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.

\[ y_i(x) = \ldots \]