $\boldsymbol{\theta}$ which minimizes $d(\boldsymbol{\theta}, \boldsymbol{Z})$ if there is such a unique $\boldsymbol{\theta}$; otherwise it is defined as the center of gravity of the deepest region.

Like a univariate boxplot, the bivariate bagplot has a central point (the Tukey median), an inner region (the 'bag') and an outer region, beyond which outliers are shown as individual points. The bag is defined as the smallest depth region containing at least 50% of the total number of observations. The outer region (or 'fence') of the bagplot is the convex hull of the points contained within the region obtained by inflating the bag (relative to the Tukey median) by a factor ρ. Rousseeuw et al. (1999) used a value of $\rho = 3$. However, we prefer $\rho = 2.58$, as that will allow the fence to contain 99% of the observations when the projected bivariate scores follow standard normal distributions. A proof of this result is given in the Appendix.

The functional bagplot is a mapping of the bagplot of the first two robust principal component scores to the functional curves. The functional bagplot displays the median curve (the curve with the greatest depth), and the inner and outer regions. The inner region is defined as the region bounded by all curves corresponding to points in the bivariate bag. Thus, 50% of curves are in the inner region. The outer region is similarly defined as the region bounded by all curves corresponding to points within the bivariate fence region.

Two examples are shown in Figure 3, using the French male mortality data and the El Niño indexes. In the left panels, the dark grey region shows the 50% bag and the light grey region exhibits the 99% fence. These convex hulls correspond directly to the regions of similar shading in the functional bagplot on the right. Points outside these regions are identified as outliers. The different colors for these outliers enable the individual functional curves on the right to be matched to the bivariate principal component scores on the left. The red asterisk marks the Tukey median of the bivariate principal component scores, and the solid black curve in each of the panels on the right shows the median curve. The dotted blue lines in the right panels give 95% pointwise confidence intervals for the median curve (similar to the notched boxplot of Tukey (1977)).

