The rainbow Package

Han Lin Shang
Monash University

Abstract

Recent advances in computer technology have tremendously increased the usage of functional data, whose graphical representation can be infinite-dimensional curves, images or shapes. This article aims to describe four methods for visualizing functional time series using an R add-on package. These methods are demonstrated using the age-specific Australian fertility data from 1921 to 2006 and monthly sea surface temperature from January 1950 to December 2006.

Keywords: functional time series visualization, singular value decomposition plot, rainbow plot, functional boxplot, functional bagplot.

Introduction

Recent advances in computer technology have enabled researchers to collect and store high-dimensional data. When the high-dimensional data are repeatedly measured over a period of time, a time series of functions can be observed. Although one can display high-dimensional time series by adapting multivariate techniques, it is important to take smoothness of functions into account (Ramsay and Dalzell 1991). It is the smooth property of functions that separates functional time series from multivariate time series. Unlike longitudinal time series, functional time series mitigates the problem of missing values by an interpolation or smoothing technique, thus functional time series is continuous. It is the smooth and continuous properties that separate functional time series from longitudinal time series. Visualization methods help the discovery of characteristics in data that might not have been apparent in mathematical models and summary statistics. Yet this area of research has not received much attention in the literature of functional data analysis to date. However, notable exceptions are the phase-plane plot of Ramsay and Ramsey (2002), which highlights important distributional characteristics using the first and second derivatives of functional data; and the singular value decomposition (SVD) plot of Zhang, Marron, Shen, and Zhu (2007), which displays the changes in latent components in relation to the increases of the sample size or dimensionality. Another exception is the rainbow plot of Hyndman and Shang (2010), which can simultaneously provide graphical display of functional data and identify possible outliers. The aim of this article is to collect the R code that facilitate the implementation of these graphical techniques. The R code of phase-plane plot is included in the fda package (Ramsay, Wickham, Graves, and Hooker 2011), while others are included in the **rainbow** package (Shang and Hyndman 2011). In addition, this article also presents the use of animation, which can easily be embedded in all three graphical techniques in order to visualize the time-varying features of data. The outline of this article is described as follows. Visualization methods of functional time series are first reviewed.

Then, illustrated by two data sets, the visualization methods are demonstrated using the **rainbow** package. Conclusions are given in the end.

Data sets

The visualization methods are demonstrated using age-specific Australian fertility rates and monthly sea surface temperatures. The detail of these two data sets are described below. Figure 1 shows annual age-specific Australian fertility rates between ages 15 and 49 observed from 1921 to 2006. These data were obtained from the Australian Bureau of Statistics (Cat No, 3105.0.65.001, Table 38), and have been included in the **rainbow** package. The fertility rates are defined as the number of live births at 30^{th} June each year, per 1000 of the female resident population of the same age.

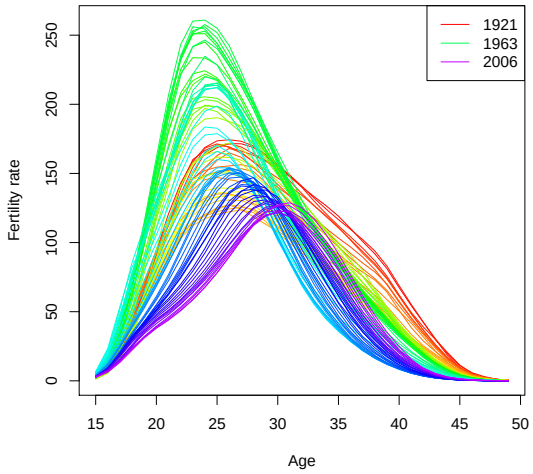


Figure 1: Smoothed Australian fertility rates between ages 15 and 49 observed from 1921 to 2006.

Although the four graphical techniques work equally well for plotting un-smoothed multivariate data, functional data ought to be smooth in nature. Therefore, the fertility rates were smoothed using a weighted median smoothing B-spline, constrained to be concave (see He and Ng 1999; Hyndman and Ullah 2007, for details).

