# Chapter 31
# Bagplots, Boxplots and Outlier Detection for Functional Data

Rob Hyndman and Han Lin Shang

**Abstract** We propose some new tools for visualizing functional data and for identifying functional outliers. The proposed tools make use of robust principal component analysis, data depth and highest density regions. We compare the proposed outlier detection methods with the existing "functional depth" method, and show that our methods have better performance on identifying outliers in French male age-specific mortality data.

## 31.1 Introduction

Although the presence of outliers has a serious effect on the modeling and forecasting of functional data, the problem has so far received little attention. In this paper, we propose the functional bagplot and a functional boxplot in order to visualize functional data and to detect any outliers present.

Recently, two papers have considered the problem of outlier detection in functional data. Hyndman & Ullah (2007) used a method based on robust principal components analysis and the integrated squared error from a linear model while Febrero et al. (2007) considered functional outlier detection using functional depth, a likelihood ratio test and smoothed bootstrapping. The method of Hyndman & Ullah involves several parameters to be specified and so is perhaps too subjective for regular use, while the method of Febrero et al. involves fewer decisions by users but is time consuming to compute and is not able to detect some types of outliers. We propose a new method that

Rob Hyndman
Department of Econometrics and Business Statistics, Monash University Clayton, VIC 3800, Australia, e-mail: Rob.Hyndman@buseco.monash.edu.au

Han Lin Shang
Department of Econometrics & Business Statistics, Monash University Clayton, VIC 3800, Australia, e-mail: Han.Shang@buseco.monash.edu.au

uses robust principal components analysis, but is simpler to apply than that of Hyndman & Ullah (2007).

Suppose we have a set of curves $\{y_i(x)\}$, $i = 1, \ldots, n$, which are realizations on the functional space $\mathcal{I}$. We are interested in visualizing these curves for large $n$ using functional equivalents of boxplots and bagplots, and we are interested in identifying outliers in the observed curves.

To illustrate the ideas, we will consider annual French male age-specific mortality rates (1899–2003) shown in Figure 31.1. These data were used by Hyndman & Ullah (2007) who obtained them from the Human Mortality Database (2007). The mortality rates are the ratio of death counts to population exposure in the relevant period of age and time. The data were first scaled using natural logarithms. The colours reflect the years of observation in "rainbow" order, with the oldest curves in red and the most recent curves in purple. There are some apparent outliers (in yellow and green) which show an unusual increase in mortality rates between ages 20 and 40. These are mainly due to the First and Second World Wars, as well as the Spanish influenza which occurred in 1918.

Before proceeding further, we need to define the notion of ordering a set of curves. López-Pintado & Romo (2007) proposed the use of "generalized band depth" to order a set of curves. The generalized band depth of a curve is the proportion (computed using Lebesgue measure) of times that the curve is entirely contained in the band defined by $J$ curves from the sample. They suggest using $J = 2$ and propose that the "median" should be defined as the curve with the highest depth. See also Ferraty & Vieu (2006, p.129) for some related discussion.


**Fig. 31.1** Functional time series plot for the French male mortality data (1899–2003).

While ordering by depth is useful in some contexts, we prefer an alternative approach to ordering obtained using a principal component decomposition of the set of observed curves. If we let

$$y_i(x) = \mu(x) + \sum_{k=1}^{n-1} z_{i,k}\phi_k(x),$$

where $\{\phi_k(x)\}$ represents the eigenfunctions, then we can use an ordering method from multivariate analysis based on the principal components scores $\{z_{i,k}\}$.

The simplest procedure is to consider only the first two scores, $\boldsymbol{z}_i = (z_{i,1}, z_{i,2})$. Then an ordering of the curves is defined using the ordering of $\{\boldsymbol{z}_i; i = 1, \ldots, n\}$. For example, bivariate depth can be used (Rousseeuw et al., 1999). Alternatively, the value of the kernel bivariate density estimate at $\boldsymbol{z}_i$ can be used to define an ordering.

There are two major advantages in ordering via the principal component scores: (1) it leads to a natural method for defining visualization methods such as functional bagplots and functional boxplots; and (2) it seems to be better able to identify outliers in real data (as we will see in the application).

Outliers will usually be more visible in the principal component space than the original (functional) space (Filzmoser et al., 2008). Thus finding outliers in the principal component scores does no worse than searching for them in the original space. Often, it is the case that the first two principal component scores suffice to convey the main modes of variation (Hall et al., 2007). We have found empirically that the first two principal component scores are adequate for outlier identification.

Because principal component decomposition is itself non-resistant to outliers, we apply a functional version of Croux & Ruiz-Gazen's (2003) robust principal component analysis which uses a projection pursuit technique. This method was described and used in Hyndman & Ullah (2007).

