Interface Foundation of America

Rainbow Plots, Bagplots, and Boxplots for Functional Data
Author(s): Rob J. Hyndman and Han Lin Shang
Source: *Journal of Computational and Graphical Statistics*, Vol. 19, No. 1 (March 2010), pp. 29-45

Published by: Taylor & Francis, Ltd. on behalf of the American Statistical Association, Institute of Mathematical Statistics, and Interface Foundation of America
Stable URL: http://www.jstor.org/stable/25651298
Accessed: 24-08-2016 22:00 UTC

Rainbow Plots, Bagplots, and Boxplots for Functional Data

Rob J. HYNDMAN and Han Lin SHANG

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, which 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 outliers.

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

Key Words: Highest density regions; Kernel density estimation; Outlier detection; Robust principal component analysis; Tukey's halfspace location 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. Functional data consist of a collection of functions—usually smooth curves or surfaces (e.g., Locantore et al. 1999; Ramsay and Silverman 2005). In this article, we are interested in visualizing data comprising smooth curves. Such functional data may be age-specific mortality or fertility rates (Hyndman and Ullah 2007), term-structured yield curves (Kargin and Onatski 2008), spectrometry data (Reiss and Ogden 2007), or one of the many applications described by Ramsay and Silverman (2002). Ramsay and Silverman (2005) and Ferraty and Vieu (2006) provided detailed surveys of 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

Rob J. Hyndman is Professor and Han Lin Shang is Ph.D. Student (E-mail: *HanLin.Shang@buseco.monash.au*), Department of Econometrics and Business Statistics, Monash University, Clayton, VIC 3800, Australia.

© *2010 American Statistical Association, Institute of Mathematical Statistics, and Interface Foundation of North America*
Journal of Computational and Graphical Statistics, Volume 19, Number 1, Pages 29–45
DOI: 10.1198/jcgs.2009.08158

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 and Ramsey (2002) and the rug plot of Hyde, Jank, and Shmueli (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 and Ullah (2007) used a method based on robust functional principal component analysis, whereas Febrero, Galeano, and Gonzalez-Manteiga (2007, 2008) considered functional outlier detection using successive likelihood ratio tests and smoothed bootstrapping.

To motivate the discussion, Figure 1 shows annual smoothed age-specific mortality curves for French males between 1899 and 2005. The data were taken from the Human Mortality Database (2009) and smoothed using penalized splines with the partial monotonic constraint, as described by Hyndman and 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 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.

As can be seen from Figure 1, mortality rates dip in early childhood, climb in the teen years, stabilize in the early 20s, and then steadily increase with age. Some years exhibit sharp increases in mortality rates between the late teens and early 20s. Some of the mortality curves shown in yellow and green indicate sudden increases in mortality rates between

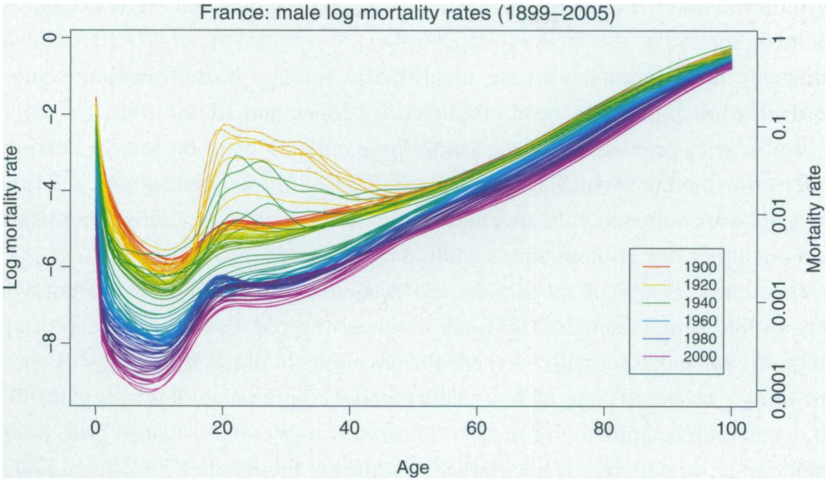

Figure 1. French male age-specific log mortality rates (1899–2005). The oldest years are shown in red, with the most recent years in violet. Curves are ordered chronologically according to the colors of the rainbow. The left vertical axis measures log mortality rates, whereas the right vertical axis adds non-log units to ease interpretation. Mortality rates dip in early childhood, climb in the teen years, stabilize in the early 20s, and then steadily increase with age. Some years exhibit sharp increases in mortality rates between the late teens and early 20s.

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.

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 visualize functional outliers if they are obscured by other curves.

