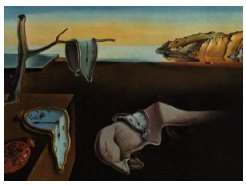

Journal of Statistical Software
June 2012, Volume 49, Issue 6. http://www.jstatsoft.org/

tourrGui: A gWidgets GUI for the Tour to Explore High-Dimensional Data Using Low-Dimensional Projections

Bei Huang
Intuit, Inc.

Dianne Cook
Iowa State University

Hadley Wickham
Rice University

Abstract

This paper describes a graphical user interface (GUI) for the **tourr** package in R. The tour is a dynamic graphical method for viewing multivariate data. The GUI allows users to interact with a tour in order to explore the data for structures like clustering, outliers, nonlinear dependence. Users can pause the tour, choose a subset of variables, color points by other variables, and switch between several different types of tours.

Keywords: dynamic graphics, interactive graphics, multivariate data visualization, visual data mining, exploratory data analysis.

1. Introduction

The tour is a method for exploring real-valued multivariate data. Users see a smooth sequence of projections of high-dimensional data. It is used to look for clusters, outliers, non-linear dependence, and to get an overview of the structures present in multivariate data. It has been defined for data exploration since Asimov (1985), and was developed further by many people, for example, Buja, Cook, Asimov, and Hurley (2005); Cook, Lee, Buja, and Wickham (2006); Wegman (1991); Wegman, Poston, and Solka (1998).

Various forms of the tour have been available in the software package **GGobi** (Swayne, Temple Lang, Buja, and Cook 2003; Temple Lang, Swayne, Wickham, and Lawrence 2011, see `http://www.ggobi.org/`). This software has been very often used for examining multivariate data but it is programmed in a way that makes it difficult to experiment with designing new types of tours. A recent R (R Development Core Team 2012) package, **tourr** (Wickham, Cook, Hofmann, and Buja 2011), reproduces all of the tours available in **GGobi**, several new tours, and makes it possible to experiment with different types of tours. How-

ever, it is a command line interface to the tour. The beauty of **GGobi** was the graphical user interface (GUI). This paper describes the development of a GUI for the **tourr** package, **tourrGui** which somewhat reproduces that in **GGobi**, but has the added capability that new tours and resulting GUIs can be developed. The GUI is programmed in the **gWidgets** package (Verzani 2007), and it is available from the Comprehensive R Archive Network at `http://CRAN.R-project.org/package=tourrGui`.

This paper describes the new tour GUI written entirely in R, **tourrGui**, with Section 2 providing details of the design, Section 3 outlining the usage, Section 4 comparing the **tourrGui** with **GGobi**, and Section 5 providing examples of its use for exploring multivariate data.

2. Design

The tour is a dynamic graphics method, for multivariate data. Thus, the GUI needs to have controls for speed of motion, for pausing and selection of variables. A slider is used to control the speed, a check box to pause and resume the tour, and a list of checkboxes for variable selection. Only real-valued variables are enabled for selection, and categorical variables are displayed in an itemized list to be used to color data in the plot.

There are also different types of tours: grand, guided, little, local. Each of these methods defines the way that projections are selected for viewing. In the grand tour the projections are chosen randomly, so that the view of the multivariate space is effectively a random walk over the space of all projections. Interpolation between random projections is done to ensure that the user sees smooth motion. The little tour restricts the choice of projections to the marginal axes, that is, variable 1 vs. variable 2, variable 1 vs. variable 3, etc. The local tour anchors the motion on a particular projection and chooses random nearby projections to tour to and back to the anchor. It allows for exploring a very local neighborhood of a projection. Finally, the guided tour searches the projection space for particularly interesting projections on the basis of a projection pursuit index. Projection pursuit is a method first decribed by Friedman and Tukey (1974). The GUI has a menu listing four indices available in this package, `holes`, `cm`, `lda_pp`, `pda_pp`. These four indices are described in Cook, Buja, and Cabrera (1993), Lee, Cook, Klinke, and Lumley (2005), Lee and Cook (2010). The pda index has an additional parameter defining the shrinkage, Λ, ranging from between 0 and 1. An additional scrollbar is used to set this value.

The tour projects data from p-dimensional space down to a lower dimensional space, d, and different types of displays might be used depending on the value of d. For $d = 2$ the common approach is to display the projections as scatterplots. For $d = 1$, histograms or density plots would be used. For $d > 2$, a choice of parallel coordinate plots or a scatterplot matrix, or for fun, Chernoff faces (Chernoff 1973) or star glyphs (Andrews 1972), might be used. In addition, there are several specialist type tours are defined, an image tour where two dimensions of geographic or spatial variables are combined with projections of multiple variables measured at each spatial location. A stereo tour can be used to show 3D renderings of projections when $d = 3$. Typically, different choices of d dictate distinct ways of examining the multivariate data, and it is not common to switch from one d to another. For this reason, separate GUIs are created for each choice of d. Figure 2 shows the GUI for $d = 2$. The different GUIs have some unique features. For example, the 2D tour GUI, has options for placement of the axes. The axes are used for observing how each variable contributes to the current projection.

