Proceedings of the 3rd International Workshop
on Distributed Statistical Computing (DSC 2003)
March 20–22, Vienna, Austria
http://www.ci.tuwien.ac.at/Conferences/DSC-2003/
K. Hornik & F. Leisch (eds.) *ISSN 1609-395X*

Visualizing Independence Using Extended Association Plots

David Meyer[†] **Achim Zeileis**[†] **Kurt Hornik**[‡]

[†]*Institut für Statistik & Wahrscheinlichkeitstheorie, Technische Universität Wien*
[‡]*Institut für Statistik, Wirtschaftsuniversität Wien*

Abstract

Association plots—a visualization technique for the independence problem in 2-way contingency tables—are extended in three directions:

1. The visualization is enhanced by using colors for the importance of the residuals.
2. Two extensions for the use of multi-way tables are proposed.
3. The implementation in the R system is improved using a more modular design and allowing a more flexible specification of plotting parameters.

1 Introduction

Given two categorical variables, or, equivalently, their contingency table, one is often interested in investigating whether they are independent or not. Usually, the χ^2 test statistic—the sum of the squared Pearson residuals—is used as a measure of association. Now, whenever the statistic is significant, one might ask for a more detailed analysis on the basis of the residuals themselves. This task can greatly be sustained by the use of graphical methods such as the association plot, for which we propose some extensions: the use of colors for the shading of the residuals, two possible extensions for multi-way tables, and a new, more flexible implementation using R's alternative graphical system, `grid`. The new functionality is provided by the R package `vcd`, inspired by the book 'Visualizing Categorical Data' (Friendly 2000).

2 Association plots for 2-way tables

To assess independence of 2 categorical variables, usually a 2-way contingency table is considered with cell frequencies $\{n_{ij}\}$ for $i = 1, \ldots, I$ and $j = 1, \ldots, J$ and row and column sums $n_{i+} = \sum_i n_{ij}$ and $n_{+j} = \sum_j n_{ij}$ respectively. For convenience, the number of observations is denoted $n = n_{++}$.

Given an underlying distribution with theoretical cell probabilities π_{ij}, the null hypothesis of independence of the two categorical variables can be formulated as

$$H_0 : \pi_{ij} = \pi_{i+}\pi_{+j}. \tag{1}$$

The expected cell frequencies in this model are $\hat{n}_{ij} = n_{i+}n_{+j}/n$. The best known and most used measure of discrepancy between observed and expected values are the Pearson residuals

$$r_{ij} = \frac{n_{ij} - \hat{n}_{ij}}{\sqrt{\hat{n}_{ij}}}. \tag{2}$$

Therefore, a rather intuitive idea is to reject the null hypothesis when there are residuals which are too extreme, i.e., not close enough to zero. The most convenient way to aggregate the $I \times J$ residuals to one test statistic is their squared sum

$$X^2 = \sum_{i,j} r_{ij}^2, \tag{3}$$

because this is known to have a limiting χ^2 distribution with $(I-1)(J-1)$ degrees of freedom under the null hypothesis. This is the well-known χ^2 test for independence in 2-way tables. Now, when the χ^2 test statistic turns out to be significant for some data, it seems natural to go back to its components, i.e. the residuals, for a more detailed analysis.

Association plots (Cohen 1980) visualize the table of Pearson residuals: each cell is represented by a rectangle that has (signed) height proportional to the corresponding Pearson residual r_{ij} and width proportional to the square root of the expected counts $\sqrt{\hat{n}_{ij}}$. Thus, the area is proportional to the raw residuals $n_{ij} - \hat{n}_{ij}$. The sign of the residual is redundantly coded by the rectangle's color and its position relative to the baseline.

Figure 1 shows the association plot (produced by the implementation in `base R`) for the variables 'Hair' and 'Eye' of the well-known 'HairEyeColor' data set (Hair and Eye colors of 264 male and 328 female statistics students). The biggest tiles are easily made out: e.g., the tiles for brown and blue eyes, given blond hair, seem to represent important residuals. But for the others, their classification in important and less important categories seems not obvious.