We consider that a curve is an outlier if it has been generated by a stochastic process with a different distribution from the large majority of the curves. For example, a functional outlier can have a different shape or magnitude from the rest of the data. With univariate data, we commonly use boxplots to identify outliers. Therefore, we aim to define functional variations on the boxplot 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 the first two principal component scores obtained by applying a robust principal component algorithm to $\{y_i(x)\}$. These bivariate points can then be ordered by applying Tukey's (1975) halfspace location depth, for example. This idea immediately lends itself to a functional bagplot obtained by applying the bivariate bagplot (Rousseeuw, Ruts, and Tukey 1999) to the first two robust principal component scores, and then mapping the features of the bagplot into the functional space. This idea has recently been used by Sood, James, and Tellis (2009) for clustering functional data.

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

Outliers in the functional data are identified as outliers in the bivariate score space. A related idea has previously been employed by Jones and 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 HDR boxplot are compared with several existing functional outlier detection methods and multivariate outlier detection methods applied to the discretized functions.

Our functional graphical techniques are related to the parallel coordinate plots of Inselberg (1985) and Wegman (1990). In essence, the parallel coordinate plot approximates a multivariate dataset with a hyperbolic envelope. It allows the display of correlation structure among observed data. Although the parallel coordinate plot is a very useful device in itself, it can only be applied to functions observed on an equally spaced grid, where each grid point can be considered as a variable. In addition, the parallel coordinate plot suffers from heavy overplotting with large high-dimensional datasets. Our functional graphical techniques are proposed to address these problems in the context of functional data.

Section 5 provides a summary of our main results, and some thoughts on how the plots might be extended. In this article, we do not address the issue of preprocessing data. For a noisy dataset or an unequally spaced dataset, it is advisable to preprocess data using techniques such as curve registration methods (Ramsay and Silverman 2005, chap. 10) or nonparametric smoothing methods (Eubank 1999), 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 datasets 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 1975) 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, Galeano, and Gonzalez-Manteiga (2007) proposed an outlier detection method that used the notion of functional depth, defined as

$$o_i = \int D(y_i(x))\,dx, \tag{2.1}$$

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 a decreasing order of $\{o_i\}$, so the first curve has the greatest functional depth and the last curve has the lowest functional depth.

However, as the functional depth is calculated by integrating the univariate depth, this method may not detect curves that have unusual shapes, but lie within the range of the majority of curves (López-Pintado and Romo 2009). Consequently, this definition of functional depth is not adequate for many functional datasets. We thus propose a measure of functional depth that is based on the first two robust principal component scores instead.

2.2 BIVARIATE SCORE DEPTH

Let $\{\phi_k(x)\}$ represent the principal components, and $\{z_{i,k}\}$ denote the principal component scores from a robust 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 their associated scores (Jones and Rice 1992; Sood, James, and Tellis 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 $\mathbf{z}_i$.

Because principal component decomposition is not resistant to outliers, we apply Croux and 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 advantage 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 1975), denoted by $d(\boldsymbol{\theta}, \mathbf{Z})$ for some point $\boldsymbol{\theta} \in R^2$ relative to the bivariate data cloud $\mathbf{Z} = \{\mathbf{z}_i; i = 1, \dots, n\}$. 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 $o_i = d(\mathbf{z}_i, \mathbf{Z})$ in an increasing order. The first curve by this ordering can be considered as a "median" curve, whereas the last curve can be considered as the outermost curve.

2.3 BIVARIATE SCORE DENSITY

The third way to order the points is by the value of a bivariate kernel density estimate (Scott 1992) at each observation. Let $o_i = \hat{f}(\mathbf{z}_i)$, where $\hat{f}(\mathbf{z})$ is a bivariate kernel density estimate calculated from all of the bivariate robust 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," whereas the last curve may be considered the most unusual curve.

Note that the last curve by this ordering may not take values that are very different from the others and its bivariate scores may not be on the edge of the scatterplot of $\{(z_{i,1}, z_{i,2})\}$. 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 Plot

For data that are not naturally ordered by time, or other indexes, the rainbow plot can still be used by constructing an ordering index, such as the data depth or data density 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 in degrees Celsius from January 1951 to December 2007, available online at *http://www.cpc.noaa.gov/data/indices/sstoi.indices*. These sea surface temperatures are measured by moored buoys in the "Niño region" defined by the coordinates 0–10° South and 90–80° West. In this case, each curve represents one year of observed sea surface temperatures. There is no time trend in these data, so a rainbow plot with time ordering is not particularly informative.

Rainbow plots using the 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 closer to the center of the dataset shown in red, whereas the outlying curves with lower depth are shown in violet. Similarly, the curves having higher density are shown in red, whereas the outlying curves with lower density are shown in violet. We plot the curves in order of depth and 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.

