INTERACTIVE RANDOM FORESTS PLOTS

by

Anna T. Quach

A report submitted in partial fulfillment
of the requirements for the degree

of

MASTER OF SCIENCE

in

Statistics

Approved:

Dr. Adele Cutler
Major Professor

Dr. Chris Corcoran
Committee Member

Dr. Heidi Wengreen
Committee Member

UTAH STATE UNIVERSITY
Logan, Utah

2012

ACKNOWLEDGMENTS

I would like to thank my advisor, Dr. Adele Cutler, for her time, knowledge, and enthusiasm. This work was partially supported by NIH grant 1R15AG037392-01.

Anna T. Quach

I dedicate this for my family.

ABSTRACT

Interactive Random Forests Plots

by

Anna T. Quach, Master of Science

Utah State University, 2012

Major Professor: Dr. Adele Cutler
Department: Mathematics and Statistics

Random Forests is a useful data mining tool that is quite popular in finding variable importance. However, many people don't make use of the Random Forests results in interactive graphs. Partly, this is because software packages that can do interactive graphs can't handle large data sets and those that use Random Forests have large data sets or many variables. A new software package in R, known as iPlots eXtreme, that is still in development makes it simple to explore large data sets interactively. I have created a function, called irfplot (interactive random forests plot) that specifically uses Random Forests to produce interactive graphs that are more informative than using raw values. I will use the interactive Random Forests plot that I've created to explore the nutrition data set from the Cache County Memory Study.

Contents

List of Figures

1 Introduction and Literature Review

In unsupervised learning, a useful graphical method used to observe patterns, clusters, outliers, and correlations of high-dimensional continuous variables is the parallel coordinate plot. Parallel coordinate plots were first introduced as a device for computational geometry by Inselberg (Inselberg, 2009). However, the application of parallel coordinate plots used as a multivariate data analysis tool in statistics was first done by Wegman (Wegman, 1990). In Wegman's paper, titled "Hyperdimensional Data Analysis Using Parallel Coordinates," he states that the scatter diagram is a fundamental tool and allows the eye to detect structures in data, such as linear and nonlinear features, outliers, and clusters. However, we cannot see the relationship among all the covariates in a compact way with scatter diagrams because they do not generalize beyond the third dimension. Parallel coordinate plots are a more desirable visualization tool because we can see many variables using less space all in one plot. What makes parallel coordinate plots even more appealing is the free software that can produce parallel coordinate plots in a way such that we have additional tools such as brushing and highlighting that are especially useful for exploring and understanding data.

The data set that I will use to illustrate parallel coordinate plots is from the Cache Valley Memory and Aging Study (Wengreen et al., 2007). The data set is made up of people 65 and over and the subset of people that will be considered only consists of those who have completed the food frequency questionnaire at baseline. Those with calorie intake below 500 calories or above 5000 calories, those diagnosed with prevalent dementia, those with missing baseline memory score, those that are protocol violators with an unknown cognitive status, and an individual consuming more than 40 servings of fruit are excluded from the final data set. Those younger than the median age (74.1) are explored because of the difference in dietary pattern are found in the younger people in Figures five to eight. The main focus of interest will be to find any dietary pattern or specific food groups that are beneficial or detrimental for those with normal cognition, cognitive impairment, or dementia of any kind. This is an example of a classification problem in which the response variable is dementia status. Covariates such as age and cognitive score that are of great impact to cognitive status and possibly diet will be explored along with the 40 energy adjusted food groups (Willett, 1989). The parallel coordinate plot is a better exploratory tool in that it can handle a larger number of variables than pairwise scatter plots. We will be able to find the variables that have more variability while also observing clustering if any.

Although, parallel coordinate plots are good tools that can deal with high dimensional data, they are more useful when the plots are used interactively. Interactive graphs are important in applied statistics because getting to know the data is part of the first step before applying any statistical methods. Some software packages that can be used to do interactive graphs and parallel coordinate plots are rggobi (Cook and Swayne, 2007) and iplots (Urbanek et al., 2005). However, rggobi cannot handle large data sets. Iplots can handle large data sets, but the response time can be quite slow. For data mining R users, to produce interactive graphs within R is of great convenience, rather than having to do any data manipulation within R then having to export the data set each time a plot is required. The fastest known graphics device in R that can handle large data sets is iPlots eXtreme (Urbanek, 2011), also known as Acinonyx, which is under development, but still useful.

One approach for visualizing classification data is to do parallel coordinate plots and color the line segments according to the class variable. It may then become obvious which variables are related to the class variable (dementia status in our data) and which are not. However, for a data set with lots of predictors, patterns may be difficult to discover using visualization alone. Random Forests (RF) is a popular data mining algorithm for classification (Izenman, 2008; Cutler et al., 2007; Cutler and Stevens, 2006; Breiman, 2001; Cutler et al., 2012, 2008). It is usually viewed as a ”black box” classifier that has excellent predictive power but is not so easy to understand or interpret. Parallel coordinate plots can help understand the RF results and help us look inside the black box. Random Forests is especially useful because it can be used on large data sets with many covariates. Random Forests can handle both categorical and continuous predictor variables. Graphics appropriate for Random Forests can be obtained using the Java package RAFT (Breiman and Cutler, 2005), but it is not implemented in R and is not able to read in or handle large data sets. Including Random Forests results such as the proximity data and measures of variable importance in parallel coordinate plots and multidimensional scaling plots gives information about the top variables that are judged to be the most important while discovering patterns within those important variables. Therefore using the information that Random Forests produces, such as the importance values and the proximity data, is advantageous in looking for dietary patterns and seeing if they are associated with cognitive status. Parallel coordinate and multidimensional scaling interactive graphs are not too common in the nutrition field they can be useful tools with the large number of food variables that all need to be taken into consideration. My goal is to create a function or package in R that can be used to explore parallel coordinate plots and multidimensional scaling plots interactively using Random Forests data.

2 Visualization Background

This section presents the general background of parallel coordinate plots, Random Forests, multidimensional scaling plots, interactive graphs, and a software package in R called iPlots eXtreme. The interpretation for parallel coordinate plots, Random Forests variable importance and proximities are mentioned along with plots to help with interpretation and the advantages of using the methods listed above are discussed.