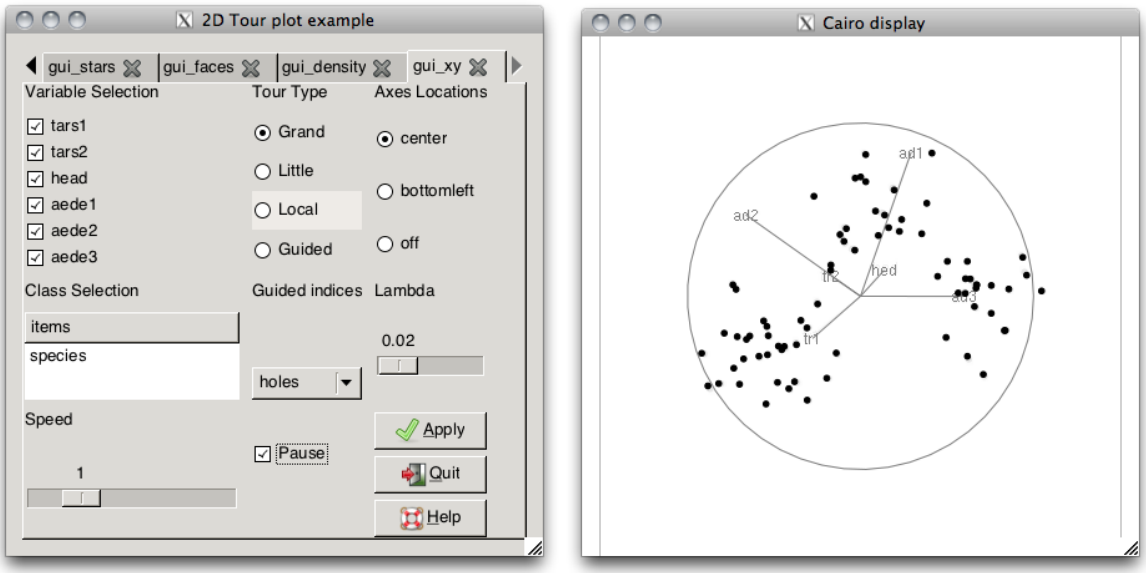

Figure 1: The single GUI for access to all tour types, shown with the xy tour tab open on the flea beetles data. Tabs at the top allow switching from one tour type to another.

In addition, three more controls are provided: `Quit`, to close the GUI; `Apply` forces changes to the parameters to take place if they didn't happen automatically; and `Help` gives some information about tours.

It should be noted that the **tourr** package has a lot more flexibility than has been exposed in the GUI. The purpose of the GUI is to make it easy to do the common things, and thus we have deliberately chosen to make a simple interface with not too many choices for the user.

3. Usage

The GUI is started by opening a single GUI accessing all tour types `gui_tour()` (Figure 1) or by selecting a particular type of tour to run, and the choice of data set. The GUI provides access to a variety of tour controls, such as variable selection, type of target base selection, `pause`, speed, and axes location.

The function `gui_xy()` starts a 2D tour, as shown in Figures 2 and 3. The function `gui_density()` starts a 1D tour, as shown in Figure 4. This GUI allows the display type to be changed from a density to a histogram. The functions `gui_pcp()` (Figure 5), `gui_scatmat()` (Figure 6), `gui_faces()`, `gui_stars()` (Figure 8) will open GUIs to the $d > 2$D tours. A 3D tour can be started using `gui_stereo()`, and the image tour is started using `gui_image()`.

Variable selection

Each of the GUIs has a checkbox list to select variables to include in the tour. By default, all real-valued variables in the data are included. Clicking on a checkbox will remove that variable from the tour, and clicking again adds it. The user will need to click on the `Apply` button to have these changes applied to the current tour.