3.2 Functional Bagplot

The functional bagplot is based on the bivariate bagplot of Rousseeuw, Ruts, and Tukey (1999), applied to the first two robust principal component scores. It uses Tukey's (1975)

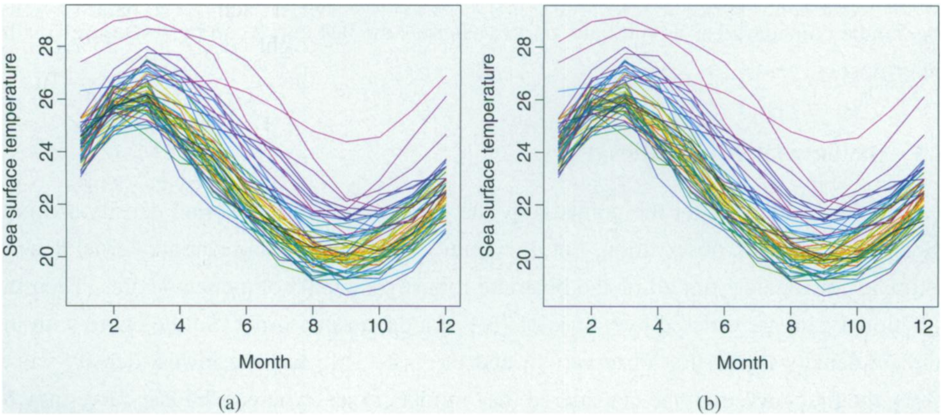

Figure 2. Rainbow plots of sea surface temperatures using different order indexes. (a) Rainbow plot of sea surface temperatures with the depth ordering. (b) Rainbow plot of sea surface temperatures with the density ordering. The black lines show the median curve in (a) and the modal curve in (b). In (a), the red lines show the curves with the higher depth, whereas the violet lines show the outlying curves with the lower depth. In (b), the red lines show the curves with the higher density, whereas the violet lines show the outlying curves with the lower density.

halfspace location depth. The depth region D_k is the set of all $\boldsymbol{\theta}$, with $d(\boldsymbol{\theta}, \mathbf{z}) \geq k$. Because 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 median is defined as the value of $\boldsymbol{\theta}$ which minimizes $d(\boldsymbol{\theta}, \mathbf{Z})$ 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 (the "fence"), 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 of the bagplot is the convex hull of the points containing the region obtained by inflating the bag (relative to the Tukey median) by a factor ρ. Rousseeuw, Ruts, and Tukey (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 a standard normal distributions. A proof of this result is given in the online supplements.

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 sea surface temperature data. In the left panels, the dark gray regions show the 50% bag and the light gray regions exhibit the 99% fence. These convex hulls correspond directly to the regions of similar shading in the functional bagplot on the right. Points outside the fence 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 robust principal component scores on the left. The red asterisk marks the Tukey median of the bivariate robust 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 curves (similar to the notched boxplot of Tukey 1977).

The outliers detected 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 (1918–1919). The detected outliers in the sea surface temperature data are the years 1982–1983 and 1997–1998. The sea surface temperatures during 1982–1983 began in June 1982 with a weak heating, then there were abnormal increases between September 1982 and June 1983 (Timmermann et al. 1999; Moran et al. 2006). The sea surface temperatures during 1997–1998 were also unusual—they became extreme in the latter half of 1997, and stayed high for the early part of 1998. Dioses, Dávalos, and Zuzunaga (2002) reported that the northern central region of Peru was strongly affected as warm waters with low salinity approached the coast, whereas 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