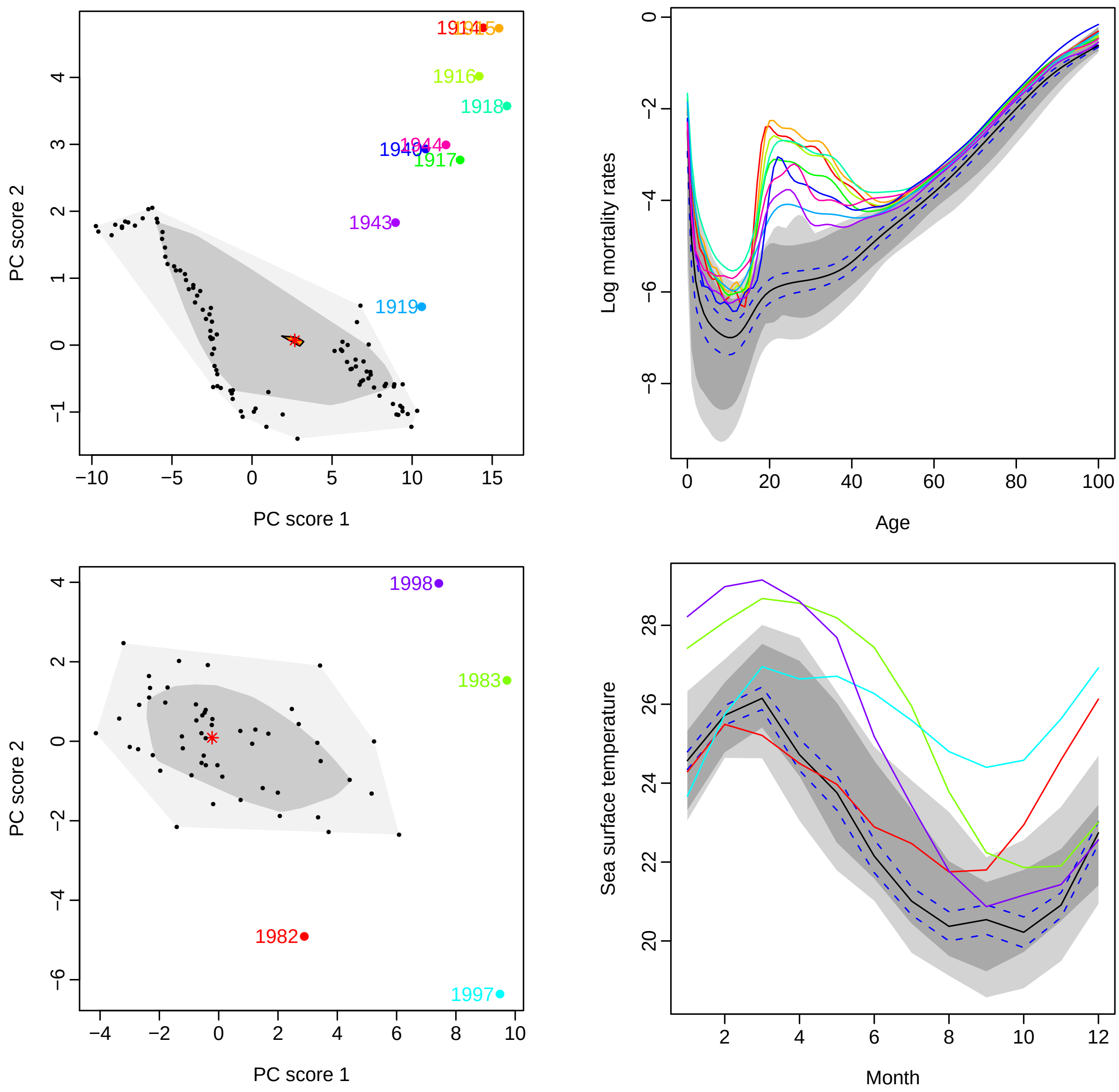


Figure 3: *Bivariate bagplot and functional bagplot for the French male mortality rates (top) and El Niño indexes (bottom). The dark and light grey regions show the bag and fence regions, respectively. The red asterisk is the Tukey depth median. In the right panels, the black line is the median curve, surrounded by 95% pointwise confidence intervals. The curves outside the outer region are shown as outliers of different colors.*

The detected outliers in the French male mortality data are the years 1914–1919, 1940, 1943–1944. They correspond to the first and second World Wars and the Spanish flu pandemic. The detected outliers in the El Niño data are the years 1982–1983 and 1997–1998. The El Niño indexes during 1982–1983 began in June 1982 with a weak heating, then there was an extreme abnormal increase in sea surface temperature between September 1982 and June 1983 (Timmermann et al., 1999; Moran et al., 2006). The El Niño indexes during 1997–1998 were also unusual, especially in March and April. Dioses et al. (2002) reported that the northern central region of Peru was strongly affected as warm waters with low salinity approached the coast, while the southern region of Peru was more influenced by oceanic waters.

The functional bagplot may be a good outlier detection method when outliers are far away from the median. However, when outliers are near the median, this depth-measure outlier detection tool can misidentify outliers, as is shown in Section 4.5 via a simulated data set. In this situation, the functional HDR boxplot is more appropriate.

3.3 Functional HDR boxplot

The functional HDR boxplot is based on the bivarate HDR boxplot (Hyndman, 1996), which is applied to the first two principal component scores. The bivariate HDR boxplot is constructed using a bivariate kernel density estimate $\hat{f}(\boldsymbol{z})$, which is defined as

$$\hat{f}(z) = \frac{1}{n}\sum_{i=1}^{n} K_{h_i}(z - Z_i),$$

where Z_i represents a set of bivariate points, $K_{h_i}(\cdot) = K(\cdot/h_i)/h_i$, K is the kernel function, and h_i is the bandwidth for the ith dimension. The bandwidths were selected using smoothed cross validation (Duong & Hazelton, 2005).

Using the kernel density estimates, a HDR is defined as

$$R_\alpha = \{\boldsymbol{z} : \hat{f}(\boldsymbol{z}) \geq f_\alpha\},$$

where f_α is such that $\int_{R_\alpha} \hat{f}(\boldsymbol{z})d\boldsymbol{z} = 1 - \alpha$; that is, it is the region with probability coverage $1-\alpha$, where all points within the region have a higher density estimate than any of the points outside the region — hence the name "highest density region". For a bivariate density, the HDRs can be considered as contours, with an expanding coverage as α decreases.

The bivariate HDR boxplot displays the mode, defined as $\arg\sup\hat{f}(\boldsymbol{z})$, along with the 50% inner and (usually) 99% highest density regions. All points excluded from the outer HDR are outliers.

The functional HDR boxplot is a mapping of the bivariate HDR boxplot of the first two robust principal component scores to the functional curves. The functional HDR boxplot displays the modal curve (the curve with the highest density), and the inner and outer regions. The inner region is defined as the region bounded by all curves corresponding to points inside the 50% bivariate HDR. Thus, 50% of curves are in the inner region. The outer region is similarly defined as the region bounded by all curves corresponding to the points within the outer bivariate HDR.

Two examples are shown in Figure 4 using the French male mortality data and the El Niño indexes. In the left panel, the dark and light grey regions show the 50% HDR and the outer HDR, respectively. These correspond directly to the regions of similar shading in the functional HDR boxplots on the right. The points outside these outer regions are identified as outliers. The use of different colors for these outliers enables the individual curves on the right to be matched with the bivariate score HDR on the left. The red dot in the left panel marks the mode of the bivariate principal component scores, and corresponds to the solid black curve in the right panel.

As with any outlier detection method, including bagplots and HDR boxplots, the coverage probability of the outer region needs to be pre-specified. With the 99% coverage probability, the outliers detected in the French male mortality data are 1919 and 1943, while the outlier detected in the El Niño data set is 1997. However, if we set the coverage probability of the outer region to be 92% and 93% in the French male mortality data and the El Niño data respectively, the outliers detected in each of the examples would then match the results obtained by the bagplot. This indicates that those outliers have different magnitudes and shapes to the rest of data.

The presence of bimodality is seen in the top row of Figure 4. This indicates that samples may come from two populations. Further investigation shows that the two regions correspond to the years before and after the end of World War II. Immediately after the war, there was a large drop in the mortality rates, which can be seen clearly in Figure 1.

