# Visualizing Independence Using Extended Association Plots

**David Meyer**         **Achim Zeileis**         **Kurt Hornik**

### 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. The implementation currently available in R is improved using a more modular design and allowing a more flexible specification of plotting parameters.

3. Two methods for the use of multi-way tables are proposed.

## 1   Introduction

Given two categorical variables, or, equivalently, their contingency table, one is often interested in whether they are independent or not. Usually, the $\chi^2$ 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 mosaic display (Hartigan and Kleiner, 1981; Friendly, 1994; Theus and Lauer, 1999; Hofmann, 2003) or the association plot (Cohen, 1980; Friendly, 2000). For the latter we propose some extensions: the use of colors for the shading of the residuals, two possible methods for multi-way tables, and a new, more flexible implementation using R's alternative graphical system `grid` (Murrell, 2002). 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 $\pi_{ij}$, the null hypothesis of independence of the two categorical variables can be formulated as

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

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

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

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 into one test statistic is their squared sum

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

which is known to have a limiting $\chi^2$ distribution with $(I - 1)(J - 1)$ degrees of freedom under the null hypothesis. This is the well-known $\chi^2$ test for independence in 2-way tables. Now, when the $\chi^2$ 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 $n_{ij} - \hat{n}_{ij}$. In available implementations, the sign of the residual is typically 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; see Snee, 1974). 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 assessing the "importance" of the others is far from obvious.
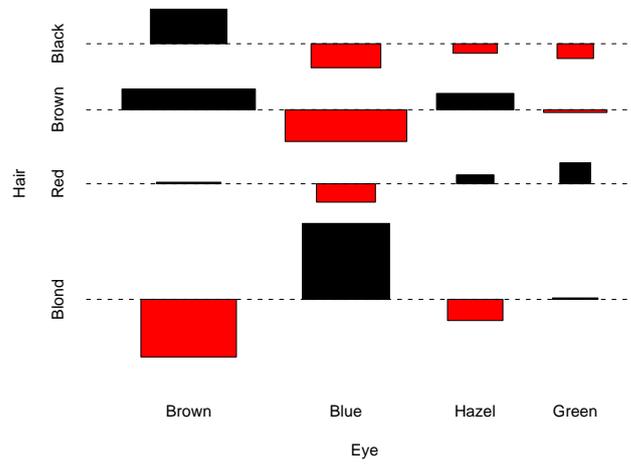
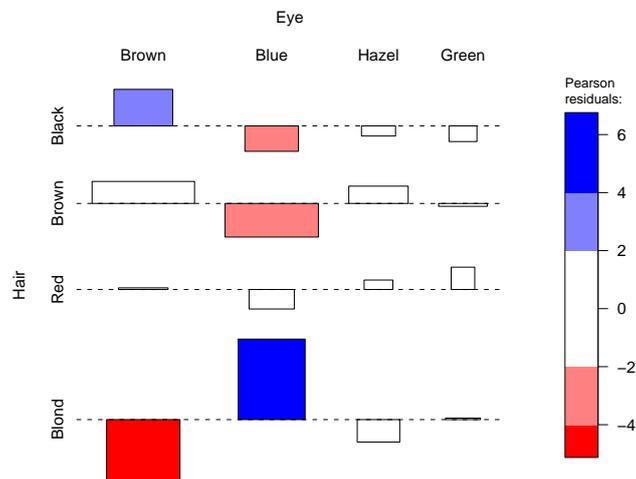Figure 1: Association plot for 'Hair' and 'Eye' colors of female students.



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

# 3 Improved visualization through residual shading

We propose to use a color-shading similar to the enhancements of Friendly (1994) for the mosaic plot (a graphical tool for the visualization of the observed frequencies in a contingency table; see Hartigan and Kleiner, 1981, 1984). This extension uses a color coding of the mosaic tiles to visualize deviations (residuals) from independence. The sign and size of residuals are coded by color shadings of the corresponding rectangles: by default, residuals with an absolute value smaller than 2 are white, positive and negative residuals exceeding an absolute value of 2 are shaded light blue and red respectively, and 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 $\chi^2$ test statistic is not significant at a 5% level: in this example, its value is 6.221, and the corresponding $p$ value is approximately 0.101. 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 see from the legend, the overall range of the residuals is smaller than $[-1.5, 1.5]$.