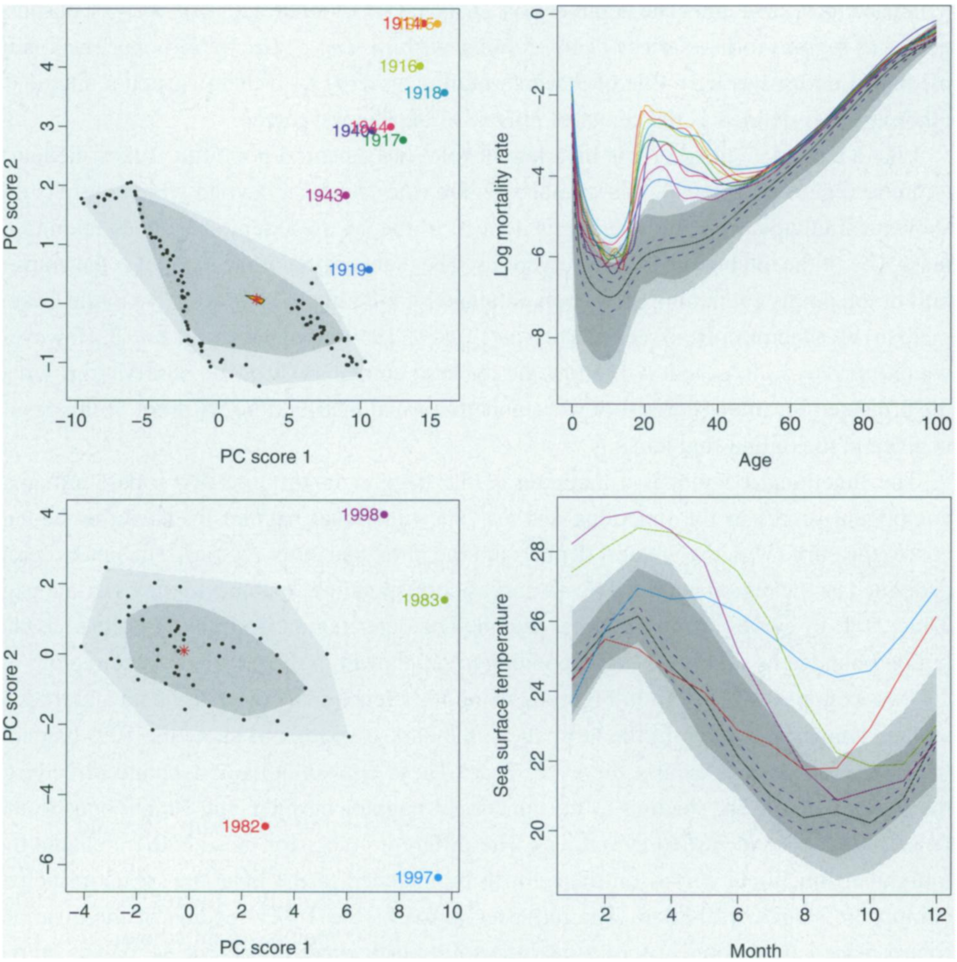

Figure 3. Bivariate bagplots and functional bagplots for the French male log mortality rates (top) and the sea surface temperatures (bottom). The dark and light gray regions show the bag and fence regions, respectively. In the left panels, the red asterisks denote the Tukey medians. The points outside the fence regions are outliers. In the right panels, the black lines are the median curves, surrounded by 95% pointwise confidence intervals. The curves outside the fence regions are shown as outliers of different colors.

outlier detection tool can misidentify outliers, as is shown in Section 4.5 via a simulated dataset. 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 robust principal component scores. The bivariate HDR boxplot is constructed using a bivariate kernel density estimate $\hat{f}(\mathbf{z})$, which is defined as

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

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 and Hazelton 2005).

Using the kernel density estimates, a HDR is defined as

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

where f_α is such that $\int_{R_\alpha} \hat{f}(\mathbf{z})\, d\mathbf{z} = 1 - \alpha$; that is, it is the region with coverage probability $1 - \alpha$, 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 highest density regions can be considered as contours, with an expanding coverage as α decreases.

The bivariate HDR boxplot displays the mode (the highest density point), defined as $\arg\sup \hat{f}(\mathbf{z})$, along with the 50% inner and (usually) 99% outer 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 sea surface temperature data. In the left panels, the dark and light gray 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 the outer regions are identified as outliers. The use of different colors for these outliers enables the individual curves on the right to be matched to the bivariate scores on the left. The red asterisk in each of the left panels marks the mode of the bivariate robust principal component scores, and corresponds to the solid black curve in each of the right panels.

As with any outlier detection method, including bagplots and HDR boxplots, the coverage probability of the outer region needs to be prespecified. With the 99% coverage probability, the outliers detected in the French male mortality data are 1919 and 1943, whereas the outlier detected in the sea surface temperature data 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 sea surface temperature 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.