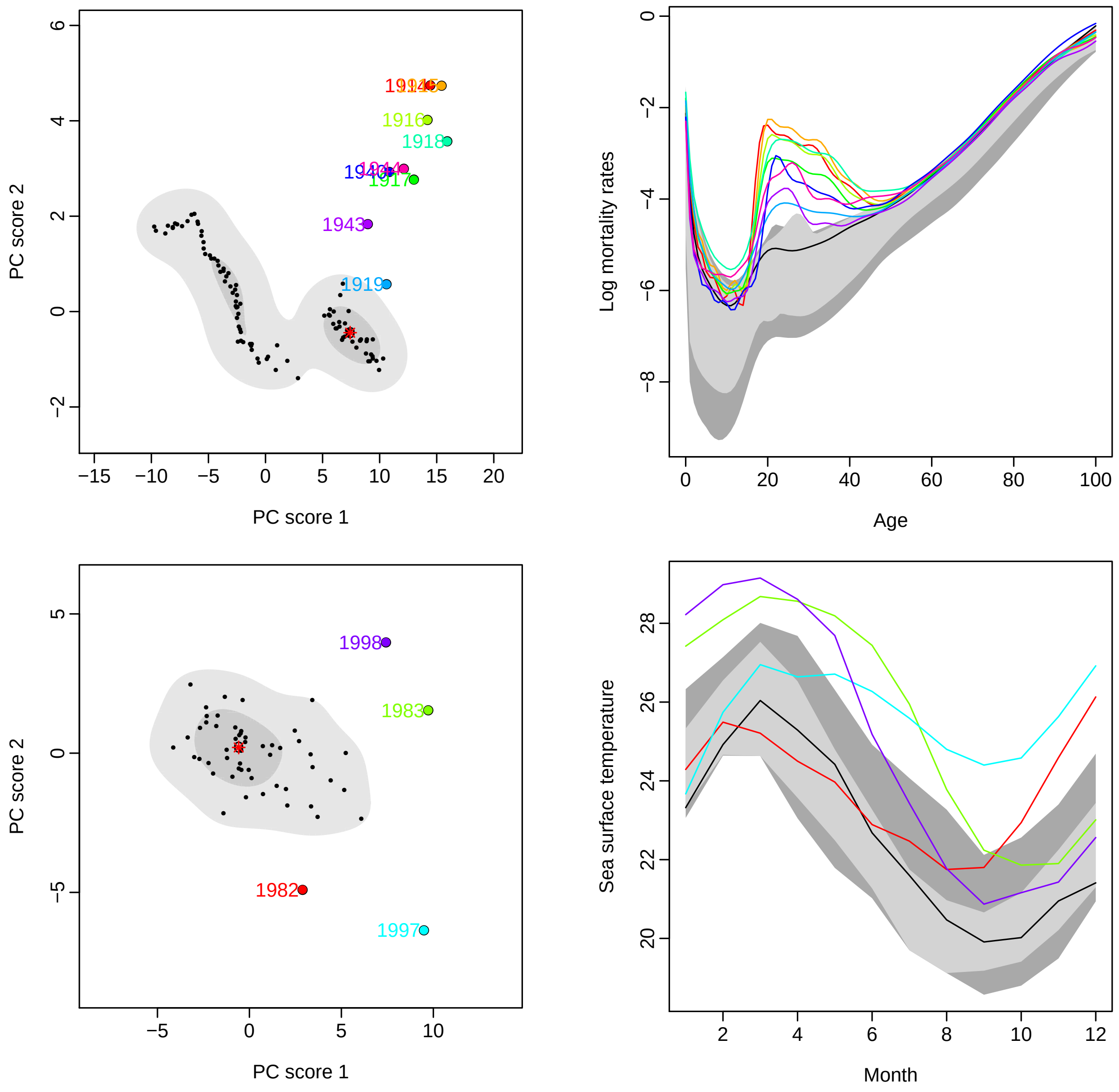


Figure 4: *Bivariate HDR boxplots and functional HDR boxplots for the French male mortality rates (top) and El Niño indexes (bottom). The dark and light grey regions show the 50% HDR and outer HDR, respectively. The black line is the modal curve. The curves outside the outer region are outliers.*

4 Outlier detection methods

The functional bagplot and functional HDR boxplot identify the outliers in the functional data. In this section we compare these outlier identification methods with other published methods.

4.1 Functional depth method

Febrero et al. (2007) proposed an outlier detection method that calculates a likelihood ratio test statistic for each curve $y_i(x)$. A point is determined to be an outlier if the maximum of the test statistics is larger than a given critical value c. This outlier is then omitted, and the remaining data are tested for another outlier. The procedure continues until no more outliers can be found. This test is based on the functional depth given by equation (1), and so is not sensitive to shape outliers.

4.2 Integrated squared error method

Hyndman & Ullah (2007) proposed an outlier detection method which utilizes robust functional principal component analysis. Let the integrated squared error for observation i be

$$v_i(x) = \int_x e_i^2(x)\,dx = \int_x \left(y_i(x) - \sum_{k=1}^{K} z_{i,k}\phi_k(x) \right)^2 dx,$$

where K is a pre-specified number of components (usually 2); $\{\phi_k(x)\}$ are the principal components functions; and $z_{i,k}$ are their associated scores. This gives a measure of the accuracy of the principal component approximation for observation i. High integrated squared errors indicate a high likelihood of the points being detected as outliers.

If $e_i(x)$ is normally distributed, then $v_i(x)$ follows a χ^2 distribution with $\mathrm{E}(v_i(x)) = 0.5\mathrm{Var}(v_i(x))$. Then, the probability that $v_i < c$, where $c = s + \lambda\sqrt{s}$ and $s = \text{median}(\{v_1, \dots, v_n\})$, is approximately $\Phi(\lambda/\sqrt{2})$, where $\Phi(\cdot)$ is the distribution function of a standard normal distribution. For example, with $\lambda = 3.29$, $\Phi(3.29/\sqrt{2}) = 99\%$.