Figure 2 shows monthly sea surface temperatures (in °C) from January 1950 to December 2006. These data were obtained from National Oceanic and Atmospheric Administration (`http://www.cpc.noaa.gov/data/indices/sstoi.indices`) and have also been included in the **rainbow** package (Shang and Hyndman 2011). These sea surface temperatures were measured by moore buoys in the "Niño region", which is defined as the area within the coordinate $0 - 10°$ South and $90 - 80°$ West.

The sea surface temperatures were smoothed using a smoothing spline with the smoothing parameter determined by generalized cross validation. Each curve represents smoothed sea surface temperatures in each year.

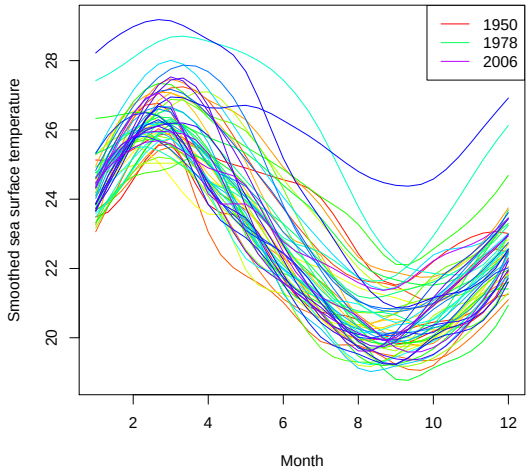


Figure 2: Smoothed monthly sea surface temperatures (in °C) from January 1950 to December 2006.

Functional time series visualization methods and their demonstrations

Rainbow plot

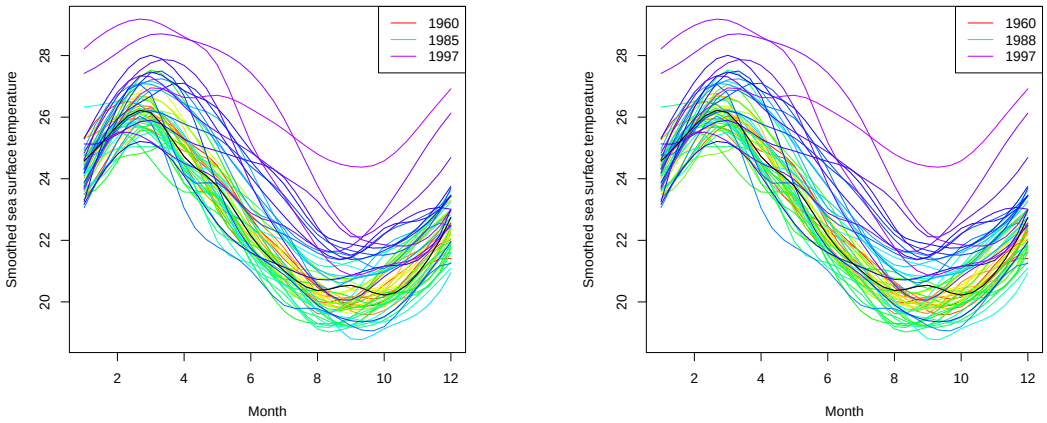
The rainbow plot is a graphical display of all the functional data, with the only additional feature being a rainbow color palette based on an ordering of the data. By default, the rainbow plot displays functional data that are naturally ordered by time. Functional data can also be ordered by halfspace location depth (Tukey 1975) and highest density regions (Hyndman 1996). The depth and density orderings lead to the developments of functional bagplot and functional HDR boxplot, described in the next subsections.

As the referees pointed out, the rainbow plot (with the default rainbow color palette) may not be suitable for readers who suffer from color blindness. To mitigate this problem, the `plot.fds` function allows users to specify their preferred color, ranging from the heat color to the terrain color. In addition to the computer-screen based RGB colors, the `plot.fds` function allows users to utilize the perceptually-based Hue-Chroma-Luminance (HCL) colors included in the **colorspace** package (Ihaka, Murrell, Hornik, and Zeileis 2011). The use of HCL colors is superior to RGB colors for readability and color separation; and it is thus preferred (Zeileis, Hornik, and Murrell 2009).