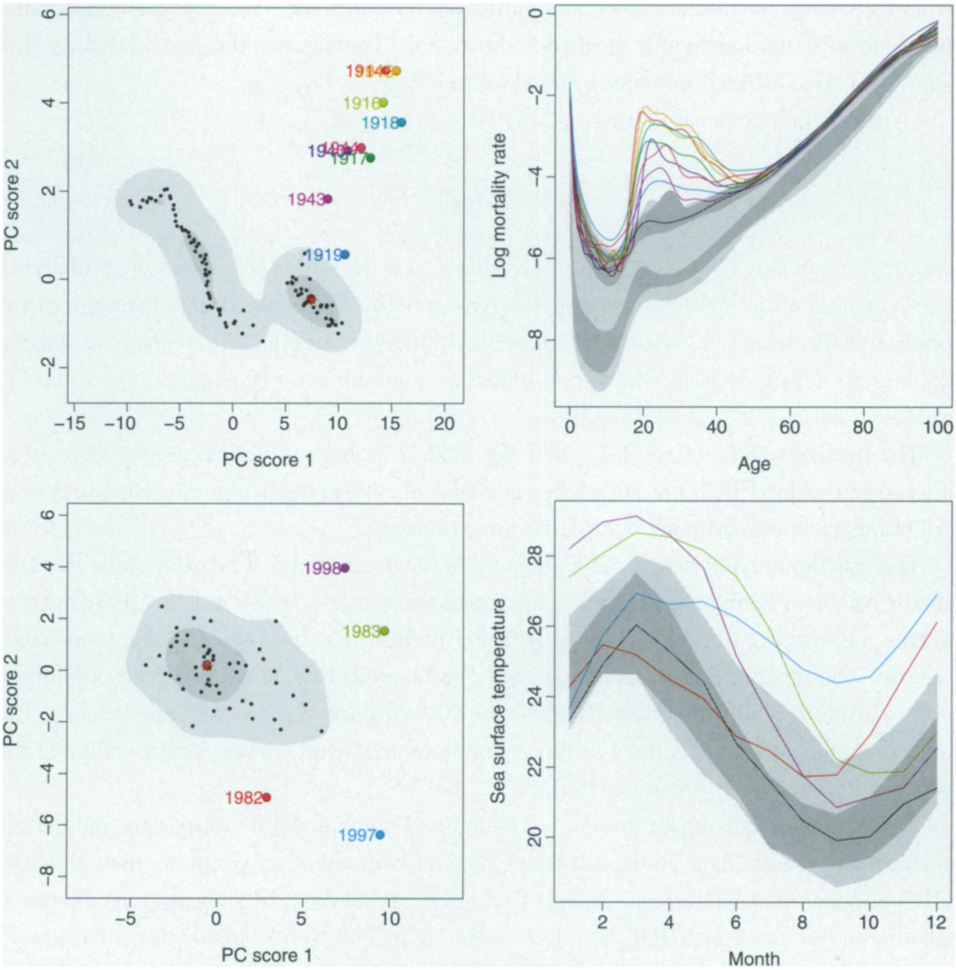

Figure 4. Bivariate HDR boxplots and functional HDR boxplots for the French male log mortality rates (top) and the sea surface temperatures (bottom). The dark and light gray regions show the 50% HDR and outer HDR, respectively. The red asterisks are the modes (the highest density points). The points outside the outer HDR are outliers. In the right panels, the black lines are the modal curves. The curves outside the outer HDR are shown as outliers of different colors.

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, Galeano, and Gonzalez-Manteiga (2007) proposed an outlier detection method that calculates a likelihood ratio test statistic for each curve $y_i(x)$. A curve 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 (2.1), and so is not sensitive to shape outliers.

4.2 INTEGRATED SQUARED ERROR METHOD

Hyndman and Ullah (2007) proposed an outlier detection method that utilizes robust functional principal component analysis. Let the integrated squared error for each observation 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 prespecified number of components (usually 2); $\{\phi_k(x), k = 1, \ldots, K\}$ is a set of principal component functions; and $z_{i,1}, \ldots, z_{i,K}$ are their associated scores. This gives a measure of the accuracy of the principal component approximation for each observation. High integrated squared errors indicate a high likelihood of the curves 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 $s = \mathrm{median}(\{v_1, \ldots, v_n\})$, 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 DISTANCE METHOD

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

$$r_i = (y_i(x_j) - \hat{\mu}(x_j))'\hat{\Sigma}^{-1}(y_i(x_j) - \hat{\mu}(x_j)), \qquad j = 1, \ldots, p, \tag{4.1}$$

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 squared robust Mahalanobis distances greater than the critical value $\chi^2_{0.99,p}$. Becker and Gather (2001) and Hardin and Rocke (2005) discussed 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, Maronna, and Werner (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

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

Method	Detected outliers	Time (sec)
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

scaled the retained principal components, Filzmoser, Maronna, and Werner (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 speeds (using a Pentium 4 CPU 3.20 GHz, 512 MB 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, and this 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 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 sea surface temperature data are the years 1982–1983 and 1997–1998. As in Table 2, the functional depth method and the integrated squared error method have failed to detect outliers correctly, and the location-scale

Table 2. A comparison of the outlier detection performances and computational speeds of the various methods applied to the sea surface temperature data.

Method	Detected outliers	Time (sec)
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

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 and Romo (2009), 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 slow, because the integrated 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 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. The curves are colored in a random order except that the outliers are shown in black.