3 Improved visualization through residual shading

We propose to use a color-shading similar to enhancements of Friendly (1994) for the mosaic plot (a graphical tool for the visualization of the observed frequencies

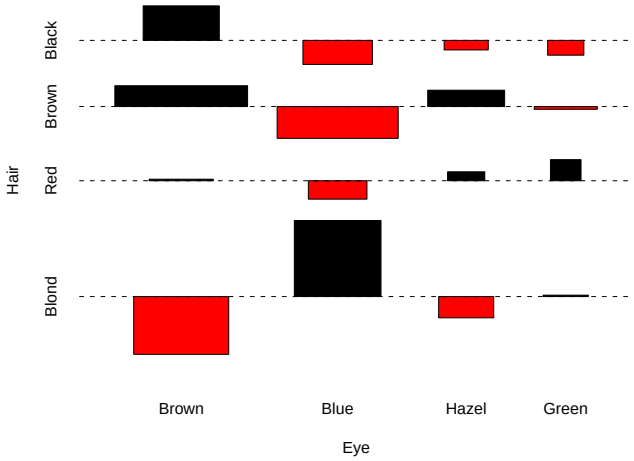

Figure 1: Association plot for 'Hair' and 'Eye' colors of female students.

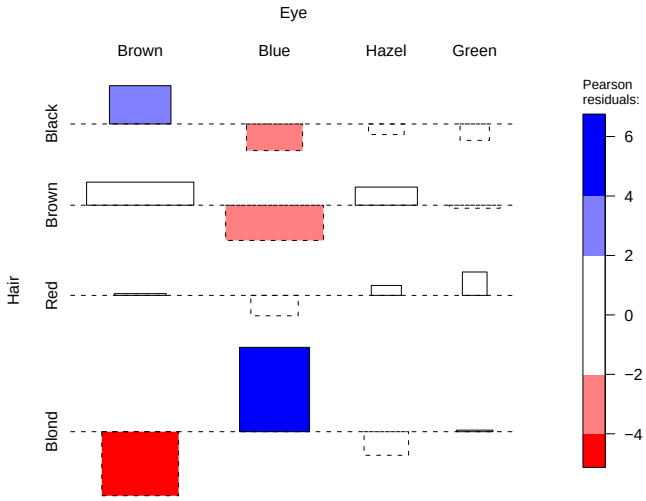

Figure 2: Extended association plot for 'Hair' and 'Eye' colors of female students.

in a contingency table; see Hartigan and Kleiner 1984): this extension uses a color coding of the mosaic tiles to visualize deviations (residuals) from a given log-linear model fitted to the table, that is, from the expected frequencies under independence. Positive and negative signs of the residuals are coded by rectangles with solid and dashed borders respectively. Furthermore, residuals can be classified according to specified shading levels: per default, residuals exceeding an absolute value of 2 are shaded light blue and red respectively, those that even exceed an absolute value of 4 are shaded with full saturation. The heuristic behind this shading is that the Pearson residuals are approximately standard normal which implies that the highlighted cells are those with residuals *individually* significant at approximately the 5% and 0.01% level. But the main purpose of the shading is not to visualize significance but the *pattern* of deviation from independence (Friendly 2000, p. 109). This color scheme similarly facilitates the detection of (in)dependence patterns in association plots. Consider the 'HairEyeColor' example introduced in the previous section: Figure 2 shows the same residuals as Figure 1, but now, the color shading helps us to sort the residuals into the three categories: important–less important–unimportant and thus emphasizes the associational pattern in the underlying data. Another example is given in Figure 3: this time, we cross-tabulate the variables 'Eye' and 'Sex'. For example, the big tiles for blond hair convey the impression of an asymmetry between male and female students, and hence that hypothesis of independence is to be rejected. However, the χ^2 test statistic is not significant at a 5% level: in this example, its value is 6.221, and the corresponding p value approximately 0.1013. Apparently, the relative size of the tiles is insufficient to assess the "strength" of association between two variables; again, information on the absolute size of the residuals would be useful. Using the extended shading (Figure 4), we see at first glance that *no* residual is bigger than 2 (or smaller than -2 for negative ones, respectively) because simply no tile is shaded. In fact, as we can depict from the legend, the overall range of the residuals is smaller than $[-1.5, 1.5]$.