## 31.2 Functional bagplot

The functional bagplot is based on the bivariate bagplot of Rousseeuw et al. (1999) applied to the first two (robust) principal component scores.

The bagplot is constructed on the basis of the halfspace location depth denoted by $d(\boldsymbol{\theta}, \boldsymbol{z})$ of some point $\boldsymbol{\theta} \in R^2$ relative to the bivariate data cloud $\{\boldsymbol{z}_i; i = 1, \ldots, n\}$. The depth region $D_k$ is the set of all $\boldsymbol{\theta}$ with $d(\boldsymbol{\theta}, \boldsymbol{z}) \geq k$. Since the depth measurements are convex polygons, we have $D_{k+1} \subset D_k$. This concept is somewhat similar to the notion of a ball used in Ferraty and Vieu (2006). For a fixed center, the regions grow as the radius increases.

Thus, the data points are ranked according to their depth. The bivariate bagplot displays the median point (the deepest location), along with the

selected percentages of convex hulls. Any point beyond the highest percentage of the convex hulls is considered as an outlier. Each point in the scores bagplot corresponds to a curve in the functional bagplot. The functional bagplot also displays the median curve which is the deepest location, the 95% confidence intervals for the median, and the 50% and 95% of surrounding curves ranking by depth. Any curve beyond the 95% convex hull is flagged as a functional outlier.

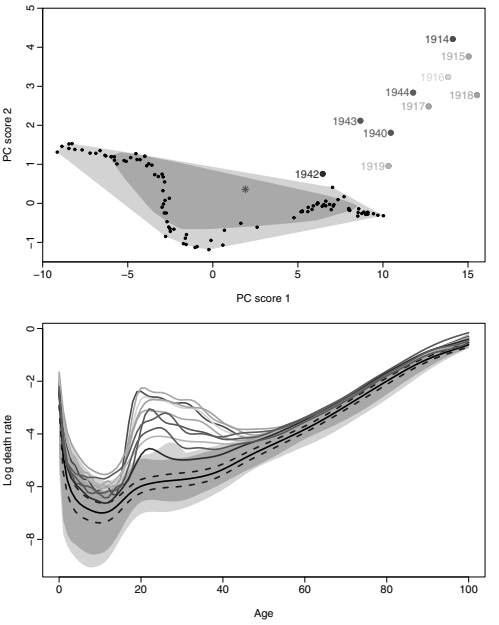


**Fig. 31.2** The scores bagplot and functional bagplot for the French male mortality data.

An example is shown in Figure 31.2 using the French male mortality data. The red asterisk marks the median of the bivariate scores and corresponds to the solid black functional observation in the right panel. The dotted blue lines give 95% confidence intervals for the median curve. In the left panel, the dark grey regions show the 50% convex hull and the light grey regions show the 95% convex hull. These correspond directly with the regions of similar shading in the functional plot on the right. Points outside these regions are identified as outliers. The different colours for these outliers enable the individual curves on the right to be matched to the scores on the left.

## 31.3 Functional HDR boxplot

The functional highest density region (HDR) boxplot is based on the bivariate HDR boxplot of Hyndman (1996) applied to the first two (robust) principal component scores.

The HDR boxplot is constructed using the Parzen-Rosenblatt bivariate kernel density estimate $\hat{f}(\boldsymbol{w}; a, b)$. For a bivariate random sample $\{\boldsymbol{z}_i; i = 1, \ldots, n\}$, drawn from a density $f$, the product kernel density estimate is defined by (Scott, 1992)

$$\hat{f}(\boldsymbol{w}; a, b) = \frac{1}{nab} \sum_{i=1}^{n} K\left(\frac{w_1 - z_{i,1}}{a}\right) K\left(\frac{w_2 - z_{i,2}}{b}\right), \tag{31.1}$$

where $\boldsymbol{w} = (w_1, w_2)'$, $K$ is a symmetric univariate kernel function such that $\int K(u)du = 1$ and $(a, b)$ is a bivariate bandwidth parameter such that $a > 0$, $b > 0$, $a \to 0$ and $b \to 0$ as $n \to \infty$. The contribution of data point $\boldsymbol{z}_i$ to the estimate at some point $\boldsymbol{w}$ depends on how distant $\boldsymbol{z}_i$ and $\boldsymbol{w}$ are.

A highest density region is defined as

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

where $f_\alpha$ is such that $\int_{R_\alpha} \hat{f}(\boldsymbol{z}; a, b)d\boldsymbol{z} = 1 - \alpha$. That is, it is the region with probability coverage $1 - \alpha$ where every point within the region has higher density estimate than every point outside the region.