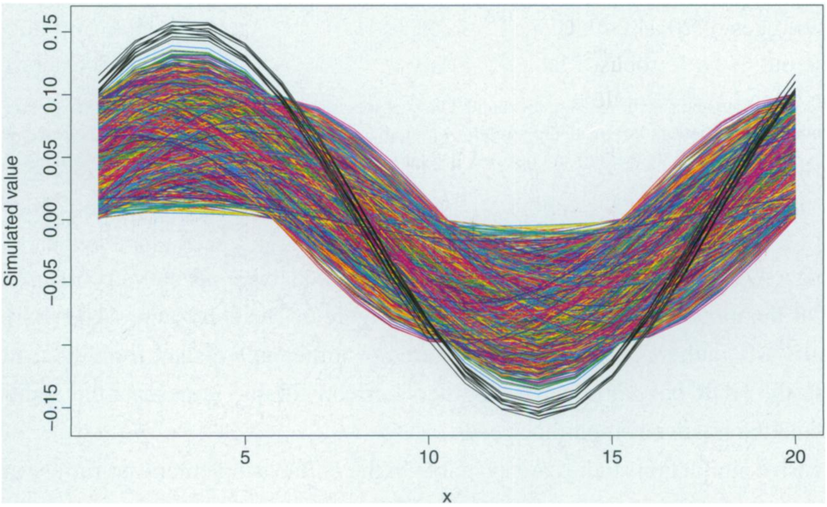

Figure 5. One thousand simulated functional curves of the form $y_i(x) = a_i \sin(x) + b_i \cos(x)$, where $0 < x < 2\pi$. In the first 990 curves, a_i and b_i follow independent uniform distributions with limits of 0 and 0.1. In the last 10 curves shown in black, a_i and b_i follow independent uniform distributions with limits of 0.1 and 0.12.

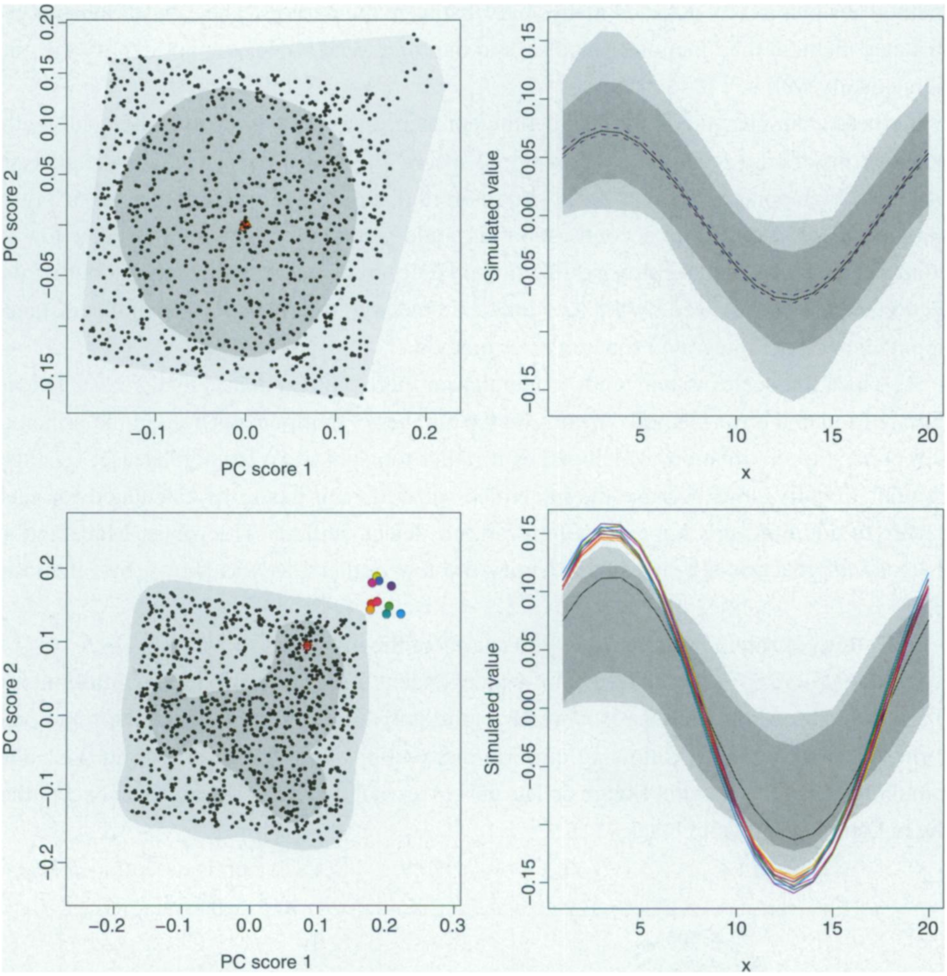

Figure 6. Bivariate and functional bagplots and HDR boxplots for the 1000 simulated functional curves. The 10 colored points in the bottom left panel are outliers, which match the 10 colored curves in the bottom right panel. This is an example where the outliers are not very distant from the median, but have much lower density.

