Visualizing Categorical Data Arising in the Health Sciences Using Hammock Plots

Matthias Schonlau

RAND Corporation

Abstract

I introduce the hammock plot, a new plot that can be used to visualize categorical data and mixed categorical / continuous data. It is well suited for all types of categorical data: both unordered and ordered categorical data as well as interval data. It can be viewed as a generalization of parallel coordinate plots where the lines are replaced by rectangles that are proportional to the number of observations they represent. In addition to the rectangles, hammock plots also incorporate univariate descriptors such as category labels into the graph. I illustrate the hammock plot with examples from the health sciences.

Introduction

There are four commonly used plots for visualizing high dimensional data: scatter plot matrices (Hartigan 1975), Mosaic plots (Hartigan and Kleiner, 1991, Friendly, 1994, Theus, 1996), parallel coordinate plots (Inselberg 1984, Wegman 1990) and Trellis displays (Becker et al. 1996, Theus, 1999).

Scatter plot matrices and parallel coordinate plots are best suited for continuous data. Over plotting is sometimes a problem for these plots. Over plotting occurs when more than one observation is assigned to the same physical space on the plot. A common way to deal with (unordered) categorical data is to assign each category a numerical value. If these values are then plotted an extreme amount of over plotting occurs. Consequently, neither scatter plot matrices nor parallel coordinate plots do well with categorical data either. Jittering, i.e. adding spherical noise to an observation, can be used to alleviate the over plotting problem somewhat.

Mosaic plots were conceived to display categorical data. Mosaic plots are not suited for continuous data at all. There are several types of categorical data: unordered categorical data, ordered categorical data, and interval data. Interval data are ordered data for which the separation between data points has meaning (Agresti, p.3). The number of comorbidities of a patient, or the number of children of a parent are examples of interval data. Mosaic plots are well suited for unordered and ordered categorical data. Mosaic plots treat interval data like ordered categorical data - the distance between categories is not visually apparent.

Neither Mosaic plots, scatter plot matrices nor parallel coordinate plots are well suited for data that have both categorical and continuous variables. In Trellis displays one specific plot (e.g. scatter plot or a box plot) is displayed for different subsets of conditioning variables. These plots are then arranged as a panel. For example, one might display two continuous and one categorical variable as a panel of scatter plots – one for

each category of the categorical variables. Therefore Trellis displays are suitable for displaying mixed continuous / categorical data.

For survey researchers missing data are very important. Most plots do not to accommodate missing data, presumably because the researchers who conceived these plots did not work with surveys. Mosaic plots are an exception: missing values have sometimes been added as an extra category. A similar approach is possible with scatter plot matrices or parallel coordinate plots, but I have never seen this being done.

I introduce a new plot for the visualization of categorical data that also handles interval data and mixed categorical /continuous data. I introduce the hammock plot in the next section. The following section gives several examples. The paper concludes with a brief discussion.

The Hammock Plot

It is easiest to introduce the hammock plot by means of an example. Figure 1 shows a hammock plot with four variables from a survey of children, adolescents and adults with asthma. The four variables displayed here are: a variable that distinguishes between children, adolescents and adults ("group"), number of nights spent in the hospital in the last 30 days ("hosp"), number of comorbidities ("comorb"), and gender.

The hammock plot consists of alternating univariate variable descriptors and bivariate graphs. In Figure 1 four variables are displayed. The labels and numbers are the univariate descriptors; the solid black rectangles and lines are the bivariate graphs. For example, the univariate descriptor for the first variable, group, corresponds to the three category labels "child", "adolescent", and "adult". For the second variable, "hosp", the univariate descriptor consists of the actual number from 0 through 20. The variable name appears at the bottom of the plot beneath the corresponding univariate descriptor. It is also clear from Figure 1 that each univariate descriptor is scaled independently from the other variables. For example, the values 20 and 7 for the variables "number of hospital nights" and "number of comorbidities" both appear at the top at the same vertical position within the graph because they both represent the maximum range for their respective variables. On the same scale these numbers would be far apart from one another.

At the bottom of the graph a category may be added for missing values for each variable. This category is separated from the others by a horizontal line.