2.1 Parallel Coordinate Plots

2.1.1 General Background

Statistical methods such as cluster analysis and principal component analysis are two of the approaches in finding or discovering dietary patterns (Lattin et al., 2003). To visualize the clustering itself is a powerful tool and can suggest that there is some sort of pattern. However, to be able to visualize the two techniques and find distinct clusters is not that simple especially in the case where we have many different diets which overlap each other.

To produce a parallel coordinate plot, a vertical axis is used for each variable. These axes are parallel to each other. An observation is represented by a point on each axis. These points are joined by a polyline, that is, a piecewise linear curve that connects the points. Each polyline represents one observation. If one of the observation's values is missing then that observation will not appear on the parallel coordinate plot. Figure 1 illustrates the basic idea of a parallel coordinate plot, where the variables say, $X_1, X_2, \ldots, X_p$ are plotted in parallel to each other. Two observations are shown in Figure 1, namely, $(x_{k1}, x_{k2}, \ldots, x_{kp})$ and $(x_{l1}, x_{l2}, \ldots, x_{lp})$. The predictor variables used for the parallel coordinate plots should be continuous variables only. Categorical variables could be used, but the variables' plotted values would not be meaningful. For example, if we had a two level education categorical variable with values 0,1, the ploylines would only go to those two points and therefore it would be difficult to see patterns. However, including such variables can help see patterns in the other variables by allowing us to highlight the value 0 or 1. Highlighting will be discussed in Section 2.4.

Parallel coordinate plots are quite intuitive when it comes to interpretation. To investigate a linear relationship between two variables, we would observe whether or not most of the lines do or do not intersect between the parallel coordinate axes. If there's an intersection, then the two variables are negatively correlated. The two variables will be positively

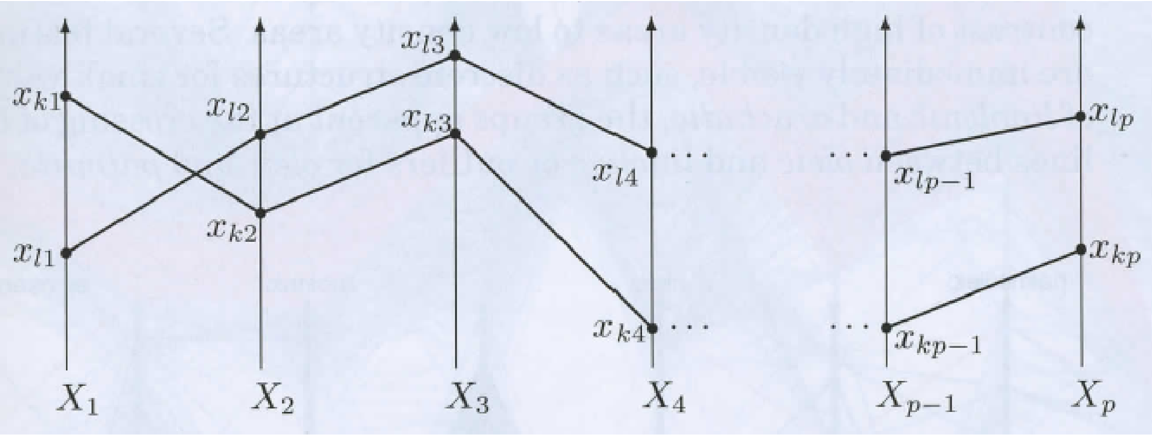

Figure 1: From Theus and Urbanek book (Theus and Urbanek, 2008)

correlated if the the lines hardly intersect at all. Figure 2 from Wegman's paper (Wegman, 1990) illustrates what the parallel coordinate plot would look like with correlations going from a highly positive correlation ($\rho = 1$) to a highly negative correlation ($\rho = -1$).

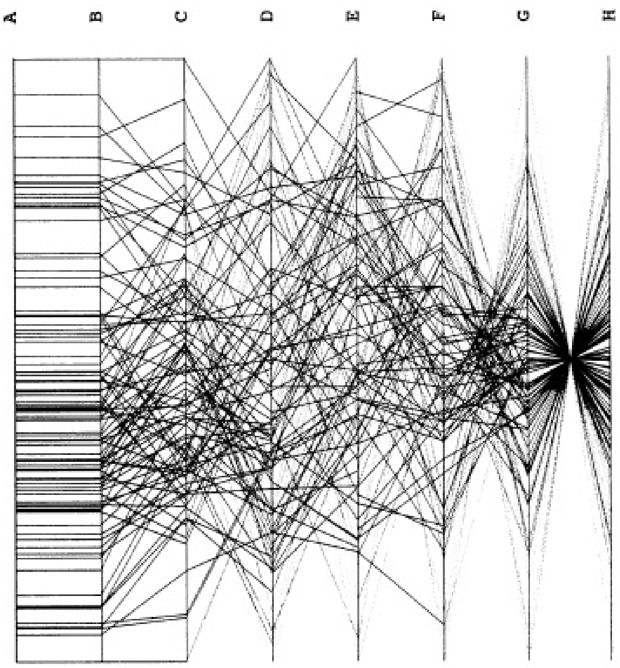

Figure 2: From Wegman's 1990 paper (Wegman, 1990) illustrating parallel coordinate plot of six-dimensional data with correlation equal to 1, .8, .2, 0, -.2, -.8, and -1.

We can also identify clusters in parallel coordinate plots. Any axis or axes with some separation between the polylines suggests clusters and it's easy to see if the clustering propagates through other dimensions by highlighting the cluster. Highlighting will be discussed in Section 2.4. Another useful interpretation is that each axis shows the distribution of each variable. With the adjustment of density (Section 2.4) of the lines, it can be visualized if a variable is normally distributed, that is, there's a tendency of the polylines or points to be in the center of the vertical axis, which will have darker shaded areas in the center when the opacity is decreased, or skewed left or right by the overplotting observed on each axis. Figure 3 illustrates the data analysis features described above, that is, the ability to diagnose one-dimensional features such as marginal densities (the axis labeled 1), two-dimensional features revealing correlations and nonlinear structures (axis 2 and 3), three-dimensional features revealing clustering (axis 3, 4, and 5), and a five-dimensional mode (axis 1 to 5).