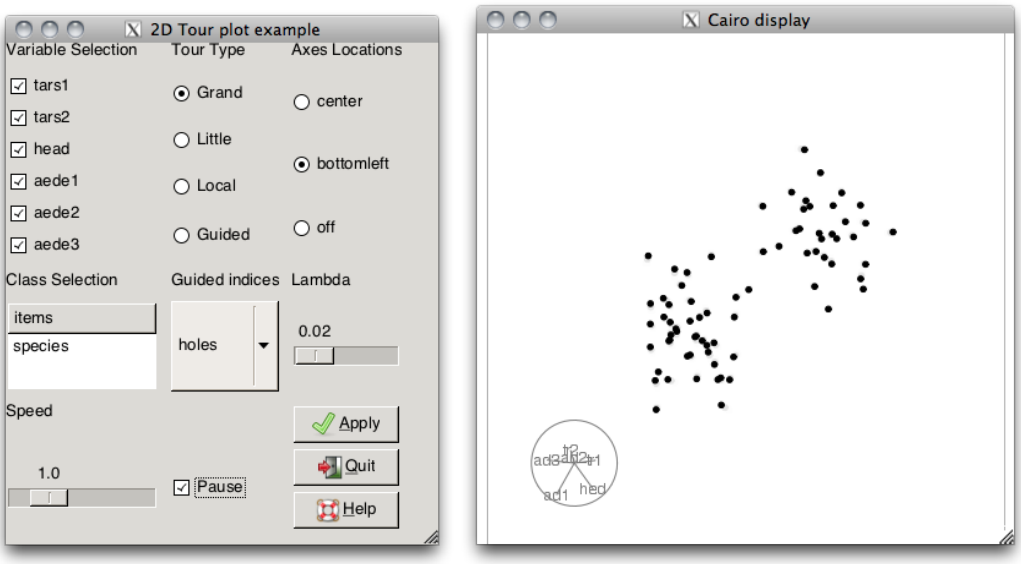

Figure 2: The GUI for the 2D tour, $d = 2$, on the default data, flea beetles. All 6 variables in the data are selected. The grand tour method is being used, but it is paused, and axes are placed in the lower left of the plot. Two circular clusters are visible in the projection, and it is a combination of most of the variables.

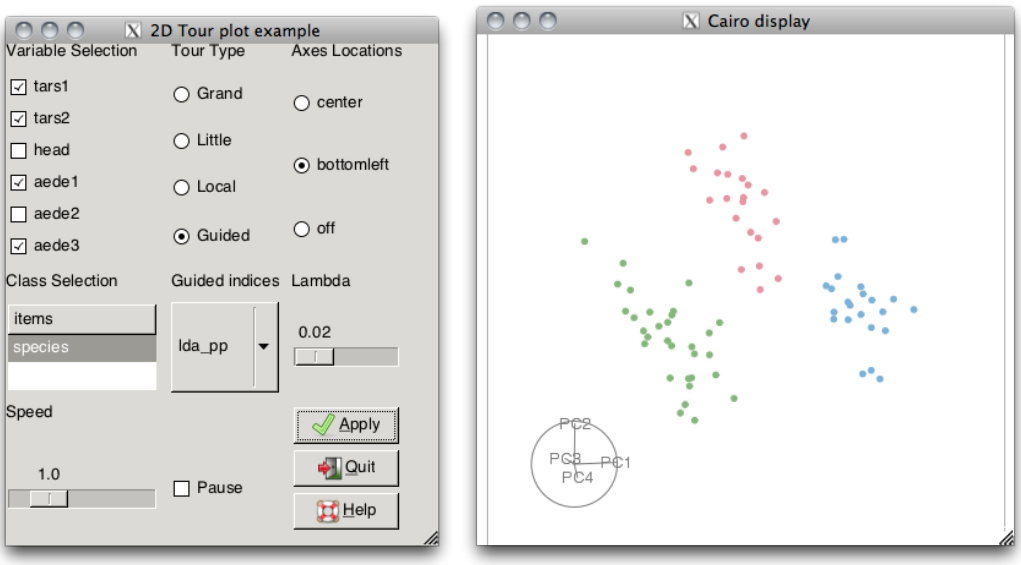

Figure 3: The GUI for the 2D tour, $d = 2$, on the default data, flea beetles. A subset of 4 variables are chosen, with points colored by the categorical variable species. A guided tour using the lda index is being viewed, and it is stopped at a projection which reveals three clusters corresponding to the three different species.

Checkboxes are a deliberate choice. The tour is best used when the data dimension is reasonably small around 20 or less variables, in which case the simplicity of toggling variables in and out suggests the checkbox interface.

Color by categorical variable

Categorical variables are not included in the tour. Instead they are available to use to color the points in the plot, and for using with the guided tour. The Class Selection interface is a list of the categorical variables in the data set, from which the user can select one. For some displays, notably the 1D tour, the observations are not colored, but the variable will be used if the guided tour with one of the class indexes, `lda_pp` or `pda_pp`, is used.

Speed, pause

All of the GUIs have a scrollbar to control the speed. Sliding to the left slows it, forcing it to take smaller steps between projections, and to the right speeds it up, effectively taking larger steps. The `Pause` checkbox allows the tour to be paused and resumed. Changing these parameters forces changes immediately in the plot, unlike other GUI items that require the `Apply` button to the clicked before the change is realized.

Target basis selection

Four different types of target basis selection methods are available through radio buttons: Grand, Little, Local and Guided. Only one can be chosen at any time. If a guided tour is chosen, the user also needs to select an index to use. By default it is the `holes` index, which optimizes the choice of projections for those which have few data points in the middle, for example a donut shape. The menu under the tour type radio buttons allows the user to choose indices. If the user chooses either the `lda_pp` or `pda_pp` index a categorical variable must be chosen also. These indices find projections of the data where the classes are most distinct from each other. The `cm` index is good for finding outliers in the data.

Axes location

The 1D and 2D GUIs have some type of axis display. The axes are displaying the projection coefficients. For each of these GUIs there are choices for how to display the axes. Axes are useful for interpreting structure visible in a particular projection. For example, in Figure 4 the axes show that major contributions to the projection come from four variables, and the structure in the projection is bimodality. This indicates that these four variables are responsible for some clustering of the data in the high-dimensional space.