Figure 1 presents the rainbow plot of the smoothed fertility rates in Australia between ages 15 and 49 observed from 1921 to 2006. The fertility rates from the distant past years are shown in red, while the most recent years are shown in violet. The peak of fertility rates occurred around 1961, followed by a rapid decrease during the 1980s, due to the increasing use of contraceptive pills. Then, there is an increase in fertility rates at higher ages in the most recent years, which may be caused by a tendency to postpone child-bearing while pursuing careers. The rainbow plot is useful to reveal pattern changes for functional time series with a trend. It was produced using the following code.

```
# load the package used throughout this article
library("rainbow")
# plot.type = "function", curves are plotted by time
# the most recent curve is shown in purple
# the distant past cure is shown in red
plot(Australiasmoothfertility, plot.type = "functions", plotlegend = TRUE)
plot(ElNinosmooth, plot.type = "functions", plotlegend = TRUE)
```

For functional time series without a trend (e.g., Figure 2), the rainbow plot can still be used by constructing other order indexes, such as halfspace location depth and highest density regions. The colors of curves are then chosen in a rainbow color according to the ordering of depth or density.



(a) Rainbow plot with depth ordering. The median curve is shown in black.

(b) Rainbow plot with density ordering. The mode curve is shown in black.

Figure 3: Rainbow plot with depth and density orderings.

Figures 3a and 3b present the rainbow plots of sea surface temperatures ordered by halfspace location depth and highest density regions. The colors reflect the ordering and follow the order of the rainbow. The curves closest to the center of the data set are shown in red, whereas the most outlying curves are shown in violet. The curves are plotted in the order of depth and density, so the red curves are mostly obscured, but the violet curves are clearly seen even if they overlap with the majority of the data. These rainbow plots were produced using the following code.

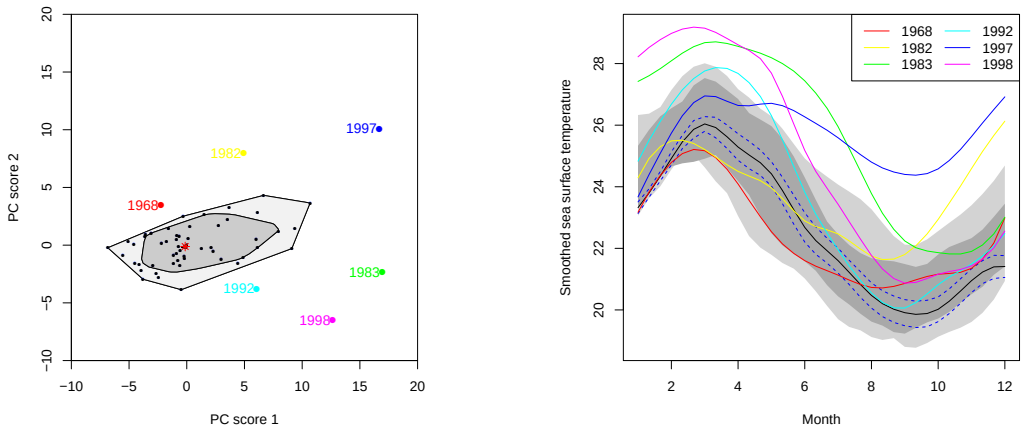
```
# plot.type="depth", curves are plotted by depth
# depth is distance between median and each curve
# median curve shown in black line is the center
plot(ElNinosmooth,plot.type="depth",plotlegend=TRUE)
# plot.type="density", curves are plotted by density
# mode shown in black line has the highest density
plot(ElNinosmooth,plot.type="density",plotlegend=TRUE)
```

Functional bagplot

Adopting from the idea of projection pursuit (Cook, Buja, Cabrera, and Hurley 1995), Hyndman and Shang (2010) use a robust functional principal component analysis to decompose functional data into the first two functional principal components and their principal component scores. As the surrogates of functional data, the bivariate principal component scores can be ordered by Tukey's halfspace location depth and plotted in a familiar two-dimensional graph.

Following Jones and Rice (1992) and Sood, James, and Tellis (2009), the functional bagplot is considered as a mapping of the bivariate bagplot (Rousseeuw, Ruts, and Tukey 1999) of the first two robust principal component scores to the functional curves. The functional bagplot displays the median curve, and the inner and outer regions. The inner region is defined as the region bounded by all curves corresponding to the points in the bivariate bag. Hence, 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 bivariate fence region. The colors of bivariate outliers are matched to the same colors of functional outliers.