The beauty of ranking by the HDR is its ability to show multimodality in the bivariate data. The HDR boxplot displays the mode, defined as $\sup_{\boldsymbol{z}} \hat{f}(\boldsymbol{z}; a, b)$, along with the 50% HDR and the 95% HDR. All points not included in the 95% HDR are shown as outliers. The functional HDR boxplot is a one-to-one mapping of the scores HDR bivariate boxplot.

An example is shown in Figure 31.3 using the French male mortality data. The black circle (left panel) marks the mode of the bivariate scores and corresponds to the solid black functional observation in the right panel. In the left panel, the dark grey regions show the 50% HDR and the light grey regions show the 95% HDR. These correspond directly with the regions of similar shading in the functional plot on the right. Points outside these regions are identified as outliers. The different colours for these outliers enable the individual curves on the right to be matched to the scores on the left.

## 31.4 Comparison

The following table presents the outlier detection results from the proposed methods along with the functional depth measure of Febrero et al. (2007).

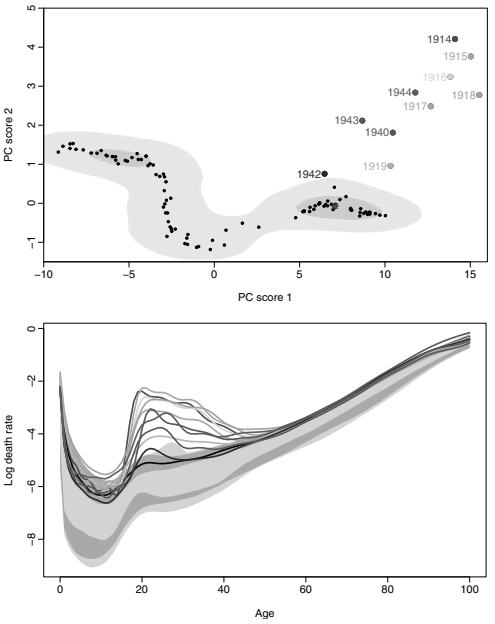


**Fig. 31.3** The scores HDR boxplot, and functional HDR boxplot for the French male mortality data.

| Method | Outlier (Year) |
|---|---|
| Functional depth | 1915 |
| Functional bagplot | 1914–1919, 1940, 1943–1945 |
| Functional HDR boxplot | 1914–1919, 1940, 1943–1945 |

**Table 31.1** Outlier detection performance between the proposed approach and the functional depth measure approach.

In this case, the functional depth measure approach performs the worst among all methods. In contrast, all of the apparent outliers in Figure 31.1 have been detected by both the functional bagplot and functional HDR boxplot methods.

Of the two new methods, we prefer the functional HDR boxplot as it also provides an additional advantage in that it can identify unusual "inliers" that full in sparse regions of the sample space.

R code for constructing the functional bagplot and HDR boxplot are available upon request from the first author.

## References


[1] Croux, C. & Ruiz-Gazen, A.: High breakdown estimators for principal components: the projection-pursuit approach revisited. Journal of Multivariate Analysis. **95**, 206–226 (2003).

[2] Febrero, M., Galeano, P. & González-Manteiga, W.: A functional analysis of NOx levels: location and scale estimation and outlier detection. Computational Statistics. **23**(3), 411–427 (2007).

[3] Ferraty, F., & Vieu, P.: Nonparametric Functional Data Analysis. Springer. (2006).

[4] Filzmoser, P., Maronna, R. & Werner, M.: Outlier identification in high dimensions. Computational Statistics & Data Analysis. **52**, 1694–1711 (2008).

[5] Hall, P.G., Lee, Y. & Park, B.: A method for projecting functional data onto a low-dimensional space. Journal of Computational and Graphical Statistics. **16**, 799–812 (2007).

[6] Human Mortality Database, University of California, Berkeley (USA), and Max Planck Institute for Demographical Research (Germany). Viewed 15/4/07, available online at <www.mortality.org> or <www.humanmortality.de>. (2007).

[7] Hyndman, R.J.: Computing and graphing highest density regions. The American Statistician. **50**(2), 120–126 (1996).

[8] Hyndman, R.J. & Ullah, Md.S.: Robust forecasting of mortality and fertility rates: a functional data approach. Computational Statistics & Data Analysis. **51**, 4942–4956 (2007).

[9] López-Pintado, S & Romo, J.: Depth-based inference for functional data. Computational Statistics & Data Analysis. **51**, 4957–4968 (2007).

[10] Rousseeuw, P., Ruts, I. & Tukey, J.: The bagplot: a bivariate boxplot. The American Statistician. **53**(4), 382–387 (1999).

[11] Scott, D. W.: Multivariate density estimation: theory, practice, and visualization. John Wiley and Sons: new York. (1992).