Apply, help, and quit

These buttons enable some general control. The `Apply` button applies changes made to tour parameters using the GUI to the current tour. `Quit` closes the GUI, and `Help` provides some information about the tour, in a popup text window.

Choosing cases, icon displays

Two of the display methods use icons to represent multivariate data: faces and stars. In these displays every row of the data matrix is coded into an icon. This can be unwieldy for large data sets, so the user has the ability to view a sample of cases.

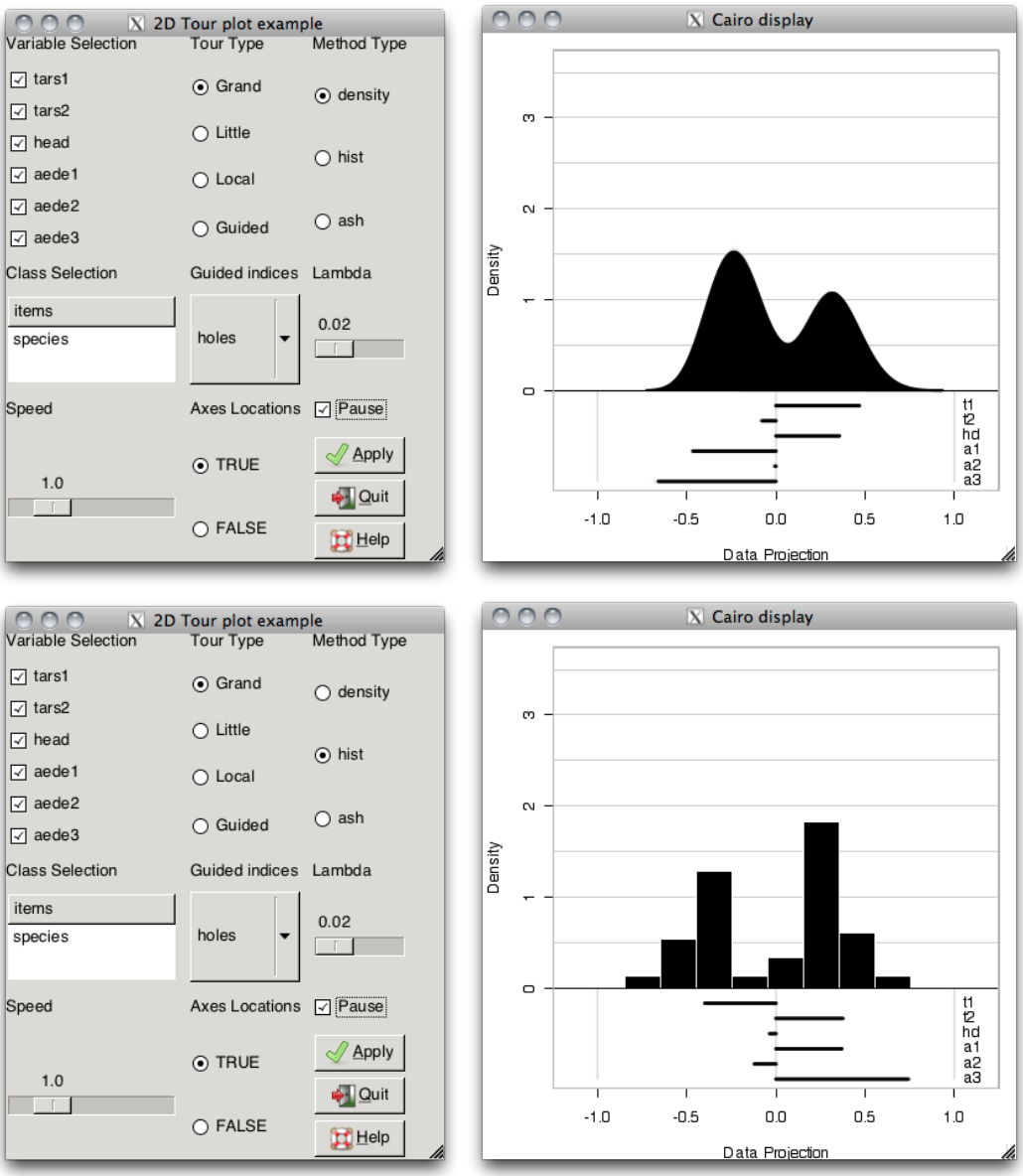

Figure 4: The GUI for the 1D tour, $d = 1$, on the default data, flea beetles. All 6 variables are included, and the current projection is displayed as a density (top). The current projection is a contrast of the variables t1, hd against a1, a3, which reveals bimodality in the data. The bottom row shows the projection as a histogram. This projection also reveals a bimodal distribution in a combination on the variables t1, t2, a1 and a3.

Miscellaneous

Each of the GUIs for $d > 2$-dimensional projections and the ones for parallel coordinate plot and scatterplot matrix displays also have GUI items for choosing d, ranging from 2 to p.