Figures 4a and 4b display the bivariate and functional bagplots of the sea surface temperature data.



(a) The bivariate bagplot.

(b) The functional bagplot.

Figure 4: The bivariate and functional bagplots.

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 moderate increase, then there were abnormal increases between September 1982 and June 1983 (Timmermann, Oberhuber, Bacher, Esch, Latif, and Roeckner 1999). The sea surface temperatures during 1997-1998 were also unusual — they became extremely warm in the latter half of 1997, and stayed high for the early part of 1998.

In Figure 4a, the dark gray region shows the 50% bag, and the light gray region exhibits the customary 99% fence. These convex hulls correspond directly to the equivalent regions with similar colors and shading in the functional bagplot (in Figure 4b). Points outside these regions are defined

as outliers. The different colors for these outliers enable the functional outliers to be matched to the bivariate outliers. The red asterisk marks the Tukey median of the bivariate principal component scores, and the solid black curve shows the median curve. The dotted blue line in the functional bagplot gives 95% pointwise confidence intervals for the median curve. These bagplots were produced using the following code.

```
# plot.type = "bivariate", the bivariate principal component scores are displayed
# type = "bag" requests the bagplot
fboxplot(ElNinosmooth, plot.type="bivariate", type="bag", projmethod = "PCAproj",
         ylim=c(-10,20), xlim=c(-10,20))
# plot.type = "functional", the bivariate pc scores are matched to corresponding curves
fboxplot(ElNinosmooth, plot.type = "functional", type = "bag", projmethod = "PCAproj")
```

Functional highest density region (HDR) boxplot

The bivariate principal component scores can also be ordered by the highest density regions. The highest density regions are quantiles of two-dimensional Parzen-Rosenblatt kernel density estimate, where the bandwidths are chosen by a plug-in method (Hyndman 1996). In comparison to a depth-measure approach, the density-measure approach is able to display multimodality if it is present in the data.

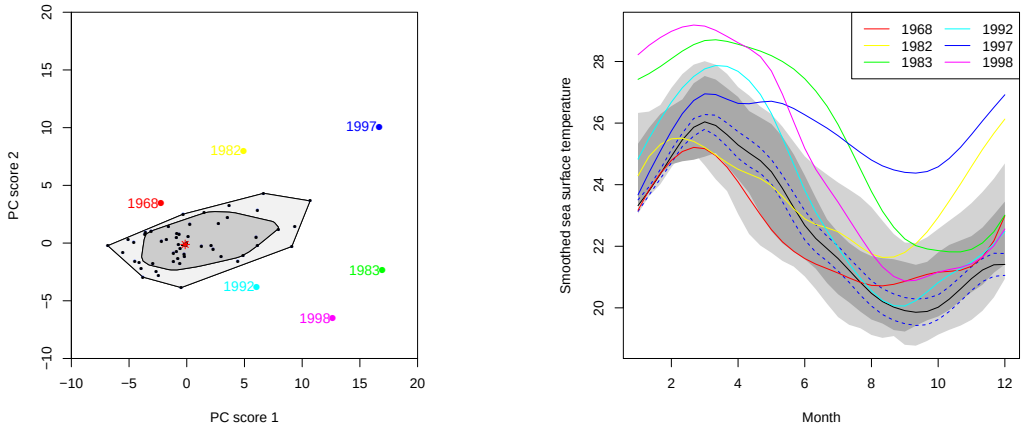
The functional HDR boxplot is a mapping of the bivariate HDR boxplot (Hyndman 1996) of the first two robust principal component scores to the functional curves. The functional HDR boxplot displays the modal curve (i.e., the curve with the highest density), and the inner and outer regions. The inner region is defined as the region bounded by all the curves corresponding to the 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 the curves corresponding to the points within the outer bivariate HDR. The colors of bivariate outliers are matched to the same colors of functional outliers.

Figures 5a and 5b display the bivariate and functional HDR boxplots of the sea surface temperature data set. As with any outlier detection methods, the coverage probability of the outer region needs to be pre-specified. If we set the coverage probability of the outer region to be 93%, then the outliers detected would match the results obtained by the bagplot. This indicates that these outliers are not only far from the median, but also have the lowest density.