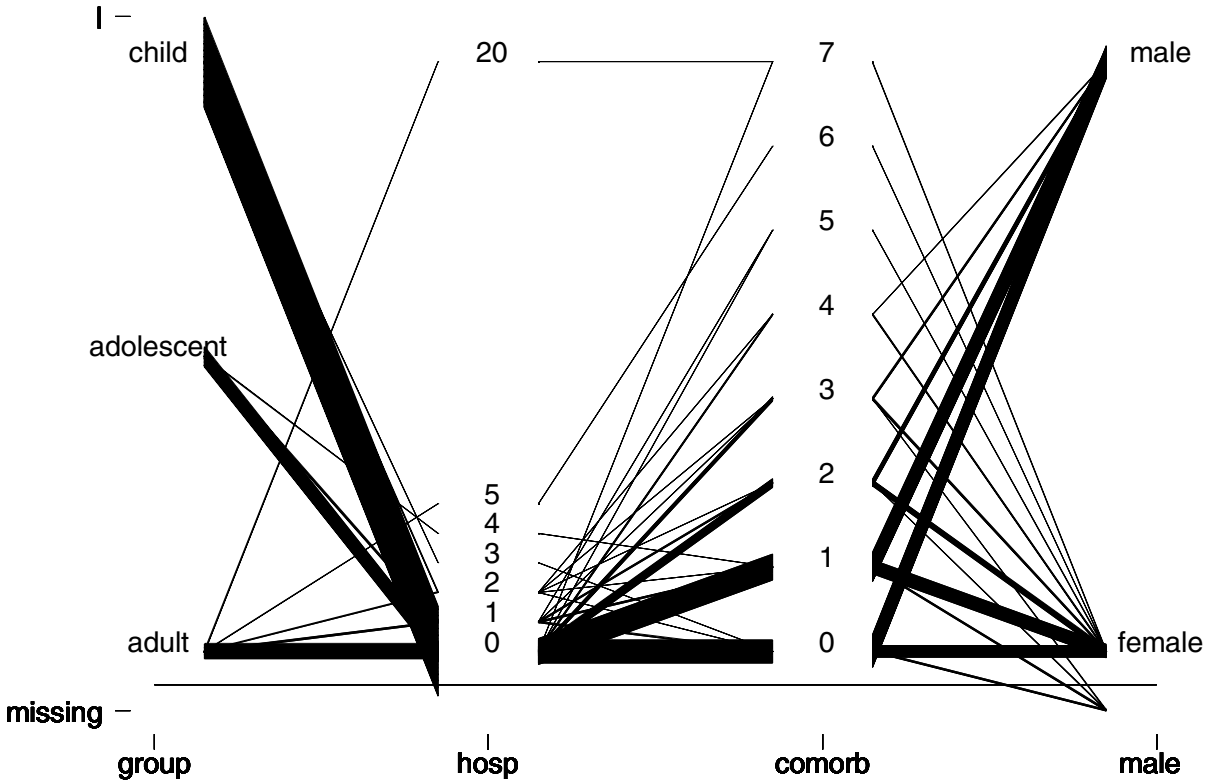

Figure 1: A Hammock plot with 4 variables

The bivariate graph consists of rectangles[1] that connect categories of adjacent variables. The "width" of a rectangle is proportional to the number of observations it represents. What is meant by "width" of a "rectangle" is illustrated in Figure 2. The "width" of the diagonal "rectangle" is defined as distance d1. To see that d1 is preferable to d2 consider that the horizontal rectangle appears wider than the diagonal rectangle. However, both rectangles have the same vertical width d2. The eye focuses on d1, not on d2. This point has previously been made by Wallgren et al (1996). If the rectangle represents only one observation, then the width of the rectangle is very small and the rectangle will appear to be a line.

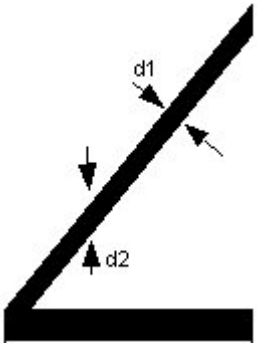

Figure 2: Two ways of defining "width". Distance d1 is preferable.

[1] Parallelograms are shown but I intend to replace them with rectangles in a future implementation.

Some of the information we can glean from Figure 1 is as follows: most people in this sample were children rather than adolescents or adults. One person had 20 hospital visits. This person was a female adult with 7 comorbidities - more comorbidities than any other person in the sample.

Examples

I give two further examples in this section. These examples are not connected with one another; they merely serve to illustrate different uses of the hammock plot. I am also not trying to fully analyze a data set with the hammock plot because I consider the hammock plot as one tool; not as the only tool. These examples convey the power of hammock plots only incompletely because they only give static views. The hammock plot is most powerful when used interactively for the analysis of data.

Experimental Design

We (Adams et al., 2003) conducted a simulation study to test the effect of three sampling design parameters on sampling coverage and effective sample size. This strategy for selecting sampling designs based on simulations is described in more detail in Adams et al. (2003). Specifically, we conducted a factorial experiment of three input parameters alpha, cutrial and cutinst. Each combination of the three input parameters forms a different sampling design.