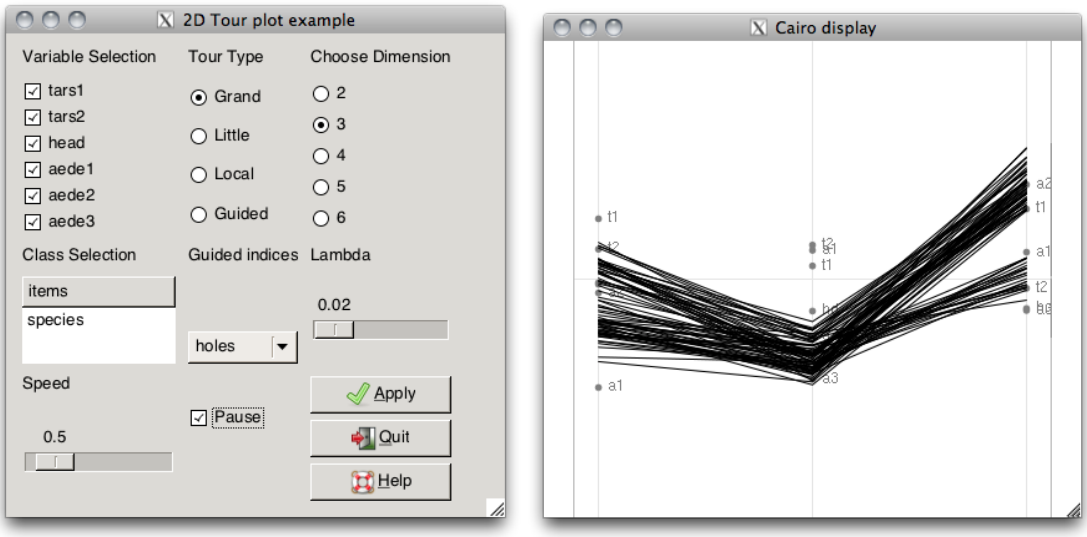

Figure 5: The GUI for the $d > 2$-dimensions using the parallel coordinate plot display. All 6 variables are included in a grand tour, and $d = 3$ is chosen for the display. Each axis of the parallel coordinate plot shows one of the three data projections, along with a representation of the coefficients of the projection, as labelled dots.

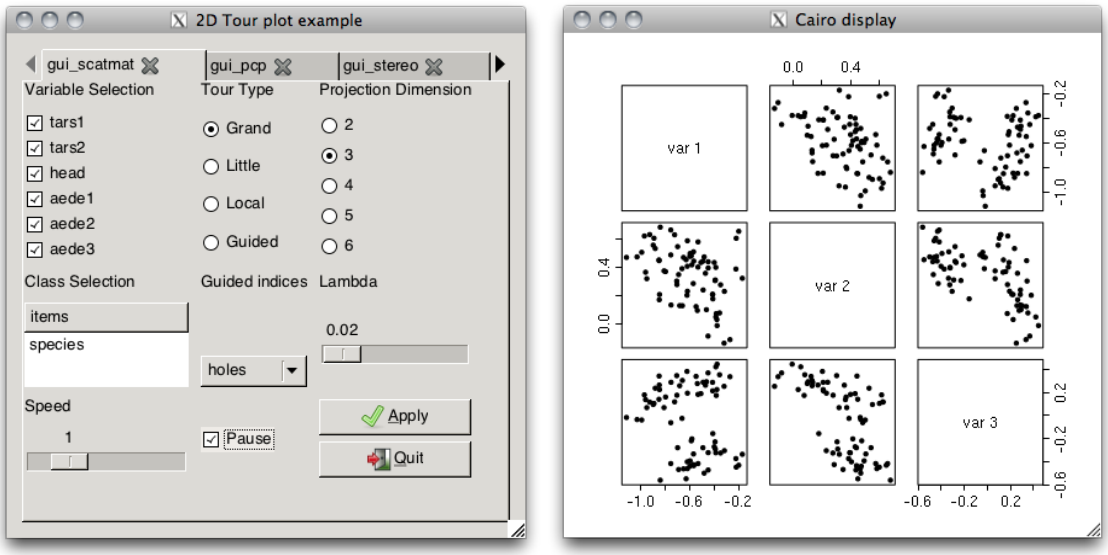

Figure 6: The GUI for the $d > 2$-dimensions using the scatterplot matrix display. All 6 variables are included in a grand tour, and $d = 3$ is chosen for the display.

4. tourrGui vs. GGobi

The big difference between **GGobi** and the **tourrGui** is that **GGobi** is programmed in C, with a main event loop listening to user events on the screen that makes truly interactive graphics possible. The **tourrGui** is programed within R, where there are no capabilities for event driven graphics. The GUI can capture screen events such as clicking on a toggle button or dragging