In Figure 5a, the dark and light gray regions show the 50% HDR and the 93% outer HDR, respectively. These correspond directly to the equivalent regions with similar colors and shading in the functional HDR boxplot (in Figure 5b). Points outside these outer regions are identified as the outliers. The use of different colors for these outliers enables the functional outliers to match with the bivariate outliers. The red dot in the bivariate HDR boxplot marks the mode of bivariate principal component scores, and it corresponds to the solid black curve in the functional HDR boxplot.

These HDR boxplots were produced using the following code.

```
# type = "hdr" requests the HDR boxplot
# alpha requests the coverage probability of inner
# and outer HDR regions, customarily c(0.05,0.5)
fboxplot(ElNinosmooth, plot.type="bivariate", type="hdr", alpha=c(0.07,0.5),
```



(a) The bivariate HDR boxplot.

(b) The functional HDR boxplot.

Figure 5: The bivariate and functional HDR boxplots.

```
        projmethod="PCAproj", ylim=c(-10,20), xlim=c(-10,20))
fboxplot(ElNinosmooth, plot.type = "functional", type = "hdr", alpha = c(0.07,0.5),
        projmethod="PCAproj")
```

Singular value decomposition (SVD) plot

Zhang *et al.* (2007) proposed an interactive plot for visualizing patterns of functional data and multivariate data. They utilize the idea of projection pursuit by finding out only low-dimensional projections that expose interesting features of high-dimensional point cloud. As a popular projection pursuit technique, singular value decomposition (SVD) decomposes high-dimensional smoothed multivariate data into singular columns, singular rows, and singular values ordered by the amount of explained variance.

Zhang *et al.* (2007) discretize a set of functional data on a dense grid, denoted as $\boldsymbol{f}(x_i) = [f_1(x_i), \cdots, f_n(x_i)]'$, for $i = 1, \cdots, p$, where p is the number of covariates, and n is the number of curves. Let $\{\boldsymbol{r}_i; i = 1, \cdots, p\}$ and $\{\boldsymbol{c}_j; j = 1, \cdots, n\}$ be the row and column vectors of the $(n \times p)$ matrix $\boldsymbol{f}(x_i)$, respectively. The SVD of $\boldsymbol{f}(x_i)$ is defined as

$$\boldsymbol{f}(x_i) = s_1\boldsymbol{u}_1\boldsymbol{v}_1^T + s_2\boldsymbol{u}_2\boldsymbol{v}_2^T + \cdots + s_K\boldsymbol{u}_K\boldsymbol{v}_K^T,$$

where the singular columns $\boldsymbol{u}_1, \cdots, \boldsymbol{u}_K$ form K orthonormal basis functions for the column space spanned by $\{\boldsymbol{c}_j\}$; the singular rows $\boldsymbol{v}_1, \cdots, \boldsymbol{v}_K$ form K orthonormal basis functions for the row space spanned by $\{\boldsymbol{r}_i\}$; and ^{T} symbolizes vector transpose. The vectors $\{\boldsymbol{u}_k\}$ and $\{\boldsymbol{v}_k\}$ are called singular column and singular row, respectively. The scalars $s_1, \cdots, s_K$ are called singular values. The matrix $\{s_k\boldsymbol{u}_k\boldsymbol{v}_k^T; k = 1, \cdots, K\}$ is referred to as the SVD component.

The interactive plot of Zhang *et al.* (2007) captures the changes in the singular columns, as the number of curves gradually increases. Similarly, it also captures the changes in the singular rows,

as the number of covariates gradually increases. The interactive plot simultaneously presents the column and row information of a two-way matrix, to relate the matrix to the corresponding curves, to show local variation, and to highlight interactions between columns and rows of a two-way matrix.