4 Extension to multi-way tables

Independence problems do not only occur in 2-way tables, but are also important in tables of higher dimensionality where they can follow much more complex patterns. These are again defined based on the underlying table of theoretical cell probabilities ($\pi_{ij\ldots}$) with more than two dimensions. Apart from the analysis of all pairwise associations, models of interest include, e.g., the null hypotheses of conditional independence:

$$\pi_{ijk\ldots} = \pi_{i|k\ldots}\pi_{j|k\ldots}$$

A general approach for visualizing such kind of hypotheses are extended mosaic plots. In this section, we give two examples for the applicability of association plots to multi-way tables. A quick overview of the mutual independence structure of a multi-way table (i.e., all pairwise *marginal* associations) can be given by a

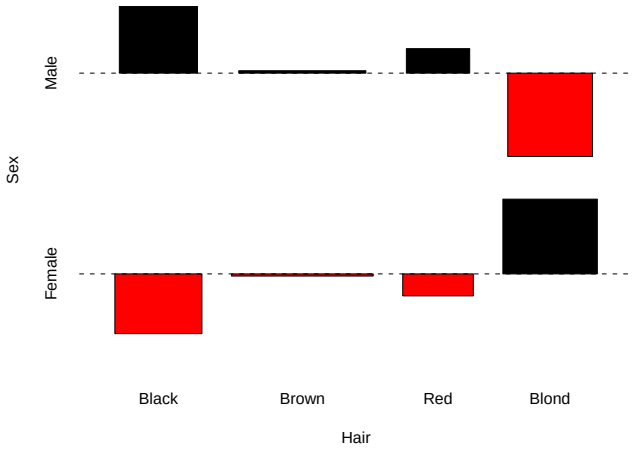

Figure 3: Association plot for 'Hair' and 'Sex' in the 'HairEyeColor' data.

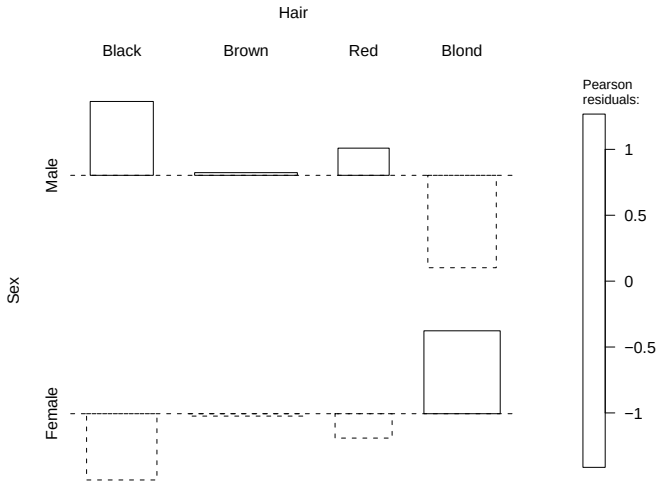

Figure 4: Extended association plot for 'Hair' and 'Sex' in the 'HairEyeColor' data.

matrix of plots where each cell contains an association plot of the corresponding row and column variables. Such an *association pairsplot* allows us to quickly scan the association plots for all variable combinations. As an example, we again use the HairEyeColor data (see Figure 5): only the cells corresponding to the variables 'Hair' and 'Eye' contain an "interesting" association plot with colored residuals.

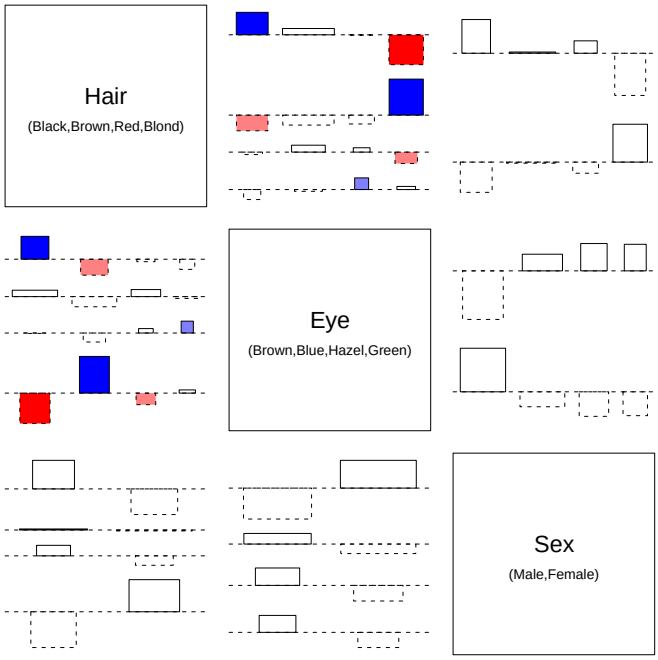

Figure 5: Extended association pairs for the 'HairEyeColor' data, visualizing all pairwise marginal associations.