4.3 Robust Mahalanobis distances

The robust Mahalanobis distance is a well-known multivariate outlier detection method which we can apply to discretized curves $\{y_i(x_j); j = 1, \dots, p\}$. Assuming the functional data are observed on an equally spaced dense grid $\{x_1, \dots, x_p\}$, the squared robust Mahalanobis distance is defined by

$$r_i = (y_i(x_j) - \hat{\mu}(x_j))'\hat{\Sigma}^{-1}(y_i(x_j) - \hat{\mu}(x_j)), \quad j = 1, \dots, p, \tag{2}$$

where $\hat{\mu}(x_j)$ is the sample mean, and $\hat{\Sigma}$ is a robust estimate of the covariance matrix of $\{y_i(x_j)\}$. We assume that $\hat{\Sigma}$ is positive definite, so that $\hat{\Sigma}^{-1}$ is nonsingular. The resultant distances are compared to a critical value, following a χ^2 distribution with p degrees of freedom. For a predefined $\alpha = 99\%$ level, outliers are observations that have Mahalanobis squared distances greater than the critical value $\chi^2_{0.99,p}$. Becker & Gather (2001) and Hardin & Rocke (2005) discuss the variations of the robust Mahalanobis distance further.

4.4 Location-scale method

Another multivariate outlier detection method is the location-scale approach of Filzmoser et al. (2008). This approach begins by robustly scaling the p equally-spaced discretized functions by the pointwise median and the median absolute deviation. They applied a robust principal component analysis and retained the number of principal components that can explain at least 99% of the total variation. Having robustly scaled the retained principal components, Filzmoser et al. (2008) calculated the robust Mahalanobis distance for each curve. They ordered the robust Mahalanobis distances using Rocke's (1996) translated biweight function, and assigned the weight $w_{1,i}$ to each observation. They repeated the above steps with a kurtosis weighted principal component analysis, and obtained the weight $w_{2,i}$. Outliers are detected when the weights $w_{1,i}$ and $w_{2,i}$ are both zero.

4.5 Outlier detection performance comparison

We applied the various outlier detection methods discussed above to the French male mortality data. Table 1 presents the comparative results and the relative computing speed (using a Pentium 4 CPU 3.20GHz, 512MB of RAM).

Based on historical information, we suspect that the functional outliers would be the time periods of the first and second World Wars (1914–1918 and 1939–1945) and the Spanish flu epidemic (1918–1919). These factors have affected the mortality pattern significantly, which provides an explanation for the sudden increases in mortality rates between the ages of 20 and 40. Clearly, the functional depth method has failed to detect any of these outliers, the robust Mahalanobis

Method	**Detected outliers**	**Time (secs)**
Functional depth	None	18.83
Integrated squared error	1914–1918, 1940, 1943–1945	3.41
Functional bagplot	1914–1919, 1940, 1943–1944	0.30
Functional HDR boxplot	1914–1919, 1940, 1943–1944	0.04
Location-scale	1914–1918, 1940, 1943–1944, 1953, 1960, 1992–2003	0.09
Robust Mahalanobis distance	1914–1918, 1940, 1944	1.42

Table 1: *A comparison of the outlier detection performances and computational speeds of the various methods applied to the French male mortality data.*

distance method has failed to detect some outliers, and the location-scale method has incorrectly detected a large number of years that were not outliers. The remaining methods all do quite well at identifying the outliers.

As discussed earlier, the expected outliers in the El Niño data set are the years 1982, 1983, 1997 and 1998. As can be seen in Table 2, the functional depth method and the integrated squared error method have failed to detect outliers correctly, and the location-scale method has incorrectly detected many years that were not outliers. The robust Mahalanobis distance method, the functional bagplot and the functional HDR boxplot identify the outliers equally well.

In these examples, the functional depth approach performed the worst among all methods in terms of both accuracy and computing speed. This is in accordance with the analysis of López-Pintado & Romo (2006), who inferred that the functional depth approach does not take shape outliers into account, and thus could potentially result in false detection. In contrast, the functional bagplot and functional HDR boxplot achieve the highest outlier detection accuracy, followed by the location–scale method, the robust Mahalanobis distance approach, and the integrated squared error method.

The location–scale method tends to identify far more outliers than actually exist. The integrated squared error method depends on the number of components used in the principal component approximation, which makes it rather too subjective for regular use; it is also computationally

Method	**Detected outliers**	**Time (secs)**
Functional depth	1983, 1997	15.7
Integrated squared error	1973, 1982–1983, 1997–1998	0.85
Functional bagplot	1982–1983, 1997–1998	0.33
Functional HDR boxplot	1982–1983, 1997–1998	0.02
Location–scale	1968, 1972–1973, 1982–1983, 1987, 1992, 1997–1998, 2007	0.05
Robust Mahalanobis distance	1982–1983, 1997–1998	8.00

Table 2: *A comparison of the outlier detection performances and computational speeds of the various methods applied to the El Niño indexes.*

slow, because the integrated sum squared error has to be calculated for each curve. In addition, this approach often fails to detect outliers. The robust Mahalanobis distance approach is also likely to identify too few outliers, as was shown by Filzmoser (2004).