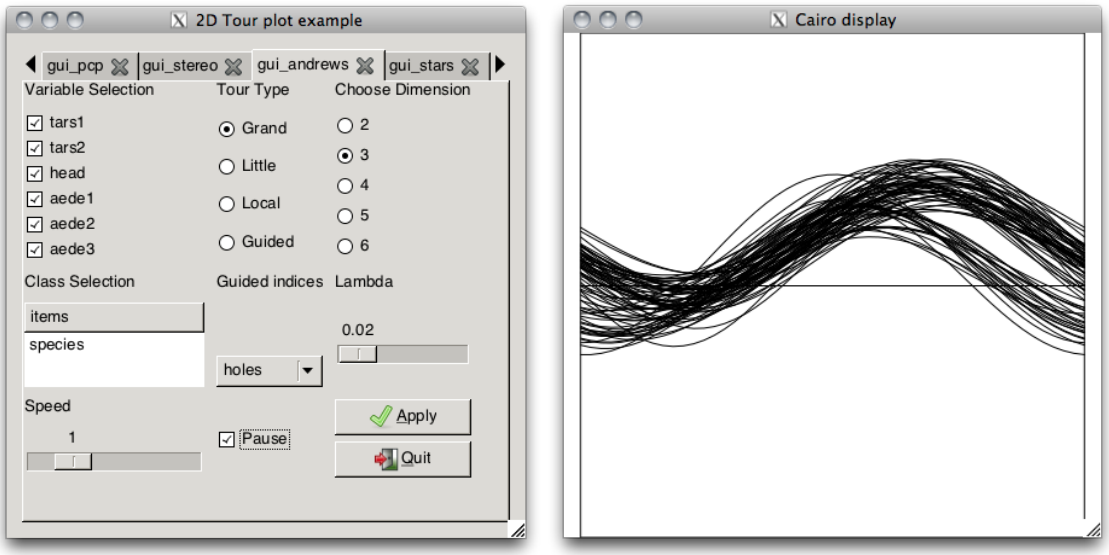

Figure 7: The GUI for the $d > 2$-dimensions using the Andrews curves display. All 6 variables are included in a grand tour, and $d = 3$ is chosen for the display.

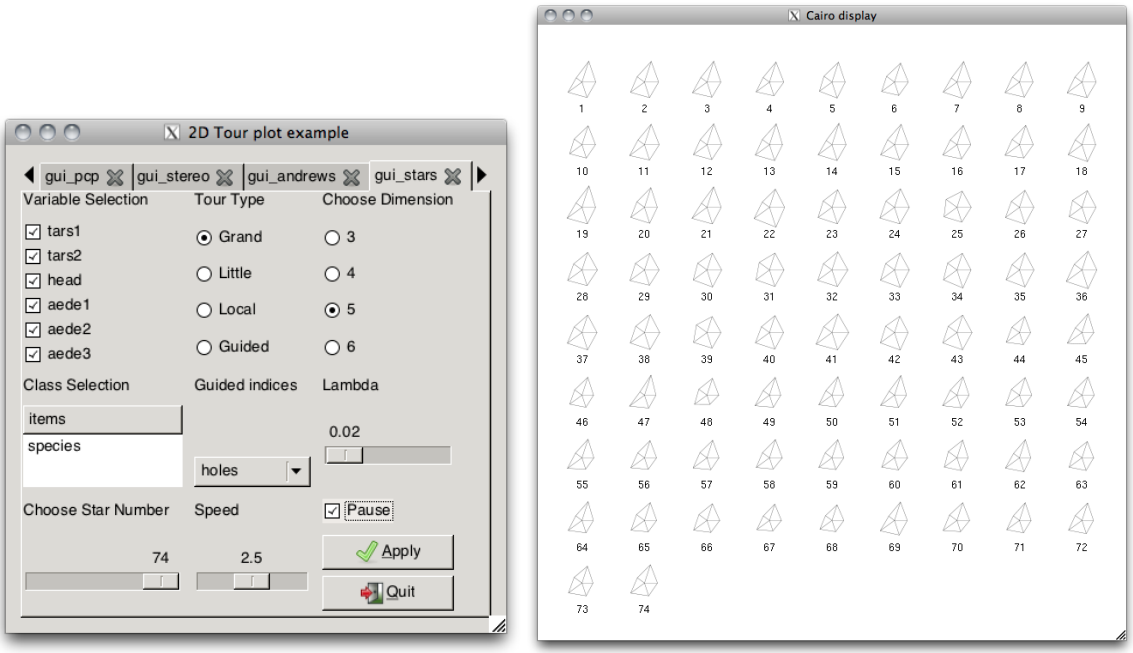

Figure 8: The GUI for the $d > 2$-dimensions using the stars display. All 6 variables are included in a grand tour, and $d = 5$ is chosen for the display. All 74 cases in the data are displayed as icons.

a scrollbar, but the user cannot brush points or bars in the plot. To facilitate some interaction on the GUI categorical variables are listed so that the user can select these to color the points by these variables. The user might also notice that some GUI actions will restart the tour,

for example choosing a new basis generation method, or a different subset of variables. It is essentially calling the tour afresh, to make these changes in tour type.

5. Examples

The default data set for the package is the flea beetles data (Lubischew 1962). This data set has been used in the plots shown in Figures 2–8. There are 6 variables which describe the physical characteristics of the beetles and a categorical variable containing the species information. The three species of beetles correspond to three clusters in the data in the 6D data space. The clusters are effectively elliptical and homogeneous, and well-separated. There is no more multivariate structure in this data – it is very clean and simple.