Figure 3 displays a hammock plot of the three input parameters and the two output parameters. One can clearly see the regular structure of the factorial design of the

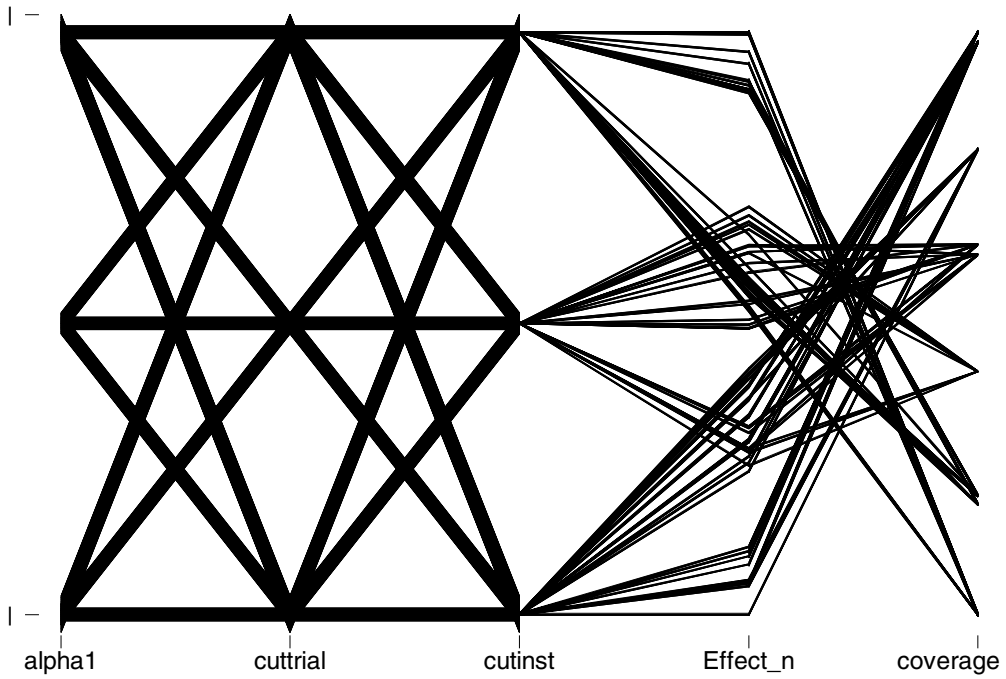

Figure 3: Hammock plot of three input parameters and two output parameters in a simulation study for sampling designs

input parameters. Further, the input parameter "custinst" has a strong influence on the effective sample size called "effect_n" in Figure 3. Moreover, the effective sample size and coverage appear strongly negatively correlated with one another because low effective sample sizes lead to high coverage values and vice versa. This makes sense: to achieve a high coverage one has to include even very small clusters of sampling units in the sample frame. This will typically lead to a larger (meaning worse) design effect implying a smaller effective sample size. The effective sample size is defined as the sample size divided by the design effect.

Outlier Detection

In a diabetes study chart data of diabetics were obtained. From the chart data we computed the number of times each patient had had measurements of A1c (hemoglobin) in the past 12 months ("num_hba1"). We also computed how many months prior to the chart review the last measurement was taken ("mn_lst_h"), the last A1c measurement "lst_hba1", and the average value of all measurements over the past 12 months ("mean_hba"). A hammock plot of these 4 variables can be seen in Figure 4.

The plot clearly shows that there were patients with an A1c value of 60. Because the rectangle that connects into the label "60" is very thin - it appears to be a line – we suspect that there is only one patient with this value. This turns out to be true. This A1c value is much higher than all other A1c values. The medical doctor in our team explained to us that A1c values above roughly 20 were medically not possible. Consequently, both the values 60 and 31.63 were errors.

The plot also shows a large number of patients with missing values. Almost all of these patients had missing values for all four variables. Our sample corresponded to a survey. It turned out that medical charts were only available for some survey respondents; the remaining respondents had missing values. There was one person with medical chart data but he/she had no A1c measurements over the past 12 months.