Using a 99% coverage probability for the outer region, bivariate and functional bagplots and HDR boxplots for these 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 speeds.

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

Table 3. A comparison of the outlier detection performances and computational speeds of the various methods applied to the simulated data.

Method	No. of detected outliers	Computing time (sec)
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

5. DISCUSSIONS AND CONCLUSIONS

In this article, we have proposed three graphical tools for visualizing functional data and identifying functional outliers. Ranking robust principal component scores by data depth or data density was done in a 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 advantage of the proposed approaches is that they detect outliers accurately with a fast computational speed, while simultaneously providing a graphical representation. As it has been illustrated using two real datasets, the proposed methods perform better than the existing approaches for outlier detection, which either identify spurious outliers or miss obvious outliers.

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

The methods presented in this article can easily be extended in several directions. The principal component 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, Song, and Hopke 2003). Other methods for ordering functional data or determining the functional median and mode can be utilized, such as the methods proposed by Gasser, Hall, and Presnell (1998) and Ferraty and Vieu (2006, chap. 9). Tukey's (1975) halfspace location 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 ellipse, then it is more efficient to use the Mahalanobis depth. Finally, it is possible to extend the proposed techniques from two-dimensional functional curves to three- or higher-dimensional functional surfaces. The idea is to apply the principal component analysis to an array rather than a matrix to obtain the order indexes of functional surfaces. With three-dimensional surfaces, it is possible to produce a movie with the ordered surface images. Computationally, this can be achieved using XGobi (Swayne, Cook, and Buja 1998) or the rgl package (Adler and Murdoch 2009) in R.

SUPPLEMENTAL MATERIALS

Proof and simulation results: Proof of the coverage probability for bagplot; refer to Section 3.2. Extensive simulation results to further compare the performances of outlier detection methods; refer to Section 4.5. (Proof and simulation results.pdf)

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 datasets used as examples in this article. The R-package can also be obtained from CRAN (*http://cran.r-project.org/package=rainbow*). (rainbow_1.4.zip)

ACKNOWLEDGMENTS

The authors thank the editor, an associate editor, and two reviewers for their insightful comments which led to a substantial improvement of the manuscript.

[Received November 2008. Revised September 2009.]

REFERENCES

Adler, D., and Murdoch, D. (2009), "rgl: 3D Visualization Device System (OpenGL)," R package version 0.87, available at *http://CRAN.R-project.org/package=rgl.* [43]

Becker, C., and Gather, U. (2001), "The Largest Nonidentifiable Outlier: A Comparison of Multivariate Simultaneous Outlier Identification Rules," *Computational Statistics & Data Analysis*, 36 (1), 119–127. [39]

Croux, C., and Ruiz-Gazen, A. (2005), "High Breakdown Estimators for Principal Components: The Projection-Pursuit Approach Revisited," *Journal of Multivariate Analysis*, 95 (1), 206–226. [33]

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. [35]

Duong, T., and Hazelton, M. L. (2005), "Cross-Validation Bandwidth Matrices for Multivariate Kernel Density Estimation," *Scandinavian Journal of Statistics*, 32 (3), 485–506. [37]

Epifanio, I. (2008), "Shape Descriptors for Classification of Functional Data," *Technometrics*, 50 (3), 284–294. [43]

Eubank, R. L. (1999), *Nonparametric Regression and Spline Smoothing* (2nd ed.), New York: CRC. [32]

Faber, N. M., Song, X.-H., and Hopke, P. K. (2003), "Sample-Specific Standard Error of Prediction for Partial Least Squares Regression," *TrAC Trends in Analytical Chemistry*, 22 (5), 330–334. [43]

Febrero, M., Galeano, P., and Gonzalez-Manteiga, W. (2007), "A Functional Analysis of NOx Levels: Location and Scale Estimation and Outlier Detection," *Computational Statistics*, 22 (3), 411–427. [30,32,38]

——— (2008), "Outlier Detection in Functional Data by Depth Measures, With Application to Identify Abnormal NOx Levels," *Environmetrics*, 19 (4), 331–345. [30]

Ferraty, F., and Vieu, P. (2006), *Nonparametric Functional Data Analysis: Theory and Practice*, New York: Springer. [29,43]

Filzmoser, P. (2004), "A Multivariate Outlier Detection Method," in *Proceedings of the Seventh International Conference on Computer Data Analysis and Modeling*, Vol. 1, eds. S. Aivazian, P. Filzmoser, and Y. Kharin, Minsk: Belarusian State University, pp. 18–22. [41]

Filzmoser, P., Maronna, R., and Werner, M. (2008), "Outlier Identification in High Dimensions," *Computational Statistics & Data Analysis*, 52 (3), 1694–1711. [39,40]