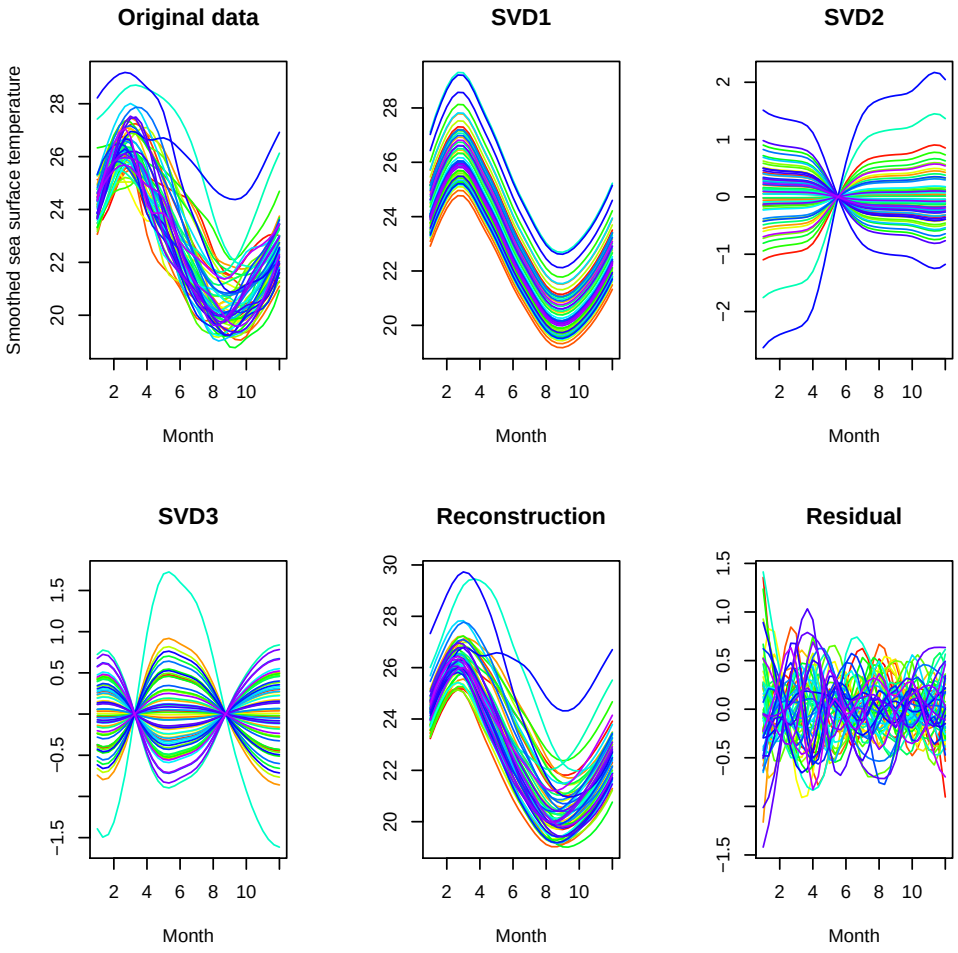


Figure 6: Singular value decomposition for the smoothed monthly sea surface temperatures (in °C) from January 1950 to December 2006.

Figure 6 shows the SVD plot of the sea surface temperature data set. The first SVD component captures the seasonal pattern, while the second and third SVD components show the contrasts of sea surface temperatures among different months. Note that the SVD2 and SVD3 are on a much smaller scale in comparison to the SVD1, because the SVD1 accounts for the most of curves' variation. The functional time series can be approximated by the summation of the first three SVD components. From the residual plot, outliers can be identified if they differ significantly from zero.

In R, the non-animated SVD plot can be produced using the following code.

```
# order represents the number of SVD components as the number of SVD components increases
# the residuals should be centered around zero plot can be suppressed by setting plot=FALSE
SVDplot(ElNinosmooth, order = 3, plot = TRUE)
```

Conclusions

This article revisited four graphical methods included in the **rainbow** package, for visualizing functional time series. These methods can be categorized into graphical techniques using projection pursuit. Each of these methods has its unique advantages for revealing the characteristics of functional time series. Some of the methods enable us to identify and analyze abnormal observations, while others can be very useful in visualizing trend. Overall, these graphical methods present a summary of functional time series, and should be considered as the first step of functional time series analysis.

Acknowledgements

Thanks to the editors and reviewers for constructive comments and suggestions that have significantly improved this article. Thanks also to Professor Rob Hyndman for helping with R code.