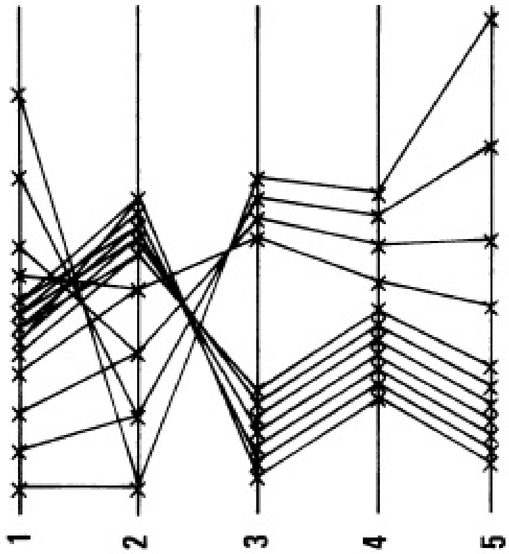

Figure 3: From Wegman's 1990 paper (Wegman, 1990) illustrating the five-dimensional parallel coordinates marginal densities, correlations, three-dimensional clustering and hyperplanes, and a five-dimensional mode.

Interpretation of parallel coordinate plots raises the question of the ordering of the variables. The ordering of the variables in Figure 3 made it easy to compare axis 1 with 2, 2 with 3, and so forth. However, the pairwise comparisons between 1 and 3 or 1 and 5 are not easy to see since those axes are not adjacent to one another. Ordering the variables by setting the variables with the highest correlation next to each other can help finding clusters since adjacent variables show the most information when it comes to interpreting correlations between variables. In the paper by Wegman (Wegman, 1990), it would take (n+1)/2 for n odd and n/2 for n even permutations, with n being the number of variables, to view all the potential adjacencies in the plot. In the nutrition data set, with 40 food groups to be analyzed, there would be 40/2 = 20 permutations. Instead of trying to look at every possible position, we can start with one variable, say the energy adjusted refined grains, and see which food group is most strongly correlated either negatively or positively with refined grains. In this case, whole grains is most strongly correlated with refined grains so whole grains would be put beside refined grains. Then the procedure would be repeated to find which variable was most strongly correlated with whole grains and that variable would be plotted on the third axis and so on.

Parallel coordinate plots can also be more informative when the variables share the same scale; it reveals the variables with larger variability if all the variables have the same units. It is important to use the same scale for each variable so that the variables are comparable. For example, in the nutrition data set we would expect high variablity for whole grains, but not the alcoholic beverages.

2.1.2 Illustration of Parallel Coordinates for Nutrition Data

The downfall in parallel coordinate plots, which is an issue in any graphics plot of large data sets, is overplotting. I found that even adjusting the opacity of the lines, it was still difficult to actually see the whole picture of what was going on.

Figure 4 is a parallel coordinate plot applied to the nutrition data set. The variables explored are, from left to right: cumdemx12z (cognitive status with those with dementia at any time in the study at the top then following is the cognitively impaired and those with normal cognition), v1msadj (baseline memory score), v1age (age at first interview), energy adjusted refined grains, and energy adjusted whole grains. Notice that I have highlighted, in red, the youngest people; those people score high on the memory test, are mostly cognitively normal, mostly consume less than the average or the median energy adjusted refined grains. Also, those that consume smaller amounts of refined grains, consume more whole grains and vice versa. Thus, there is a negative correlation between whole grains and refined grains. We can also observe that the distribution of age is right skewed since darker areas are closer to the younger ages. Without the tool to highlight polylines and to adjust the opacity, the interpretation is limited. Discussion of the interactive tools are in Section 2.4 and 2.5.

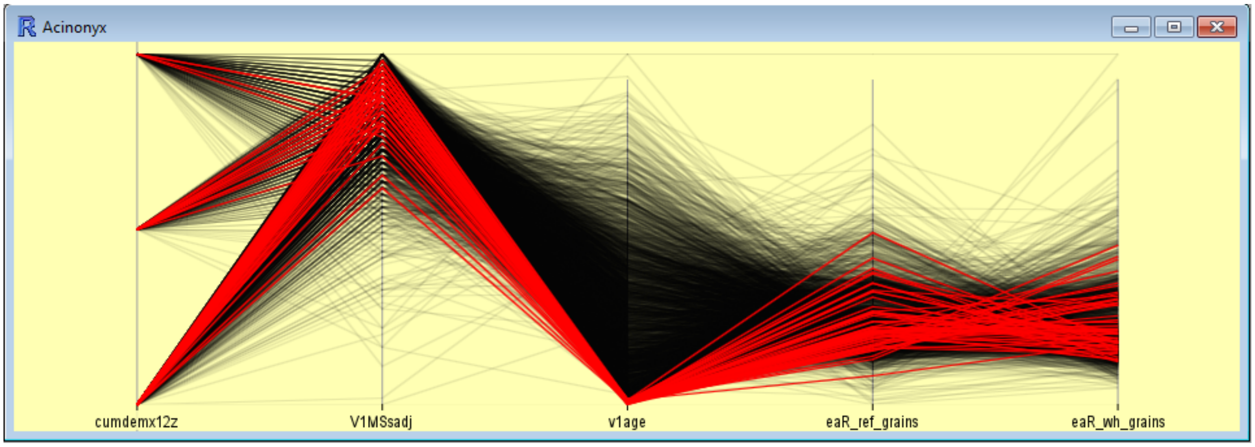

Figure 4: Example using the nutrition data set. The order of the levels of the cumdemx12z are from top to bottom: people with any kind of dementia, people that are cognitively impaired, and those that have normal cognition.

2.2 Variable Importance

2.2.1 General Background

Random Forests is a new and powerful statistical classifier that uses bootstrap samples (repeated sampling with replacement from the learning set) and randomness in the tree-building procedure. Leo Breiman (Breiman, 2001) defined Random Forests (RF) as a combination of tree predictors such that each tree depends on the values of a random vector sampled independently and with the same distribution for all trees in the forest. The advantages of using Random Forests in comparison to other statistical classifiers are: high classification accuracy, a novel method of determining variable importance, ability to model complex interactions among predictor variables, flexibility to perform several types of statistical data analysis, including regression, classification, survival analysis, and unsupervised learning; an algorithm for imputing missing values, robustness to outliers in the predictor variables, insensitivity to monotone transformations of the predictor variables, scaling well for large sample sizes, and dealing with irrelevant predictor variables.

The Random Forests algorithm begins with growing a forest of many trees to a data set; the default number of trees is 500 in R. A tree is grown on each independent bootstrap sample from the training data. At each node, a small number of randomly selected variables, say the square root of the number of variables, is used for binary partitioning and the best split on the selected variables is found. The trees are grown to maximum depth and each tree is used to predict the observations that were not in the bootstrap sample (”out-of-bag” observations). The predicted class of an observation is calculated by the majority vote of the out-of-bag predictions for that observation.