A data set with richer multivariate data is the music data, which was collected by the author from her music CDs. This data contains five variables which measure different aspects of the sound in the first 40 seconds of the music clip. Figure 9 shows several projections of this data. The projection in the top right plot, two outliers (one at lower right, and the other nearly hidden by the variable axes), nonlinear dependence, and a hint of clustering, are visible.The non-linear dependence is seen from the curvature in the trend of points from top right to lower left. And the hint of clustering, because there is one long cluster of about 12 points separated slightly from the main body of points - which one would find is a distinct cluster when more projections are examined.

If we examine the data in relation to the two categorical variables included with the data, artist and type, we learn that the cluster structure is related to these, and that the categories of these variables partition the data. There are 3 types of music measured, rock, classical and new wave, from 7 artists, Beatles, Abba, Eels, Vivaldi, Mozart, Beethoven, and Enya. The bottom two plots of Figure 9 show guided tours using the `lda_pp` index separately for each of these groups. (Note that these indices force the change to principal component space, which is why the axes say PC1, . . . , PC6.) The two main types of music, rock and classical (red and blue) are essentially separated in the data space, with new wave (green) falling right in the middle (bottom left plot). There are also some differences between the artists (bottom right plot), most notably the separated red cluster, which corresponds to the Abba music clips. This cluster is the same as that noticed in the projection shown in the top right plot. The data is also available on the **GGobi** web site, see Swayne *et al.* (2003), on the publications page.

6. Future work

The **tourrGui** package represents a start, demonstrating that it is possible to build a GUI to provide simple control of different types of tours in the **tourr** package. Some immediately new directions are to include backtracking and saving of projections. Backtracking will allow users to navigate forward and backwards in the sequence of projections displayed in any tour. Saving a projection is useful when some interesting structure is found in the data, and the user would like to save this particular combination of the data for further analysis. Missing from this application are interactive tools that were extremely useful in **GGobi**, brushing and identification of observations. These tools are not available with the current graphics devices in R. New work is being conducted by two different groups (`http://www.rforge.net/`