As a third example, we simulated 990 curves of the form $y_i(x) = a_i \sin(x) + b_i \cos(x)$, where $0 < x < 2\pi$, and a_i and b_i follow independent uniform distributions with limits of 0.0 and 0.1. Ten additional curves were also randomly simulated with the same functional form, but with a_i and b_i following uniform distributions with limits of 0.1 and 0.12. The simulated data are shown in Figure 5.

Using a 99% coverage probability for the outer region, a bagplot and a boxplot for the data are shown in Figure 6. The depth approaches have failed to identify any outliers because the curves are not sufficiently distant from the median. In contrast, the HDR boxplot identifies outliers correctly. Table 3 presents the comparative results and the relative computing speed.

Extensive simulation studies are available in the online supplements to further examine the performance of outlier detection methods. In the case of extreme outliers, the robust Mahalanobis distance approach improves the performance of outlier detection significantly. However, there is no difference between the functional bagplot and the functional HDR boxplot, which detect the same outliers.

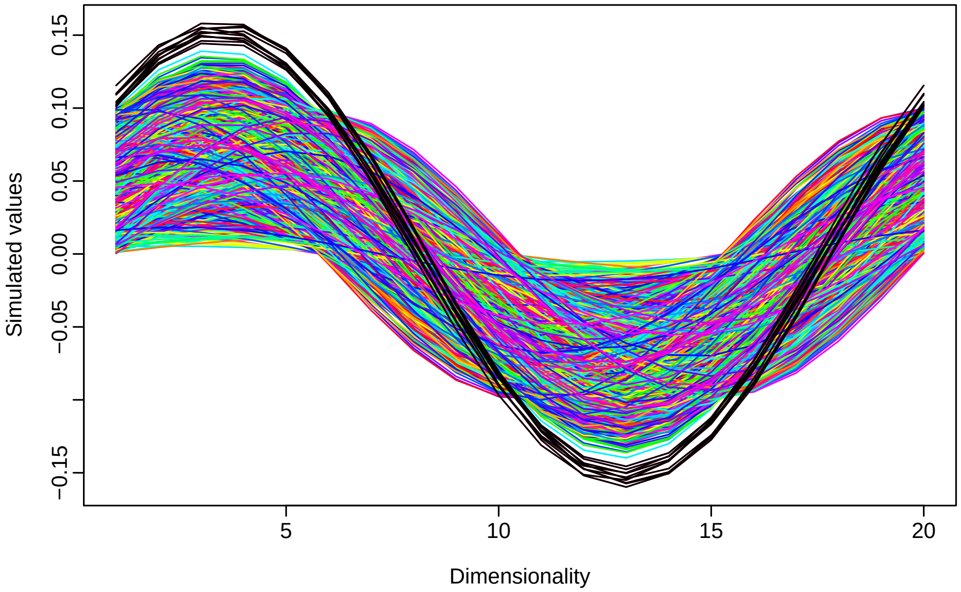


Figure 5: *1000 simulated functional curves with a functional form of $y_i(x) = a_i sin(x) + b_i cos(x)$. The 10 black curves are outliers.*

Method	**No. of detected outliers**	**Computing time (secs)**
Functional depth	None	28.5
Integrated squared error	None	18.82
Functional bagplot	None	0.56
Functional HDR boxplot	10	0.02
Location–scale	10	0.08
Robust Mahalanobis distance	1	9.72

Table 3: *A comparison of the outlier detection performances and computational speeds of two proposed approaches, two existing functional methods and two high-dimensional multivariate outlier detection methods, based on a simulated data set.*

5 Discussion and conclusions

In this paper, we have proposed three graphical tools for visualizing functional data and identifying functional outliers. Ranking principal component scores by data depth or data density is done in a