To measure the permutation importance of some variable, say k, consider a single tree and the observations that are out of bag. The out of bag observations are passed down the tree and the out-of-bag error rate for the tree is obtained. Then randomly permute the values of variable k for the out-of-bag data, so each out-of-bag observation gets a random value for variable k and all the other observations are kept at their original values. Pass the modified out-of-bag data down the tree and compute the error rate. If the new error rate is about the same as before this means that the variable does not appear to be contributing to the accuracy of the classification. If the new error rate is higher than before, that variable's values were useful for accurate classification. The average increase in the error rate over all the trees in the forest is used to rank the variables.

Random Forests can handle situations where an important predictor variable is correlated with other predictor variables, that is, there exists some multicollinearity among the predictor variables. The reason that Random Forests can handle the multicollinearity issues is because the algorithm tends to identify all of the correlated predictors as important if any one of them is important. Random Forests will sometimes split on one variable and sometimes on another variable due to the random choice of predictors at each node.

An attractive feature of Random Forests and of all tree-based methods is their ability to capture complex interactions between predictors. If an interaction between two important variables exists the variables involved are likely to show up as important in Random Forest because randomly permuting one of the variables destroys the predictive power of the interaction.

2.2.2 Illustration of Variable Importance for Nutrition Data

The Cache Valley Aging and Memory Study nutrition data set is unique for its demographics in that it consists mainly of elderly that are affiliated with the Later Day Saints (LDS) church in which they don't consume alcoholic beverages, coffee, or caffeinated tea. Therefore, low variability of some particular drinks is expected. People in the Cache Valley area also do not consume much fish and seafood, so it can be difficult to detect whether these food groups are of any importance especially if principal component analysis is used to find important variables. The reason is that principal component analysis is looking for the largest variability, which doesn't necessarily correspond to variable importance. It is also difficult to interpret the components. Finding which foods or food groups are associated with dementia is crucial because foods that affect the elderly dementia status can delay onset of dementia. Dementia is the 3rd most costly disease and it's said that dementia is increasing because the survival rate of Americans is increasing. Random Forests can obtain an intuitive measure of variable importance of the 40 food groups for those that are diagnosed with dementia versus those that are normal. Thus the measure of importance of the predictor variables is quite useful when it comes to variable selection and for interpretation. Although we can run principal components analysis to reduce dimensionality before fitting a classifier or regression predictor, there is the possibility that principal components do not capture the most important information for prediction. For the nutrition data set, it is preferable to obtain variable importance from Random Forests and then, if desired, we can fit principal component analysis on the most important predictors.

Figure 5 is a permutation importance plot, displaying only the 30 most important variables by default, used to illustrate the behavior of Random Forest permutation importance. To obtain the plot in Figure 5, a classification forest is fit to the nutrition data set with 40 food group variables with the outcome variable being the cognitively normal people versus those that are diagnosed with any kind of dementia. The 5 variables that give the largest mean decrease in accuracy (left panel) are energy adjusted versions of: wine, liquor, coffee, desserts, and beer (in decreasing order of importance). The 5 variables that give the largest mean decrease in the Gini index (right panel) are energy adjusted versions of: potatoes, red meat, high-calorie drinks, french fries, and low-calorie drinks. The rankings of the important variables can change if the initial seed for randomization changes, if the number of variables chosen at each node changes, and if the number of bootstrap trees in the forest changes.

The Lasso (least absolute shrinkage and selection operator) (Tibshirani, 1996), a recent popular variable selection method, shrinks the parameter estimates where the ordinary least squares regression coefficients are shrunk toward the origin. A given value of the tuning parameter, c, controls the amount of shrinkage and only a subset of the coefficient estimates will have nonzero values. Note that a smaller value of c reduces the size of the subset. To see how well Lasso performs compared to Random Forests, if we fit Lasso with the 2-level cognitive status variable (0-Normal or 1-Demented) as the response and with the same predictors as above, the top 5 energy adjusted food groups that are added to the regression model indicating importance are in the following order: high energy drinks, tomatoes, coffee, desserts, and tea. Thus, there is an agreement that coffee and desserts are important predictors.

2.3 Multidimensional Scaling

2.3.1 General Background

Multidimensional scaling (MDS) is a statistical method used to study the similarity between objects. MDS is primarily used as a data visualization tool to identify any clustering of points, where the points are viewed as a particular cluster if a certain number of points are close to each other and are not so close to other points that make up another cluster. MDS consists of a family of different algorithms with each algorithm's goal being to come up with an optimal low-dimensional configuration using some type of proximity data. There are a number of MDS methods. Some of methods used are classical scaling or distance scaling. Classical scaling, also known as distance geometry, is an eigenvalue decomposition problem and is the same as principal components if the goal is dimensionality reduction. Distance scaling is also referred to as metric or nonmetric MDS and can be categorized as using metric or nonmetric distances. The distance scaling algorithm uses an iterative procedure to arrive to the solution.