Gasser, T., Hall, P., and Presnell, B. (1998), "Nonparametric Estimation of the Mode of a Distribution of Random Curves," *Journal of the Royal Statistical Society, Ser. B*, 60 (4), 681–691. [43]

Hardin, J., and Rocke, D. M. (2005), "The Distribution of Robust Distances," *Journal of Computational and Graphical Statistics*, 14 (4), 928–946. [39]

Human Mortality Database (2009), University of California, Berkeley (USA) and Max Planck Institute for Demographic Research (Germany). Available at *http://www.mortality.org/* (data downloaded on 15/4/2009). [30]

Hyde, V., Jank, W., and Shmueli, G. (2006), "Investigating Concurrency in Online Auctions Through Visualization," *The American Statistician*, 60 (3), 241–250. [30]

Hyndman, R. J. (1996), "Computing and Graphing Highest Density Regions," *The American Statistician*, 50 (2), 120–126. [31,32,36]

Hyndman, R. J., and Ullah, M. S. (2007), "Robust Forecasting of Mortality and Fertility Rates: A Functional Data Approach," *Computational Statistics & Data Analysis*, 51 (10), 4942–4956. [29,30,39]

Inselberg, A. (1985), "The Plane With Parallel Coordinates," *The Visual Computer*, 1 (2), 69–91. [32]

Jones, M. C., and Rice, J. A. (1992), "Displaying the Important Features of Large Collections of Similar Curves," *The American Statistician*, 46 (2), 140–145. [32,33]

Kargin, V., and Onatski, A. (2008), "Curve Forecasting by Functional Autoregression," *Journal of Multivariate Analysis*, 99 (10), 2508–2526. [29]

Locantore, N., Marron, J. S., Simpson, D. G., Tripoli, N., Zhang, J. T., and Cohen, K. L. (1999), "Robust Principal Component Analysis for Functional Data," *Test*, 8 (1), 1–73. [29]

López-Pintado, S., and Romo, J. (2009), "On the Concept of Depth for Functional Data," *Journal of the American Statistical Association*, 104 (486), 718–734. [33,41]

Moran, E. F., Adams, R., Bakoyéma, B., Stefano, F. T., and Boucek, B. (2006), "Human Strategies for Coping With El Niño Related Drought in Amazônia," *Climatic Change*, 77 (3–4), 343–361. [35]

Ramsay, J. O., and 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. [30]

Ramsay, J. O., and Silverman, B. W. (2002), *Applied Functional Data Analysis: Methods and Case Studies*, New York: Springer. [29]

——— (2005), *Functional Data Analysis* (2nd ed.), New York: Springer. [29,32]

Reiss, P. T., and Ogden, T. R. (2007), "Functional Principal Component Regression and Functional Partial Least Squares," *Journal of the American Statistical Association*, 102 (479), 984–996. [29]

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. [40]

Rousseeuw, P., Ruts, I., and Tukey, J. W. (1999), "The Bagplot: A Bivariate Boxplot," *The American Statistician*, 53 (4), 382–387. [31,34,35]

Scott, D. W. (1992), *Multivariate Density Estimation: Theory, Practice, and Visualization*, New York: Wiley. [31, 33]

Sood, A., James, G. M., and Tellis, G. J. (2009), "Functional Regression: A New Model for Predicting Market Penetration of New Products," *Marketing Science*, 28 (1), 36–51. [31,33]

Swayne, D. F., Cook, D., and Buja, A. (1998), "XGobi: Interactive Dynamic Data Visualization in the X Window System," *Journal of Computational and Graphical Statistics*, 7 (1), 113–130. [43]

Timmermann, A., Oberhuber, J., Bacher, A., Esch, M., Latif, M., and Roeckner, E. (1999), "Increased El Niño Frequency in a Climate Model Forced by Future Greenhouse Warming," *Nature*, 398 (6729), 694–697. [35]

Tukey, J. W. (1975), "Mathematics and the Picturing of Data," in *Proceedings of the International Congress of Mathematicians, Vol. 2, August 21–29, 1974*, ed. R. D. James, Vancouver: Canadian Mathematical Society, pp. 523–531. [31-34,43]

——— (1977), *Exploratory Data Analysis*, London: Addison-Wesley. [35]

Wegman, E. J. (1990), "Hyperdimensional Data Analysis Using Parallel Coordinates," *Journal of the American Statistical Association*, 85 (411), 664–675. [32]

Zhang, L., Marron, J. S., Shen, H., and Zhu, Z. (2007), "Singular Value Decomposition and Its Visualization," *Journal of Computational and Graphical Statistics*, 16 (4), 833–854. [30]