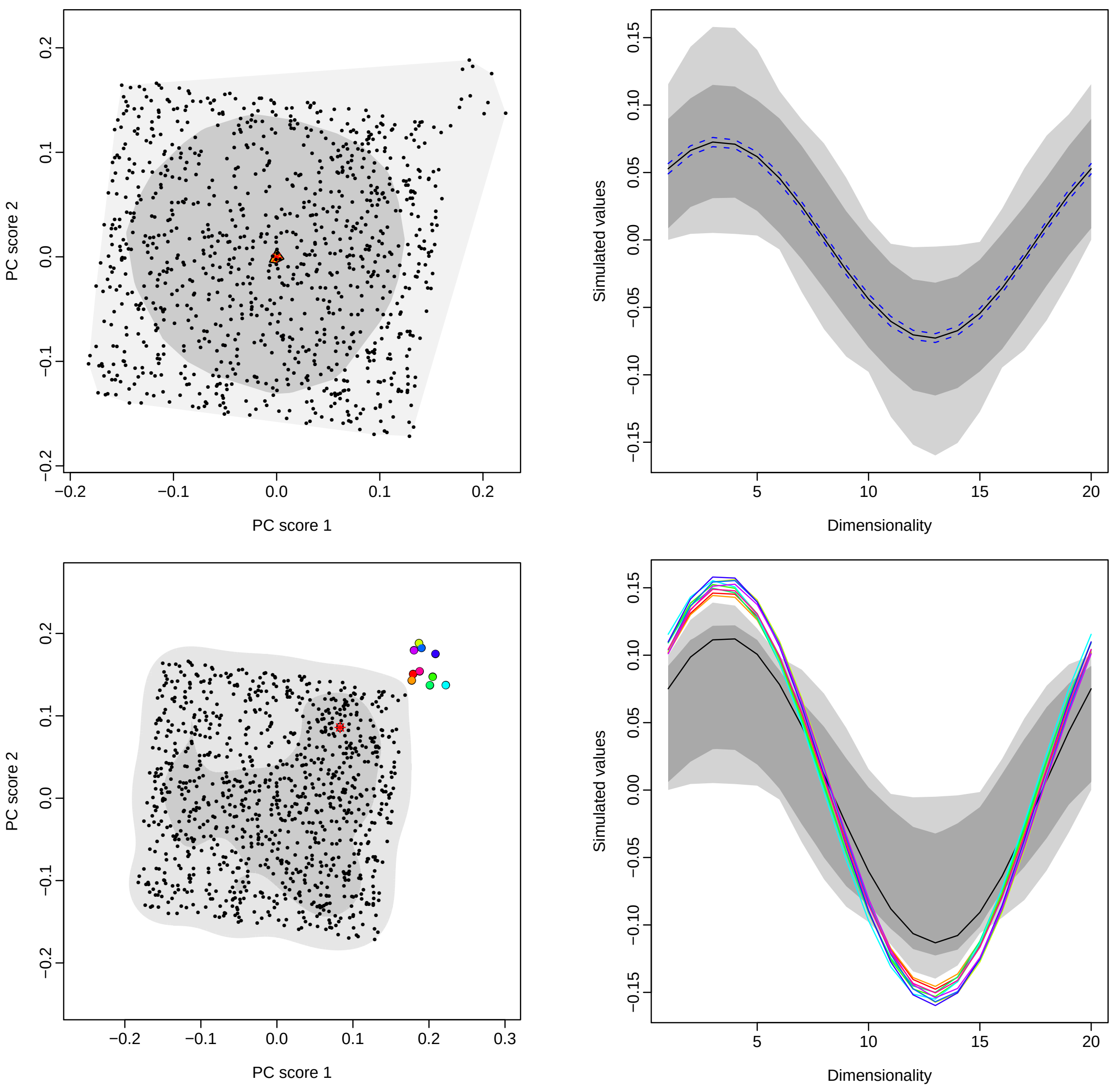


Figure 6: *1000 simulated functional curves with a functional form of $y_i(x) = a_i sin(x) + b_i cos(x)$. The 10 colored curves are outliers.*

familiar two-dimensional space, from which outliers and inliers are separated. Graphical displays are achieved by matching the scores obtained from both a bivariate bagplot and a HDR boxplot back to the functional curves.

The advantages of the proposed approaches are that they detect outliers accurately with a fast computational speed, while simultaneously providing a graphical representation. As has been illustrated using two real data sets, the proposed methods perform better than the existing approaches for outlier detection, which either identify spurious outliers or miss obvious outliers.

Through a simulated data set, we have demonstrated that the depth-based methods fail to detect some outliers that are not far from the median curve. In contrast, a density-based approach, such as the HDR boxplot, can identify such outliers correctly.

The methods presented in this paper can easily be extended in three directions. The principal components decomposition that is used in several of our proposed methods could be replaced with other dimension reduction methods such as independent component analysis (Epifanio, 2008) or partial least squares (Faber et al., 2003). Tukey's (1974) halfspace depth may also be replaced by other depth measures if they are more appropriate for capturing a certain aspect of the data. For example, if the underlying distribution is close to elliptical, then it is more efficient to use the Mahalanobis depth approach. Finally, other methods for ordering functional data or determining the functional median and mode can be utilized, such as the methods proposed by Gasser et al. (1998) and Ferraty & Vieu (2006, Chapter 9).

The R-package `rainbow` for constructing rainbow plots, functional bagplots and functional HDR boxplots is available on CRAN (`http://cran.r-project.org/`).

Appendix A. Proof on the coverage probability for bagplots

Because of the affine invariant property of depth, we assume, without loss of generality, a bivariate standard normal distribution with center $O = (0, 0)$.