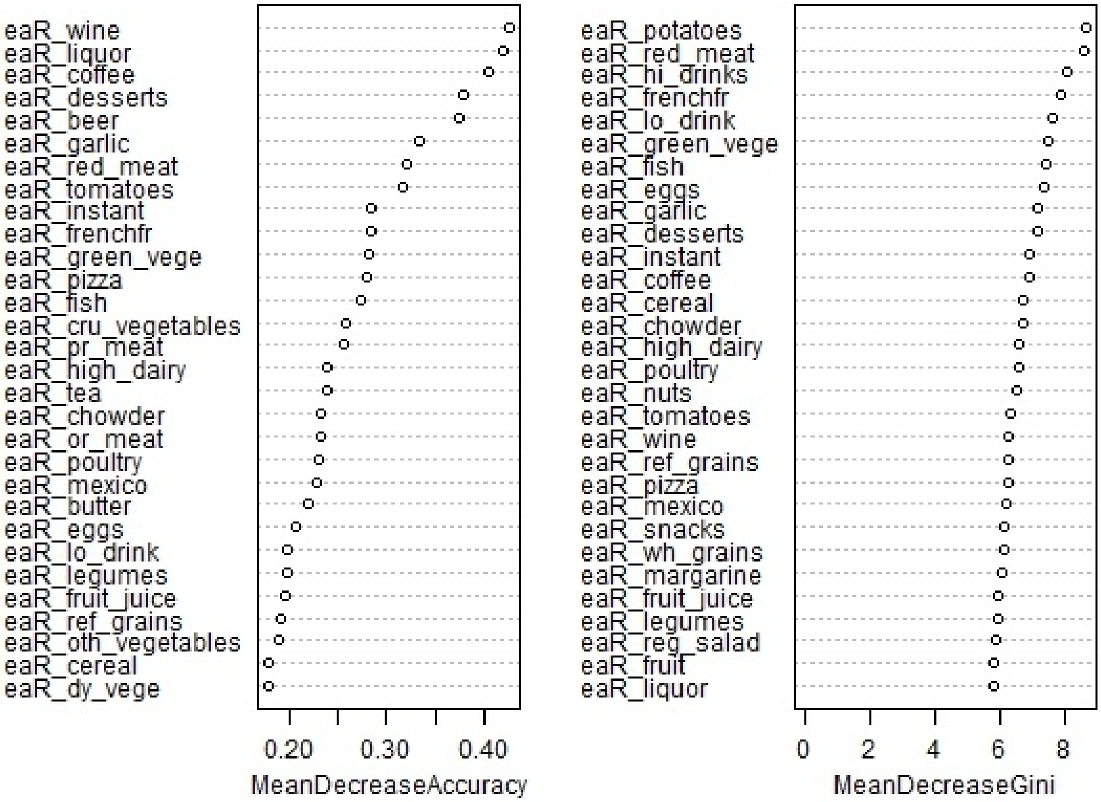

Figure 5: Permutation variable importance of 40 energy adjusted food group variables are in the left panel and the variable importance determined by the Gini Index is in the right panel.

Proximity measures of a pair of entities, such as brand-name products, nations, etc., can be a measure of association or some continuous measure of how alike the entities are. It could be in some cases that the proximity isn't continuous, but an ordinal rating of similarity on a pair of entities. A similarity rating indicates how close a pair of entities are to each other, whereas a dissimilarity rating is the opposite and shows how unalike the entities are. Proximities can be obtained from dissimilarities by, for example, subtracting from the maximum value.

MDS uses a symmetric matrix of proximities. Points representing objects that are close in proximity are objects that are similar to each other, and points that are quite far from each other are dissimilar. Pairs of objects that are very similar will have large values (close to 1) and pairs of objects that are very dissimilar will have small values (close to 0). The points in the MDS plot are close together when the proximity value for the points is high.

Random Forests provide a proximity matrix which can be used to obtain MDS plots in two or three dimensions, which is helpful in identifying multivariate outliers if there are any. Random Forest first computes the proximities between each pair of observations. The proximity in Random Forests is defined to be the proportion, taken over all the trees in the forest, of the time that two observations end up in the same terminal node. More specifically, if a Random Forest of trees is constructed from the learning set and all observations are dropped down all the trees in the forest, the measure of how often a pair of observations occupy the same terminal node will be the measure of how close in proximity that pair of observations is. The proximity matrix is used as input to the classical scaling algorithm. The algorithm will give a visual comparison, that is a scatter plot in two or three dimensions, of the n observations in a lower-dimensional setting by taking the first three principal components of the distance matrix.

2.3.2 Multidimensional Scaling of Random Forests Proximities for Nutrition Data

Figure 6 applies MDS on the 40 energy adjusted food groups, excluding the cognitively impaired, and those aged less than the median age (74). Figure 6 reveals that using the Random Forests proximities gives 2 clusters, while the MDS plot using euclidean distances shows no sign of distinct clusters. Many would want to know what those clusters are. It is difficult to answer this question without using an interactive plot (Section 2.4). Using an interactive plot along with a parallel coordinate plot (Section 4), I can discover what makes up those clusters.

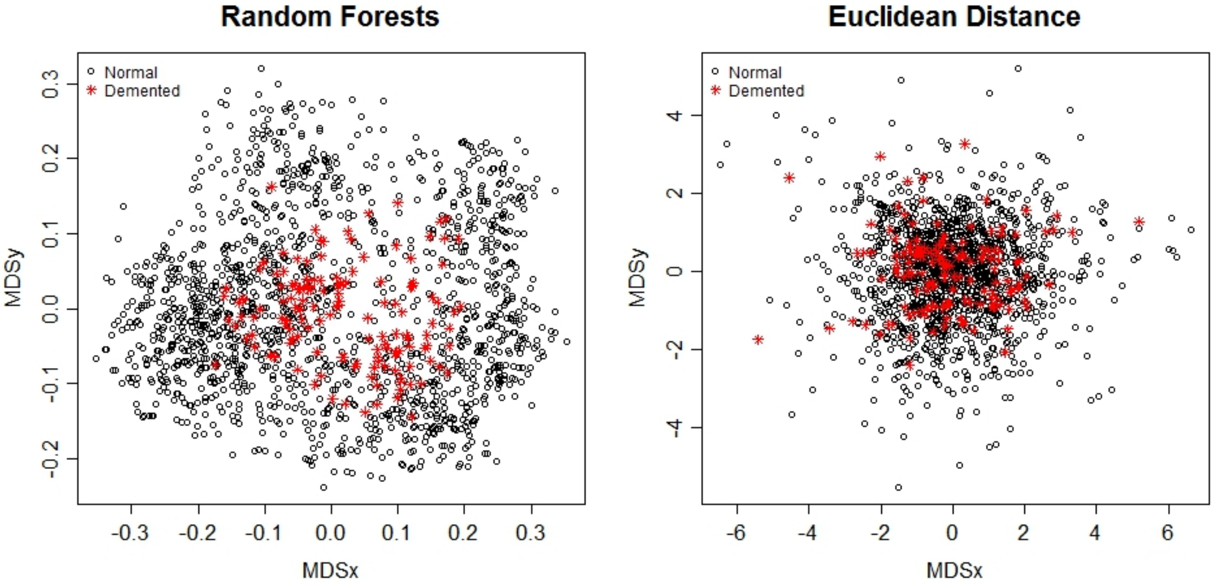

Figure 6: Left panel illustrates the MDS plot using Random Forests proximities and right panel uses euclidean distances. The red stars indicate those with dementia and the open black circles are those with normal cognition. (See, for example, Lattin et al.2003 (Lattin et al., 2003))

2.4 Interactive Graphs

Graphics and plots that allow direct interaction by allowing the user to manipulate the visible graphics by input devices such as a keyboard or mouse, that makes changes on the visible result are known as interactive graphics. Interactive graphics are a useful unsupervised learning tool to get the user acquainted with the data set. Some of the most basic functionalities that an interactive statistical graph may have are selection or highlighting, queries, zooming, direct change of parameters, and multiple views. Selection and highlighting changes the state of points or lines in all components. In other words, if a selection of points or lines is highlighted, the corresponding points or lines in other plots will be highlighted as well. Queries allow the user to obtain information without having to have labels which can create clutter. Zooming allows the user to look at a complex plot without any loss of context. The ability to change parameters lets the user interactively influence properties of the plot that will allow a different view of the data.