References

Cook D, Buja A, Cabrera J, Hurley C (1995). "Grand tour and projection pursuit." *Journal of Computational and Graphical Statistics*, **4**(3), 155–172. URL http://www.jstor.org/stable/1390844.

He X, Ng P (1999). "COBS: qualitatively constrained smoothing via linear programming." *Computational Statistics*, **14**(3), 315–337. URL http://papers.ssrn.com/sol3/papers.cfm?abstract_id=185108.

Hyndman RJ (1996). "Computing and Graphing Highest Density Regions." *The American Statistician*, **50**(2), 120–126. URL http://www.jstor.org/stable/2684423.

Hyndman RJ, Shang HL (2010). "Rainbow Plots, Bagplots, and Boxplots for Functional Data." *Journal of Computational and Graphical Statistics*, **19**(1), 29–45. URL http://pubs.amstat.org/doi/pdf/10.1198/jcgs.2009.08158.

Hyndman RJ, Ullah MS (2007). "Robust Forecasting of Mortality and Fertility Rates: A Functional Data Approach." *Computational Statistics & Data Analysis*, **51**(10), 4942–4956. URL http://portal.acm.org/citation.cfm?id=1241107.1241182.

Ihaka R, Murrell P, Hornik K, Zeileis A (2011). *colorspace: Color Space Manipulation.* R package version 1.1-0., URL http://CRAN.R-project.org/package=colorspace.

Jones MC, Rice JA (1992). "Displaying the Important Features of Large Collections of Similar Curves." *The American Statistician*, **46**(2), 140–145. URL http://www.jstor.org/stable/2684184.

Ramsay JO, Dalzell CJ (1991). "Some Tools for Functional Data Analysis (with Discussion)." *Journal of the Royal Statistical Society: Series B*, **53**(3), 539–572. URL http://www.jstor.org/stable/2345586.

Ramsay JO, Ramsey JB (2002). "Functional Data Analysis of the Dynamics of the Monthly Index of Nondurable Goods Production." *Journal of Econometrics*, **107**(1-2), 327–344. URL http://ideas.repec.org/a/eee/econom/v107y2002i1-2p327-344.html.

Ramsay JO, Wickham H, Graves S, Hooker G (2011). *fda: Functional Data Analysis*. R package version 2.2.6, URL http://CRAN.R-project.org/package=fda.

Rousseeuw PJ, Ruts I, Tukey JW (1999). "The bagplot: a bivariate boxplot." *The American Statistician*, **53**(4), 382–387. URL http://www.questia.com/googleScholar.qst?docId=5001888966.

Shang HL, Hyndman RJ (2011). *rainbow: Rainbow plots, bagplots and boxplots for functional data*. R package version 2.6, URL http://CRAN.R-project.org/package=rainbow.

Sood A, James GM, Tellis GJ (2009). "Functional Regression: A New Model for Predicting Market Penetration of New Products." *Marketing Science*, **28**(1), 36–51. URL 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. URL http://www.nature.com/nature/journal/v398/n6729/abs/398694a0.html.

Tukey JW (1975). "Mathematics and the Picturing of Data." In RD James (ed.), *Proceedings of the International Congress of Mathematicians*, volume 2, pp. 523–531. Canadian mathematical congress, Vancouver.

Zeileis A, Hornik K, Murrell P (2009). "Escaping RGBland: selecting colors for statistical graphics." *Computational Statistics and Data Analysis*, **53**(9), 3259–3270. URL http://www.sciencedirect.com/science/article/B6V8V-4VM43VH-1/2/b63f4a1a558083ab2b98f12e446d1ac7.

Zhang L, Marron JS, Shen H, Zhu Z (2007). "Singular Value Decomposition and Its Visualization." *Journal of Computational and Graphical Statistics*, **16**(4), 833–854. URL http://pubs.amstat.org/doi/abs/10.1198/106186007X256080?journalCode=jcgs.

Affiliation:

Han Lin Shang
Department of Econometrics & Business Statistics
Monash University
Melbourne, VIC, 3800,
E-mail: HanLin.Shang@monash.edu
URL: http://monashforecasting.com/index.php?title=User:Han