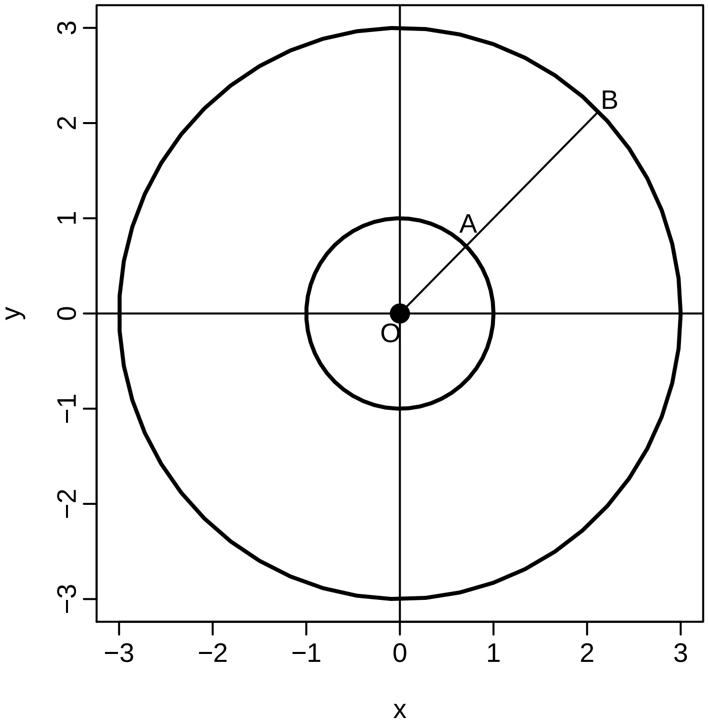


Figure 7: *A concept diagram to show the inflation factor in the bagplot.*

Let A be a point on the ellipse of a bivariate standard normal distribution associated with a constant probability density contour with coverage 0.5. Similarly, let B be a point on the ellipse associated with a probability density contour with coverage $1-\alpha$. We denote the distances between the origin and A and B as $k_{0.5}$ and k_α respectively. Let $\rho_\alpha = k_\alpha / k_{0.5}$ be the inflation factor used in constructing the fence of the bagplot. Then α is the probability of any point being an outlier when the original data are bivariate normal.

To compute k_α, we must solve the following bivariate integral:

$$\iint\limits_{x^2+y^2 \le k_\alpha^2} \frac{1}{2\pi} \exp\left[-\frac{1}{2}(x^2+y^2)\right] dx\, dy = 1 - \alpha.$$

Switching to polar coordinates, we obtain

$$\frac{1}{2\pi}\int^{k_\alpha}\int_0^{2\pi} e^{-r^2/2}\, r\, dr\, d\theta = \frac{1}{2}\int^{k_\alpha^2} e^{-u/2} du, \tag{3}$$

where $u = r^2$. The expression (3) under the integral can be recognized immediately as a χ^2 density with 2 degrees of freedom. Therefore, $k_{0.5} = 1.1774$, $k_{0.99} = 3.0349$ and $\rho_{0.01} = 2.5776$. By comparison, $\rho_{0.002} = 3$.

Appendix B. Supplemental materials

R package for rainbow The R-package "rainbow" contains functions for constructing rainbow plots, functional bagplots and functional HDR boxplots as described in this article. The package also contains all data sets used as examples in the article. The R-package can be obtained from CRAN (`http://cran.r-project.org/`).

Simulation results Extensive simulation results to further compare the performance of outlier detection methods.

References

Becker, C. & Gather, U. (2001), 'The largest nonidentifiable outlier: a comparison of multivariate simultaneous outlier identification rules', *Computational Statistics and Data Analysis* **36**(1), 119–127.
`http://ideas.repec.org/a/eee/csdana/v36y2001i1p119-127.html`

Croux, C. & Ruiz-Gazen, A. (2005), 'High breakdown estimators for principal components: the projection-pursuit approach revisited', *Journal of Multivariate Analysis* **95**(1), 206–226.
`http://ideas.repec.org/a/eee/jmvana/v95y2005i1p206-226.html`

Dioses, T., Dávalos, R. & 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.
`http://www.scielo.cl/scielo.php?pid=S0717-71782002030100077&script=sci_arttext`

Duong, T. & Hazelton, M. L. (2005), 'Cross-validation bandwidth matrices for multivariate kernel density estimation', *Scandinavian Journal of Statistics* **32**(3), 485–506.
`http://www3.interscience.wiley.com/journal/118653028/abstract`

Epifanio, I. (2008), 'Shape descriptors for classification of functional data', *Technometrics* **50**(3), 284–294.
`http://pubs.amstat.org/doi/abs/10.1198/004017008000000154?journalCode=tech`

Eubank, R. L. (1999), *Nonparametric regression and spline smoothing*, 2nd edn, CRC, New York.
`http://www.amazon.com/dp/0824793374`

Faber, N., Song, X. & Hopke, P. (2003), 'Sample-specific standard error of prediction for partial least squares regression', *TrAC Trends in Analytical Chemistry* **22**(5), 330–334.
`http://cat.inist.fr/?aModele=afficheN&cpsidt=14731612`

Febrero, M., Galeano, P. & Gonzalez-Manteiga, W. (2007), 'A functional analysis of NOx levels: location and scale estimation and outlier detection', *Computational Statistics* **22**(3), 411–427.
`http://www.springerlink.com/content/146v81216v078582/`

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**(4), 331–

345.
http://www3.interscience.wiley.com/journal/114802319/abstract

Ferraty, F. & Vieu, P. (2006), *Nonparametric functional data analysis: theory and practice*, Springer, New York.
http://www.springer.com/statistics/statistical+theory+and+methods/book/978-0-387-30369-7

Filzmoser, P. (2004), A multivariate outlier detection method, *in* S. Aivazian, P. Filzmoser & Y. Kharin, eds, 'Proceedings of the Seventh International Conference on Computer Data Analysis and Modeling', Vol. 1, Belarusian State University, Minsk, pp. 18–22.
http://www.statistik.tuwien.ac.at/public/filz/papers/minsk04.pdf

Filzmoser, P., Maronna, R. & Werner, M. (2008), 'Outlier identification in high dimensions', *Computational Statistics & Data Analysis* **52**(3), 1694–1711.
http://ideas.repec.org/a/eee/csdana/v52y2008i3p1694-1711.html

Gasser, T., Hall, P. & Presnell, B. (1998), 'Nonparametric estimation of the mode of a distribution of random curves', *Journal of the Royal Statistical Society: Series B* **60**(4), 681–691.
http://www.jstor.org/stable/2985956