A challenge and desirable aspect of interactive graphs is exploring the interactive graphs with an instant response time, using an intuitive and consistent interface for large data sets.

2.5 Tools

In interactive statistical graphs, some useful tools include selection by simply using a drag-box that selects a rectangular region, brushing that allows us to move the selected rectangular region across a plot and dynamically changing the selected subset, highlighting when a region is selected and all the corresponding cases in other plots are highlighted with the same color. Color brushing assigns a color to each observation usually defined by a categorical variable, and alpha-blending or alpha-transparency effectively handles overplotting by changing the opacity.

2.6 iPlots eXtreme

There are software packages that provide interactive graphs, however there are hardly any software packages that can handle large data sets. However, for convenience it is desirable to produce interactive graphs within R rather than having to do some data manipulation then having to export the data set. Packages in R are not all built to deal with large data sets; it tends to be quite slow. There are a limited number of packages that do interactive graphics partially because it is not easy to implement interactive graphics for data analysis and many packages or software are created to offer graphics for a particular problem and leave off the fundamental plots. A package in R that is in the process of completion that focuses on speed optimization in order to handle large data sets is called iPlots eXtreme-the next-generation interactive graphics for analysis of large data in R (Urbanek, 2011), also known as Acinonyx. This package is designed for exploring large data sets with basic plots such as bar plots, histograms, and parallel coordinate plots.

The goal of the creation of iPlots eXtreme was to focus on performance, interactive models, and intuitive use. Performance is quite important when analyzing large data sets. Since iPlots eXtreme is not only available as stand alone software, but is also available as a package in R, using code within R allows the user to share data such that no data would have to be exported to be read into a different software package to explore the data set, which reduces the memory usage dramatically. Statistical models that can be visualized in plots and modified interactively are called interactive models and it is desirable to be able to not only visualize them in a plot, but also to interact with the model. Last, intuitive use is important in interactive graphics because the user needs to be able to think about the data, not the graphical controls and interface.

Some of the modifier keys that will be useful in the interactive plots the I've created using my function (Section 3) are the following: holding shift and selecting additional observations, pressing control while hovering the mouse over the barplot will query the plot and will give the coordinates in the MDS plot, and using the left and right arrow keys will change the transparency in the parallel coordinate plot and MDS plot. There is an option in the MDS plot to go back to see what has been highlighted previously, in other words, the interactive plot records what the user highlighted before and another option that lets the user swap the x-axis and y-axis of the MDS plots. The barplot has the option of brushing the bars by class, ordering the barplots, and switching from a barplot to a spline plot.

3 irfplot

Before we get into running iPlots eXtreme, some preprocessing must be done. First of all we need to run Random Forest to be able to use the function I created to output the parallel coordinate plot using raw values, variable importance values, three MDS plots, and either a histogram or barplot depending whether the outcome variable is continuous or categorical. My function is created so that the user will just have to input the randomForest object, the predictor variables that were input to the randomForest function, the number of important variables to be shown in the parallel coordinate plot, and whether they want the parallel coordinate plots to be scaled or not. By default, my function will not scale the parallel coordinate plots and will display the 10 most important variables determined by the permutation importance.

The name of my function is called irfplot, which stands for interactive Random Forest plot. The function is displayed below with the four arguments that were described above:

```
irfplot(rf, x, n=10, scale=FALSE)
```

An example of the graphics produced by my function is shown in Figure 7. More details are given in Section 4.

4 Results

The variables of interest are the energy adjusted foods. These foods were explored in parallel coordinate plots and the MDS plots. The top ten variables that Random Forests called important using the mean decrease in accuracy criterion are used to display both parallel coordinate plots, in order from left (most important) to right (least important). Three different views of the MDS plot are given, with the first plot giving the best pattern since the first two principal components are the most informative. The plot in the lower right gives the bar plot of the outcome variable since I will be focused on looking for patterns of the normal versus demented elderly, a dichotomous response. The observations are colored by brushing the bar plot. For this particular result, the parallel coordinates plot is scaled.

In Figure 7, the people that are being explored are those aged less than or equal to 74.1. The observations are colored by which class they belong to, that is, the people colored blue are all cognitively normal and those that are demented are represented in orange. The top panel is the parallel coordinate plot of the energy adjusted foods showing that people with higher consumption of wine, liquor, coffee, desserts, beer, and tomatoes are normal. Those that are demented are not consuming as much food compared to those that are cognitively normal overall. Notice there is larger variability for desserts and less variability for the alcoholic drinks, garlic, instant foods, and french fries. We can thus assume that this subset of people consume more desserts, red meat, and tomatoes per day than the other food groups displayed. Also we can see that there are some outliers for most of the food groups. There is a positive correlation between wine and liquor, since the lines are more parallel than crossing between the two axes. The other pairs don't show much of a correlation. Modifying the transparency identified that the distribution of each food group is right skewed and verified those that look like outliers really are outliers rather than an overplot of lines. However, highlighting the observations that we think are outliers will be linked to the bar plot and would give us the counts of the number of people highlighted by hovering the mouse over the bar and pressing the ctrl key.

The second panel is the parallel coordinate plot of the variable importance values from Random Forests displaying the most important variable on the left (wine) to the least important variable (french fries). Higher importance values, which are mostly people that are normal in this case, are the observations that are correctly classified. However, there are negative importance values for those that are normal so those people would have a detrimental impact on classification. Since wine is the most important predictor in classifying whether someone is normal or demented, if we highlight the top portion of those that are normal, we find that those people don't consume much wine, liquor or beer according to the parallel coordinate plot in the top panel. Those people show up as a cluster in the MDS plots in the non-Mormon cluster, which is discussed below.

In the lower panel are the MDS plots produced by the proximities from Random Forests.