# 4 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. 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, a new implementation was written from scratch in `grid` offering much more versatility amongst some minor advantages and convenient improvements. It allows the user to specify either 2-way tables of graphics parameters, i.e., arrays 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. This makes it easy for users to implement their own shading schemes; `vcd` provides functions for shading schemes like those described in the previous sections. How association plots can be recycled using the new `grid` implementation is described in the next section.
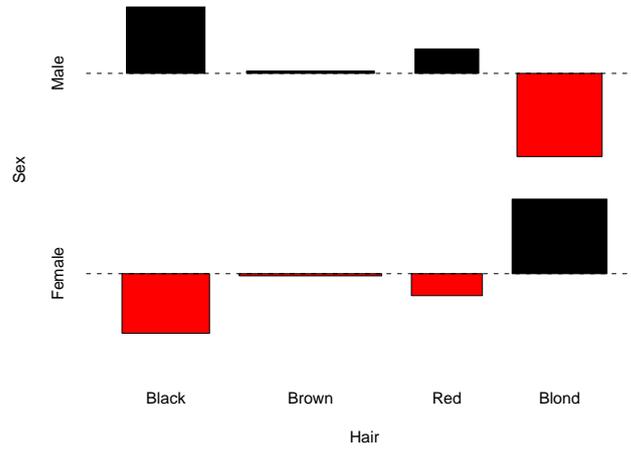
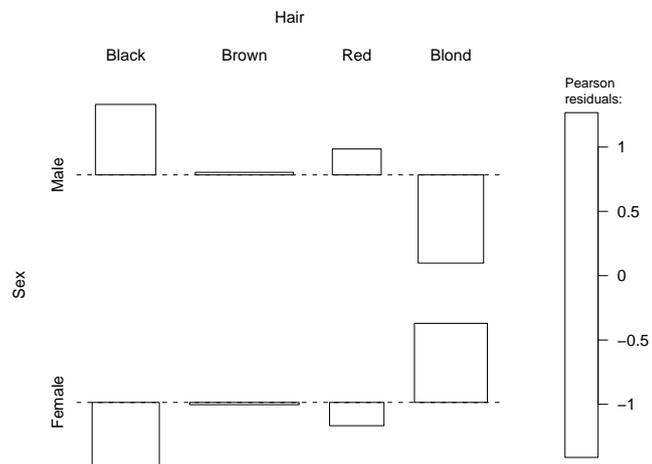Figure 3: Association plot for 'Hair' and 'Sex' in the 'HairEyeColor' data.



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

# 5 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 ($\pi_{ij...}$) 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 (for example the model: $\pi_{ijk} = \pi_{i|k}\pi_{j|k}$). 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 matrix of plots where each cell contains an association plot of the corresponding row and column variables. Such an *association pairs plot* 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.
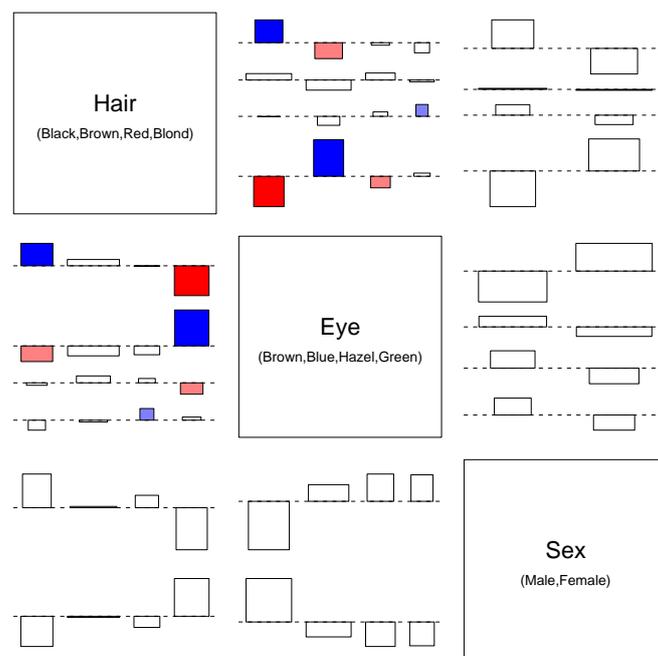


Figure 5: Extended association pairs for the 'HairEyeColor' data, visualizing all pairwise marginal associations.