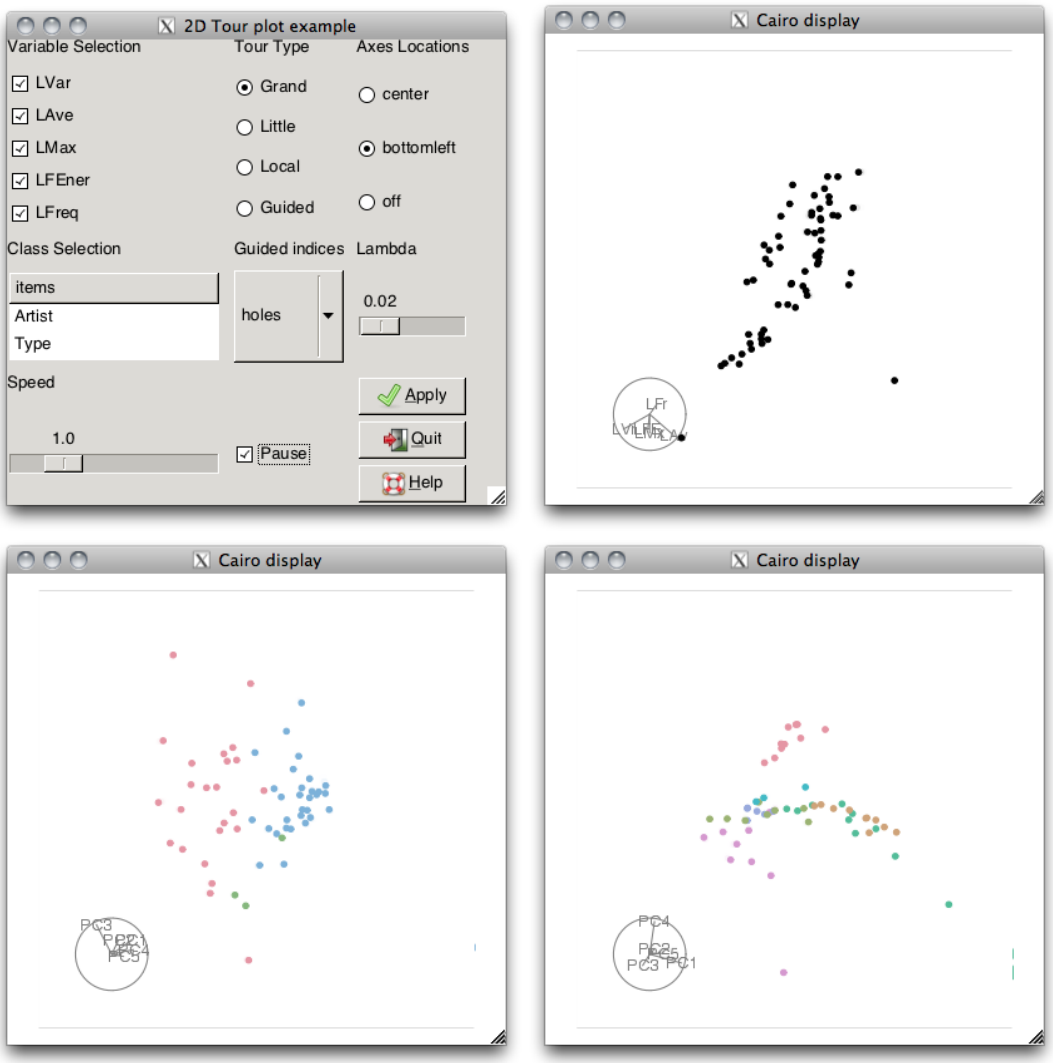

Figure 9: Exploring the music data with a 2D tour. Three different projections of the data. At top right this projection shows two outliers, nonlinear dependence and two clusters. The bottom left projection uses the music type in a guided tour to reveal that the two types of music (rock and classical) are different in these five variables. The bottom right plot shows a projection from a guided tour on the six different artists. The red cluster corresponds to Abba clips.

`Acinonyx/`, `https://github.com/ggobi/`) that may provide an interactive graphics canvas for R, and then these tools will be combined with the **tourrGui**. These are exciting new advances for data analysis because it will allow scripting of interactive and dynamic graphics in R and a tight coupling of graphics and modeling.

Acknowledgments

This work has been partly supported by the National Science Foundation grant DMS0706949.

References

Andrews DF (1972). "Plots of High-Dimensional Data." *Biometrics*, **28**, 125–136.

Asimov D (1985). "The Grand Tour: A Tool for Viewing Multidimensional Data." *SIAM Journal of Scientific and Statistical Computing*, **6**(1), 128–143.

Buja A, Cook D, Asimov D, Hurley C (2005). "Computational Methods for High-Dimensional Rotations in Data Visualization." In CR Rao, EJ Wegmanand, JL Solka (eds.), *Handbook of Statistics: Data Mining and Visualization*, pp. 391–414. Elsevier, Amsterdam.

Chernoff H (1973). "The Use of Faces to Represent Points in k-Dimensional Space Graphically." *Journal of the American Statistical Association*, **68**, 361–368.

Cook D, Buja A, Cabrera J (1993). "Projection Pursuit Indexes Based on Orthonormal Function Expansions." *Journal of Computational and Graphical Statistics*, **2**(3), 225–250.

Cook D, Lee EK, Buja A, Wickham H (2006). "Grand Tours, Projection Pursuit Guided Tours and Manual Controls." In CH Chen, W Härdle, A Unwin (eds.), *Handbook of Data Visualization*. Springer-Verlag, Berlin.

Friedman JH, Tukey JW (1974). "A Projection Pursuit Algorithm for Exploratory Data Analysis." *IEEE Transactions on Computing C*, **23**, 881–889.

Lee EK, Cook D (2010). "A Projection Pursuit Index for Large p Small n Data." *Statistics and Computing*, **20**, 381–392.

Lee EK, Cook D, Klinke S, Lumley T (2005). "Projection Pursuit for Exploratory Supervised Classification." *Journal of Computational and Graphical Statistics*, **14**(4), 831–846.

Lubischew AA (1962). "On the Use of Discriminant Functions in Taxonomy." *Biometrics*, **18**, 455–477.

R Development Core Team (2012). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0, URL http://www.R-project.org/.

Swayne DF, Temple Lang D, Buja A, Cook D (2003). "**GGobi**: Evolving from **XGobi** into an Extensible Framework for Interactive Data Visualization." *Computational Statistics & Data Analysis*, **43**, 423–444.

Temple Lang D, Swayne D, Wickham H, Lawrence M (2011). "**rggobi**: An Interface between R and **GGobi**." R package version 2.1.17, URL http://CRAN.R-project.org/package=rggobi.

Verzani J (2007). "An Introduction to **gWidgets**." *R News*, **7**(3), 26–33. URL http://CRAN.R-project.org/doc/Rnews/.

Wegman EJ (1991). "The Grand Tour in k-Dimensions." *Technical Report 68*, Center for Computational Statistics, George Mason University.

Wegman EJ, Poston WL, Solka JL (1998). "Image Grand Tour." In *Automatic Target Recognition VIII – Proceedings of SPIE, 3371*, pp. 286–294. SPIE, Bellingham.

Wickham H, Cook D, Hofmann H, Buja A (2011). "**tourr**: An R Package for Exploring Multivariate Data with Projections." *Journal of Statistical Software*, **40**(2), 1–18. URL http://www.jstatsoft.org/v40/i02/.

Affiliation:

Bei Huang
Intuit, Inc.
Mountain View, CA, United States of America
E-mail: beihuangisu@gmail.com

Dianne Cook
Department of Statistics
Iowa State University
Ames, IA, United States of America
E-mail: dicook@iastate.edu
URL: http://www.public.iastate.edu/~dicook/

Hadley Wickham
Department of Statistics
Rice University
Houston, TX, United States of America
E-mail: hadley@rice.edu
URL: http://had.co.nz/

Journal of Statistical Software http://www.jstatsoft.org/
published by the American Statistical Association http://www.amstat.org/
Volume 49, Issue 6 *Submitted:* 2011-01-20
June 2012 *Accepted:* 2012-04-16