From left to right are the first and second principal components of the distance matrix (Random Forests proximities) plotted against each other, then the first and third principal components, and last are the second and third principal components. Each of the MDS plots show that there are two clusters and also a cluster of those that are demented in the center of those that are cognitively normal. As mentioned before in Section 2.3.2 it is difficult to figure out why there are two clusters. In Figure 8, those consuming the least amount of coffee (top panel) are highlighted and we can then see that there are distinct clusters in each of the MDS plots. If the high consumers of wine, liquor, coffee, and beer are highlighted we can see that those people make up the right cluster in the third MDS plot.

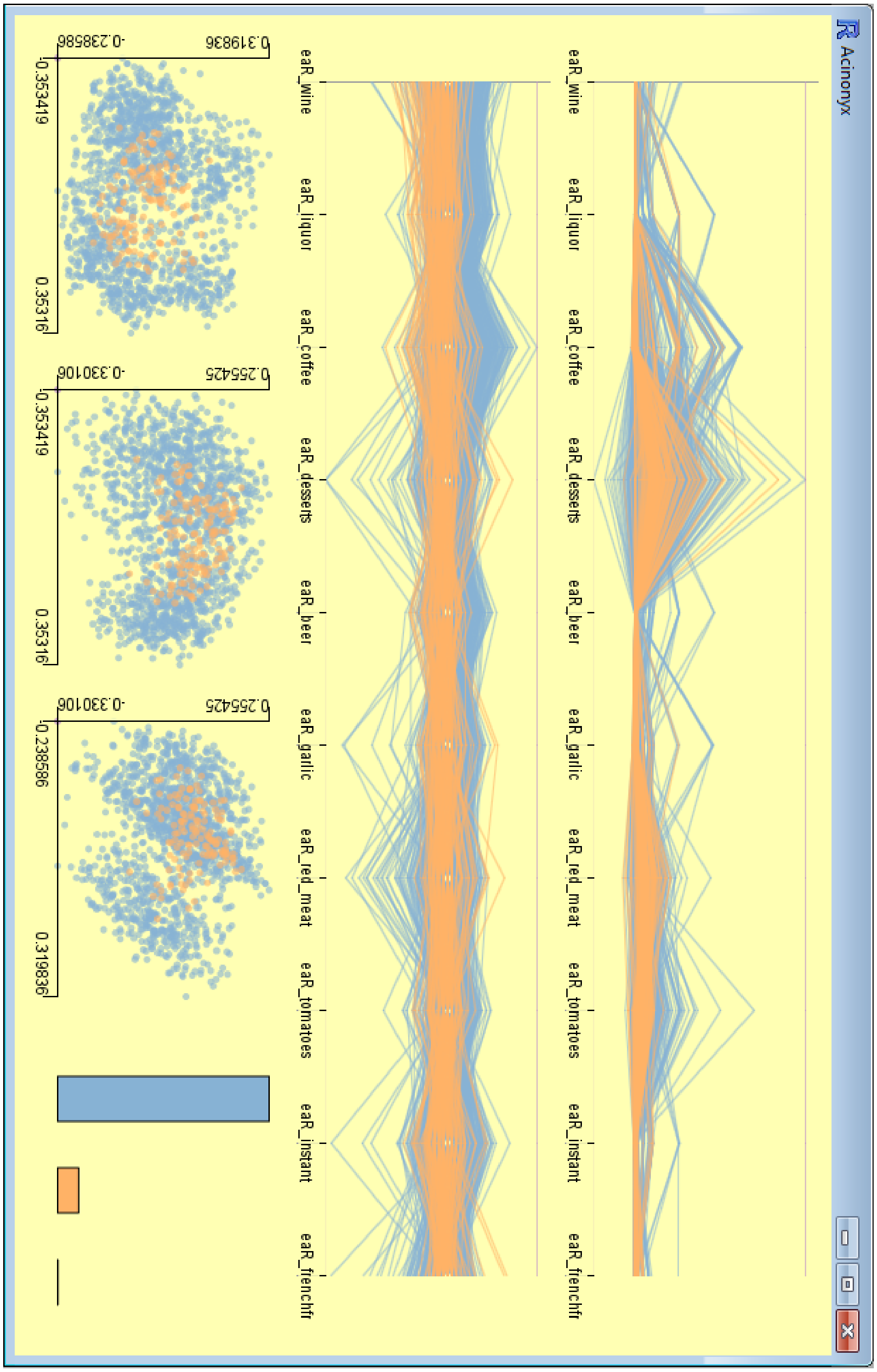

Figure 7: The top panel shows the raw values of the food groups. The second panel shows the importance values from Random Forests. The three scatterplots at the lower left are the MDS plots produced from Random Forest, and the last plot on the lower right is the bar plot of the outcome variable. The observations are colored by cognitive status (blue=normal and orange=demented).