To visualize *conditional* independence, we could use conditioning plots: like in trellis displays, we produce a separate plot for each level of a conditioning variable (in the case of two or more conditioning variables, for each level combination). This can be illustrated using the famous admissions data of the University of California at Berkley (UCB) which is available in base R. In this data, the question whether there is sex discrimination at the UCB leads to the result that although women seem to be disadvantaged at the aggregated level there is no sex discrimination in the department strata—with the very exception of one department in which women are *more* likely to be admitted than would be plausible under independence. Exactly this is illustrated in the conditioning association plot in Figure 6.

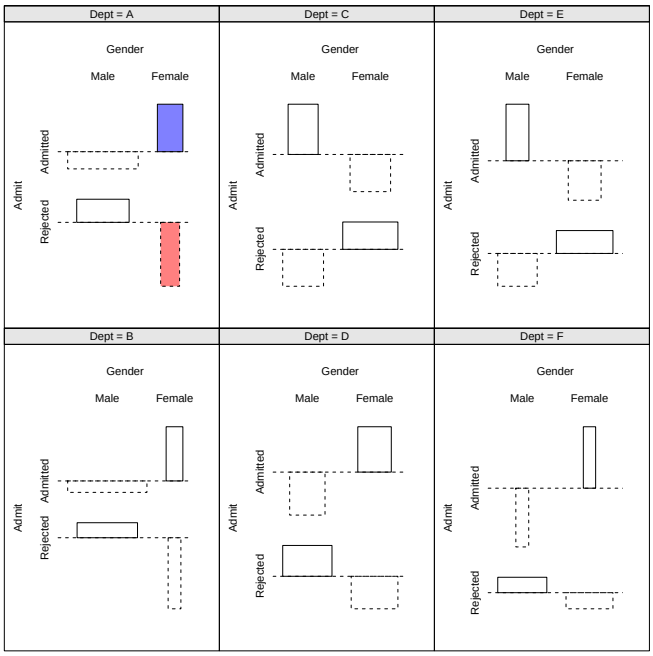

Figure 6: Extended conditioning association plots for the 'UCB Admissions' data.

5 Implementation enhancements

The current implementation of association plots in R suffers from two main disadvantages: First, it is not easy to recycle the plots in conditioning plots or pairs plots as they have been implemented using R's base graphics engine where in general plotting to relative coordinates is not supported. The new implementation was written from scratch in `grid` offering much more versatility amongst some minor advantages and convenient improvements. Second, the graphics parameters of the rectangles in the association plots, such as color and line type, cannot be specified for each cell by the user. To overcome this, the current implementation in `vcd` allows the user to specify either arrays of graphics parameters of the same dimensionality as the object being plotted or a function which computes these graphics parameters based on the original table and its Pearson residuals. Functions are provided for shading schemes like those described in the previous sections.

6 Conclusion

We suggest a set of enhancements for visualizing the independence problem in 2-way tables using association plots. The extensions aim at improving the visualization by displaying the size of the residuals, and by extending the concept of association plots to multi-way tables. Furthermore, a new implementation is outlined based on the graphics package `grid` which provides more modular design and more flexible specification of graphical parameters. Our work is still not at its end, though: for example, the plots for multi-way tables currently do not use the same scale for all subplots, and thus residuals of different subplots are not comparable. Furthermore, the current color scheme only allows the detection of "patterns" of independence. It would be nice, however, that the shading levels were chosen such that a true visual test of independence can be performed, that is, such that the presence of "significant" residuals lead to a rejection of the hypothesis of independence. Current work includes the development of such color schemes, based on an alternative test statistic—the maximum of the absolute values of the pearson residuals.

References

Cohen, A. (1980), "On the Graphical Display of the Significant Components in a Two-Way Contingency Table," *Communications in Statistics—Theory and Methods*, A9, 1025–1041.

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

— (2000), *Visualizing Categorical Data*, Carey, NC: SAS Insitute.

Hartigan, J. and Kleiner, B. (1984), "A mosaic of television ratings," *The American Statistician*, 38, 32–35.