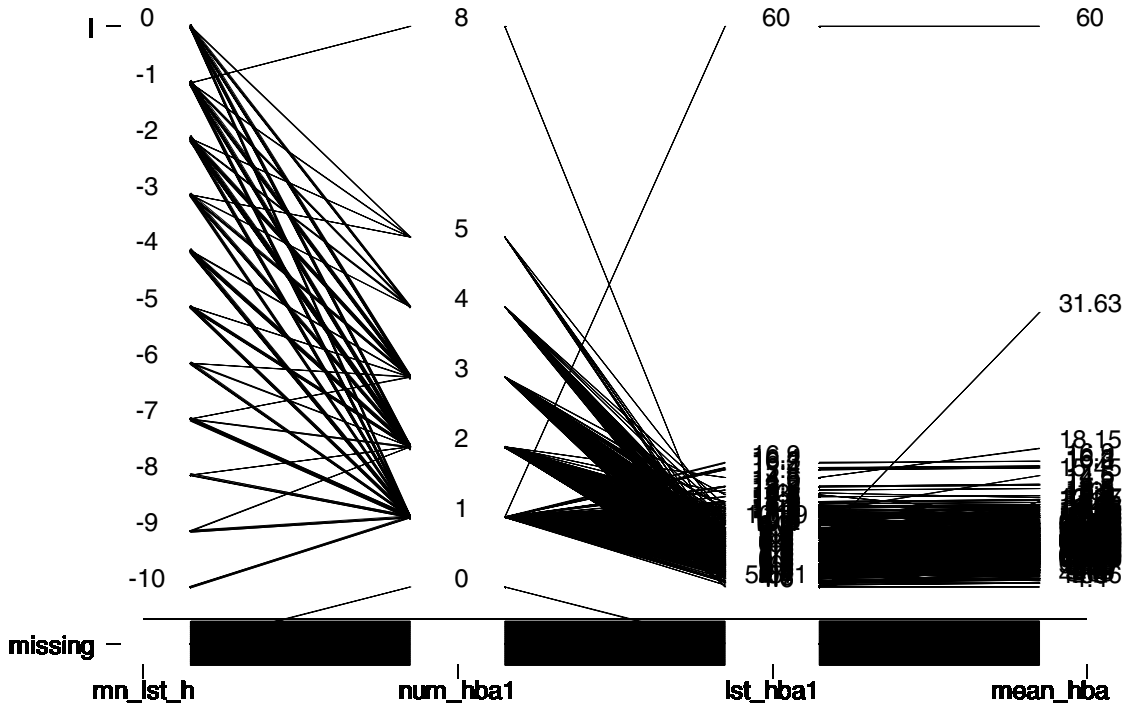

Figure 4: Hammock plot for diabetes data.

Discussion

I have introduced a new plot for visualizing categorical data. A hammock plot can handle mixed continuous/categorical data. It deals well with categorical data that consists of interval data, and it is good at handling a large number of categories.

It is relatively easy to tell independence from Mosaic plots, which is very difficult from Hammock plots. Mosaic plots tend to be problematic when large numbers of categories need be displayed. I believe that the learning curve to start interpreting data with hammock plots is less steep than the one for Mosaic plots though I do not have any evidence to support that claim.

My work on hammock plots grew out of my work on visualizing clusters (Schonlau, 2002). Parallel coordinate plots (Inselberg 1984, Wegman, 1990) emerge as a special case of the hammock plot in case the rectangles have width 0, i.e. when the rectangles degenerate into lines, and when the univariate descriptors are omitted.

Hammock plots are still work in progress. Future work will include tracing highlighted observations through the entire graph, the use of sampling weights, and the ordering of categories of unordered categorical variables.

The hammock plot is implemented in Stata Version 7. The Stata program will be downloadable from www.schonlau.net in the near future.

References

Adams JL, Schonlau M, Escarce J, Kilgore M, Schoenbaum M, Goldman DP. (2003) "Sampling Patients Within and Across Health Care Providers: Multi-Stage Non-nested Samples in Health Services Research." *Draft Paper.*

Becker, RA, Cleveland WS, and Shyu M-J (1996). "The Design and Control of Trellis Display". Journal of Computational and Statistical Graphics, 5, 123-155.

Agresti A. *Categorical Data Analysis*, John Wiley & Sons, New York, 1990.

Hartigan JA and Kleiner B (1981). "Mosaics for Contingency Tables." Computer Science and Statistics: *Proceedings of the 13th Symposium on the Interface*, 268-273, ed W.F. Eddy, Springer, New York.

Inselberg A (1984). "The Plane With Parallel Coordinates." *The Visual Computer*, 1, 69-91.

Friendly M (1994). "Mosaic Displays for Multi-Way Contingency Tables." *Journal of the American Statistical Association*, 89, 425, 190-200.

Hartigan JA (1975). "Printer graphics for clustering." *Journal of Statistical Computation and Simulation*, 4,187-213.

Schonlau M. (2002). "The Clustergram: a Graph for Visualizing Hierarchical and Non-hierarchical Cluster Analyses." *The Stata Journal*, 2 (4):391-402.

Theus M, (1999). "Trellis Displays." Entry in *Update Volume of the Encyclopedia of Statistical Science*, eds. S. Kotz & Johnson.

Theus M. *Theorie und Anwendung Interactiver Statistischer Graphik*. Augsburger Mathematisch-Naturwissenschaftliche Schriften, Augsburg, Germany, 1996 (in German).

Wallgren A, Wallgren B, Persson R, Jorner U, Haaland J, *Graphing Statistics & Data: Creating Better Charts*, Sage Publications, Thousand Oaks, CA. 1996.

Wegman E (1990) "Hyperdimensional Data Analysis Using Parallel Coordinates." *Journal of the American Statistical Association*, 85, 411, 664-675.