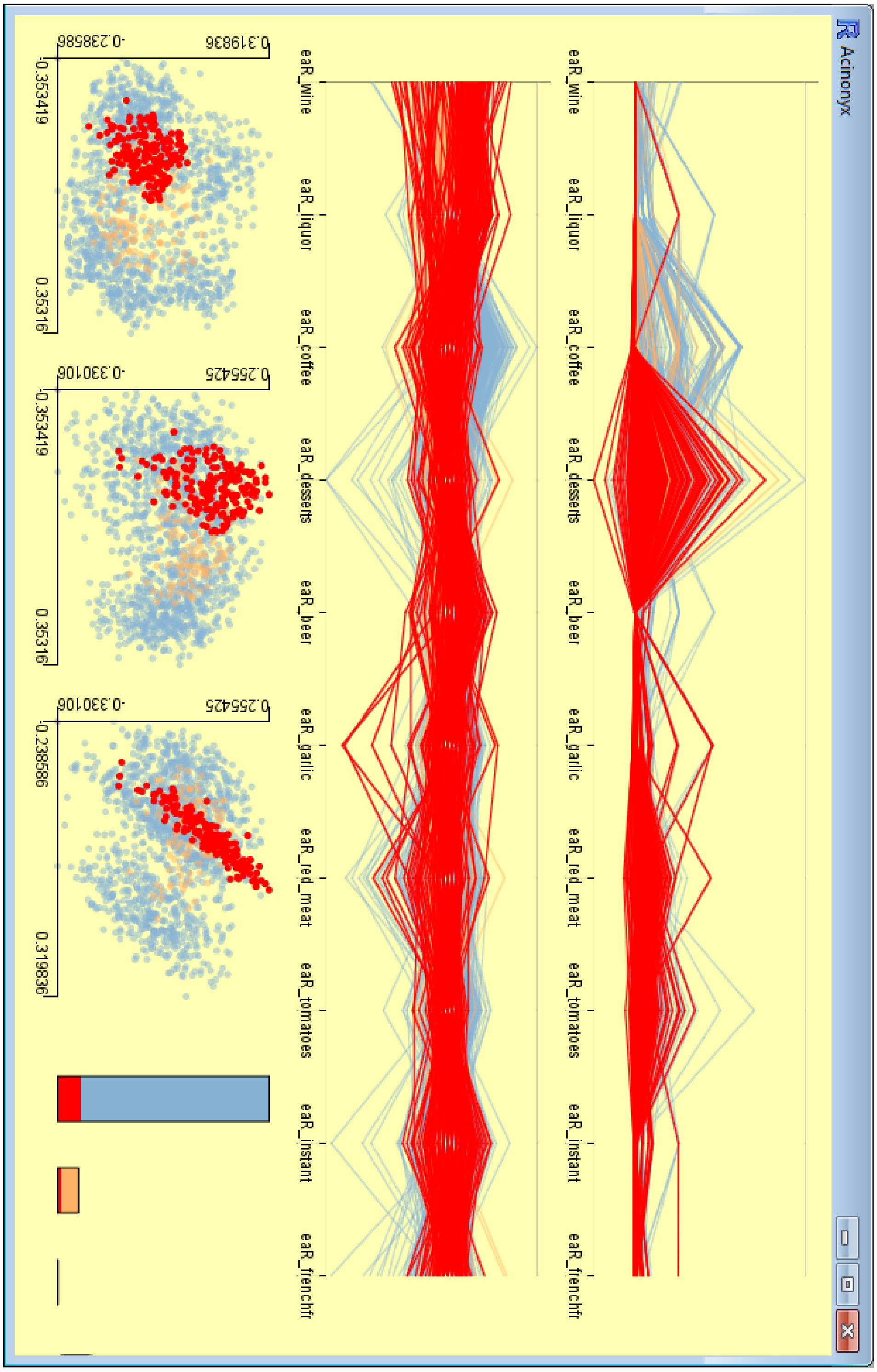

Figure 8: Figure is same as Figure 7, but the lower consumption of energy adjusted coffee is highlighted in red.

5 Conclusion

Random Forests can be difficult to understand and interpret. However, using Random Forests importance values and proximities gave more insight into finding dietary pattern and clusters when it's represented in interactive graphs. Using the iPlots eXtreme package to create the plots I was interested in in R made it easy to explore the large data set because of it's quick response when cases are highlighted and when the transparency changes. My function will be an additional data mining tool for users to gain insight into understanding Random Forests and making use of the Random Forests results through visualization.

6 Discussion and Future Work

As mentioned above, the iPlots eXtreme R package is still in development. Therefore, some of the basic graphical features are excluded. As far as I have found the following features do not work: a main title for each plot, changing the colors used for the classes, x-axis and y-axis labels. Also, the colors chosen for the classes by default are difficult to distinguish, especially if the plots are printed in black and white. Once the iPlots eXtreme package is completed, I would like to update my function.

A new package called Random Survival Forests (Ishwaran et al., 2008) is quite similar to Random Forests but takes censoring into account and is comparable to Cox model regression. The Random Survival Forests package produces proximity and variable importance values as well. Therefore, my function could be used with little change for Random Survival Forests. I would like to further investigate the nutrition data set by applying both Random Survival Forests and iPlots eXtreme together.

References

Breiman, L., 2001. Random forests.

Breiman, L., Cutler, A., 2005. Random forests.
URL www.math.usu.edu/~adele/forests

Cook, D., Swayne, D. F., 2007. Interactive and Dynamic Graphics for Data Analysis: With R and GGobi, 1st Edition. Springer, New York.

Cutler, A., Cutler, D. R., Stevens, J., 2008. High-Dimensional Data Analysis in Oncology. Springer, Ch. Tree-based Methods.

Cutler, A., Cutler, D. R., Stevens, J., 2012. Ensemble Machine Learning: Methods and Applications. Springer, Ch. Random Forests.

Cutler, A., Stevens, J. R., 2006. DNA Microarrays, Part B: Databases and Statistics. Vol. 411. Academic Press, Ch. Random Forests for Microarrays, pp. 422–432.

Cutler, D. R., Edwards Jr., T. C., Beard, K. H., Cutler, A., Hess, K. T., Gibson, J., Lawler, J. J., 2007. Random forests for classification in ecology. Ecology 88 (11), 2783–2792.

Inselberg, A., 2009. Parallel Coordinates: Visual Multidimensional Geometry and Its Applications. Springer, New York.

Ishwaran, H., Kogalur, U. B., Blackstone, E. H., Lauer, M. S., 2008. Random survival forests. The Annuals of Applied Statistics 2 (3), 841–860.

Izenman, A. J., 2008. Modern Multivariate Statistical Techniques, 1st Edition. Springer, New York.

Lattin, J., Carroll, D., Green, P., 2003. Analyzing Multivariate Data. Brooks/Cole, Pacific Groove.

Theus, M., Urbanek, S., 2008. Interactive Graphics for Data Analysis: Principles and Examples, 1st Edition. Chapman and Hall/CRC, Boca Raton.

Tibshirani, R., 1996. Regression shrinkage and selection via the lasso. Journal of the Royal Statistical Society 58 (1), 267–288.

Urbanek, S., 2011. iplots extreme - next-generation interactive graphics design and implementation of modern interactive graphic. Computational Statistics 26 (3), 381–393.

Urbanek, S., Wichtrey, T., Gouberman, A., Theus, M., 2005. iplots.
URL http://www.rosuda.org/iplots/

Wegman, E. J., 1990. Hyperdimensional data analysis using parallel coordinates. Journal of the American Statistical Association 85 (411), 664–675.

Wengreen, H. J., Munger, R. G., Corcoran, C. D., Zandi, P., Hayden, K. M., Fotuhi, M., Skoog, I., Norton, M. C., Tschanz, J., Breitner, J. C., Welsh-Bohmer, K. A., 2007. Antioxidant intake and cognitive function of elderly men and women: the cache county study. The Journal of Nutrition, Health and Aging 11 (3), 230–237.

Willett, W., 1989. Nutritional Epidemiology. Oxford University Press.