Hardin, J. & Rocke, D. M. (2005), 'The distribution of robust distances', *Journal of Computational and Graphical Statistics* **14**(4), 928–946.
http://pubs.amstat.org/doi/abs/10.1198/106186005X77685

Human Mortality Database (2008), *University of California, Berkeley (USA), and Max Planck Institute for Demographic Research (Germany).* Viewed 15/4/07, available online at <www.mortality.org> or <www.humanmortality.de>.
http://www.mortality.org/

Hyde, V., Jank, W. & Shmueli, G. (2006), 'Investigating concurrency in online auctions through visualization', *The American Statistician* **60**(3), 241–250.
http://pubs.amstat.org/doi/abs/10.1198/000313006X124163

Hyndman, R. J. (1996), ‘Computing and graphing highest density regions’, *The American Statistician* **50**(2), 120–126.
`http://www.jstor.org/stable/2684423`

Hyndman, R. J. & Ullah, M. S. (2007), ‘Robust forecasting of mortality and fertility rates: A functional data approach’, *Computational Statistics & Data Analysis* **51**(10), 4942–4956.
`http://ideas.repec.org/a/eee/csdana/v51y2007i10p4942-4956.html`

Jones, M. C. & Rice, J. A. (1992), ‘Displaying the important features of large collections of similar curves’, *The American Statistician* **46**(2), 140–145.
`http://www.jstor.org/stable/2684184`

Kargin, V. & Onatski, A. (2008), ‘Curve forecasting by functional autoregression’, *Journal of Multivariate Analysis* **99**(10), 2508–2526.
`http://ideas.repec.org/p/sce/scecf5/59.html`

Locantore, N., Marron, J. S., Simpson, D. G., Tripoli, N., Zhang, J. T. & Cohen, K. L. (1999), ‘Robust principal component analysis for functional data’, *Test* **8**(1), 1–73.
`http://www.springerlink.com/content/96126443p02r6774/`

López-Pintado, S. & Romo, J. (2006), On the concept of depth for functional data, Technical Report 06-30, Departamento de Estadística, Universidad Carlos III de Madrid.
`http://docubib.uc3m.es/WORKINGPAPERS/WS/ws063012.pdf`

Moran, E. F., Adams, R., Bakoyéma, B., Stefano, F. T. & Boucek, B. (2006), ‘Human strategies for coping with El Niño related drought in Amazônia’, *Climatic Change* **77**, 343–361.
`http://www.springerlink.com/content/l1710m30j5415270/`

Ramsay, J. O. & Ramsey, J. B. (2002), ‘Functional data analysis of the dynamics of the monthly index of nondurable goods production’, *Journal of Econometrics* **107**(1-2), 327–344.
`http://ideas.repec.org/a/eee/econom/v107y2002i1-2p327-344.html`

Ramsay, J. O. & Silverman, B. W. (2002), *Applied Functional Data Analysis: methods and case studies*, Springer, New York; London.
`http://www.amazon.com/dp/0387954147`

Ramsay, J. O. & Silverman, B. W. (2005), *Functional Data Analysis*, 2nd edn, Springer, New York.
http://www.springer.com/statistics/statistical+theory+and+methods/book/978-0-387-40080-8

Reiss, P. T. & Ogden, T. R. (2007), 'Functional principal component regression and functional partial least squares', *Journal of the American Statistical Association* **102**(479), 984–996.
http://works.bepress.com/phil_reiss/7/

Rocke, D. M. (1996), 'Robustness properties of S-estimators of multivariate location and shape in high dimension', *The Annals of Statistics* **24**(3), 1327–1345.
http://www.jstor.org/stable/2242597

Rousseeuw, P., Ruts, I. & Tukey, J. W. (1999), 'The bagplot: A bivariate boxplot', *The American Statistician* **53**(4), 382–387.
http://cat.inist.fr/?aModele=afficheN&cpsidt=1244534

Scott, D. W. (1992), *Multivariate density estimation: theory, practice, and visualization*, Wiley, New York.
http://www.wiley.com/WileyCDA/WileyTitle/productCd-0471547700.html

Sood, A., James, G. M. & Tellis, G. J. (2009), 'Functional regression: a new model for predicting market penetration of new products', *Marketing Science* **28**(1), 36–51.
http://mktsci.journal.informs.org/cgi/content/abstract/mksc.1080.0382v1

Timmermann, A., Oberhuber, J., Bacher, A., Esch, M., Latif, M. & Roeckner, E. (1999), 'Increased El Niño frequency in a climate model forced by future greenhouse warming', *Nature* **398**(6729), 694–697.
http://www.nature.com/nature/journal/v398/n6729/abs/398694a0.html

Tukey, J. W. (1974), Mathematics and the picturing of data, *in* 'Proceedings of the International Congress of Mathematicians', Vancouver, pp. 523–531.

Tukey, J. W. (1977), *Exploratory Data Analysis*, Addison-Wesley, London.
http://www.amazon.com/dp/B0007347RW

Zhang, L., Marron, J. S., Shen, H. & Zhu, Z. (2007), 'Singular value decomposition and its visualization', *Journal of Computational and Graphical Statistics* **16**(4), 833–854.
`http://pubs.amstat.org/doi/abs/10.1198/106186007X256080`