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 (Bickel, Hammel, and O'Connell, 1975) which is available in base R. In this data, the question whether there is gender discrimination at the university leads

to the result that although women seem to be disadvantaged at the aggregated level there is no gender 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.
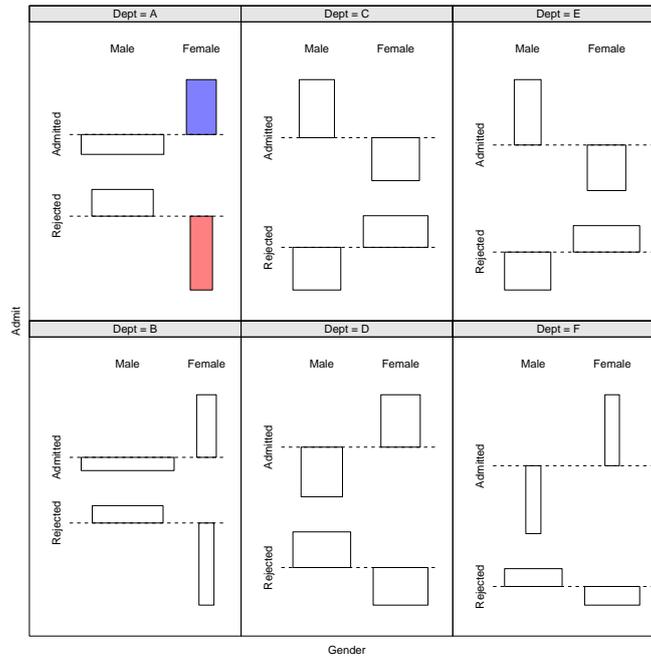


Figure 6: Extended conditioning association plots for the 'UCB Admissions' data.

## 6 Conclusion

We suggest a set of enhancements for visualizing the independence problem in contingency tables using association plots. The extensions aim at improving (a) the visualization by displaying the size of the residuals and (b) the implementation based on the graphics package `grid` which provides more modular design and more flexible specification of graphical parameters. Furthermore, some extensions of association plots to multi-way tables are outlined. 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, if the shading levels were chosen in a way that a true visual test of independence can be performed, that is, that the presence of "significant" residuals leads 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

P. Bickel, E. Hammel, and J. O'Connell. Sex bias in graduate admissions: Data from berkeley. *Science*, 187(398–403), 1975.

A. Cohen. On the graphical display of the significant components in a two-way contingency table. *Communications in Statistics — Theory and Methods*, A9:1025–1041, 1980.

M. Friendly. Mosaic displays for multi-way contingency tables. *Journal of the American Statistical Association*, 89:190–200, 1994.

M. Friendly. *Visualizing Categorical Data*. SAS Insitute, Carey, NC, 2000.

J. Hartigan and B. Kleiner. Mosaics for contingency tables. In W. Eddy, editor, *Computer Science and Statistics: Proceedings of the 13th Symposium on the Interface*, pages 268–273, New York, 1981. Springer.

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

H. Hofmann. Constructing and reading mosaicplots. *Computational Statistics & Data Analysis*, 2003. forthcoming.

P. Murrell. The `grid` graphics package. *R News*, 2(2):14–19, June 2002. URL http://CRAN.R-project.org/doc/Rnews/.

R. Snee. Graphical display of two-way contingency tables. *The American Statistician*, 28:9–12, 1974.

M. Theus and S. R. W. Lauer. Visualizing loglinear models. *Journal of Computational and Graphical Statistics*, 8(3):396–412, 1999.

## Affiliation

David Meyer, Achim Zeileis
Institut für Statistik und Wahrscheinlichkeitstheorie
Technische Universität Wien
Wiedner Hauptstraße 8–10/1071
A-1040 Wien, Austria
E-mail: David.Meyer@ci.tuwien.ac.at, Achim.Zeileis@ci.tuwien.ac.at

Kurt Hornik
Institut für Statistik
Wirtschaftsuniversität Wien
Augasse 2–6
A-1090 Wien, Austria
E-mail: Kurt.Hornik@wu-wien.ac.at