Residual-based Shadings for Visualizing (Conditional) Independence

Achim Zeileis, David Meyer, and Kurt Hornik

Wirtschaftsuniversität Wien, Augasse 2–6, 1090 Wien, Austria
E-mail: `Firstname.Lastname@wu-wien.ac.at`

Visualizing independence by association and mosaic plots.
Enhancements: diverging color palette based on HCL color space combined with visualization of the result of a significance test.
Extensions to visualizing conditional independence.

Key Words: Association plots; Conditional inference; Contingency tables; HCL colors; HSV colors; Mosaic plots; Permutation tests.

1. INTRODUCTION

independence in 2-way tables
association and mosaic displays
colors and color spaces
somewhere: mention Augsburg stuff, visualization of log-linear models via mosaics (Theus and Lauer 1999; Hofmann 2003, 2001).
Association plots as residual plots for log-linear models (Meyer, Zeileis, and Hornik 2003), especially in coplot or trellis layout.
Arthritis data (Koch and Edwards 1988)
UCB admissions data (Bickel, Hammel, and O'Connell 1975)

2. INDEPENDENCE IN 2-WAY TABLES

2.1 Tests

We consider a 2-way contingency table 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

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

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.

But this is not the only plausible way of aggregation of the Pearson residuals. There are many conceivable functionals $\lambda(\cdot)$ which lead to reasonable test statistics $\lambda(\{r_{ij}\})$, the sum of squares is just one of them. Another functional suitable for identifying the cells which cause the dependence (if any) is the maximum of the absolute values

$$M \quad = \quad \max_{i,j} |r_{ij}|. \tag{4}$$

Given a critical value c_α for this test statistic, all residuals whose absolute value exceeds c_α violate the hypothesis of indendence at level α (Mazanec and Strasser 2000, ch. 7). Thus, the interesting cells causing the dependence can easily be identified.

Furthermore, the main reason for using the unconditional limiting distribution for the χ^2 statistic (3) was the closed form result for the distribution. Recently, with the improving perfomance of computers, performing permutation tests—either by simulation or by computation of the permutation distribution—became more and more popular. In particular for the independence hypothesis (1), using a permutation test is very intuitive due to the permutation invariance (given row and column sums) of this problem. By employing this approach the permutation distribution of statistics of type $\lambda(\{r_{ij}\})$ (also including the χ^2 statistic) can be derived.

Other classical tests for the hypothesis of independence not fitting in the Pearson residual-based framework described above include Fisher's exact test, asymptotic tests of the odds ratio (for 2×2 tables) and the Mantel-Haenzel test.

2.2 Visualizations

ONLY SHAPE! not color

The two best known visualization techniques for independence in 2-way tables are association plots and mosaic plots. Both are suitable to bring out departures of an observed table (n_{ij}) from the expected table ($\hat{n}_{ij}$) in a graphical way. The former focuses on the visualization of the Pearson residuals r_{ij} (under independence) while the latter primarily displays the observed frequencies n_{ij}.

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.

Mosaic Plots (Hartigan and Kleiner 1981) can be seen as an extension of grouped bar charts where width and height of the bars show the relative frequencies of the two variables: a mosaic plot simply consists of a collection of tiles whose sizes are proportional to the observed cell frequencies (see Figure **??**).

Sequential horizontal and vertical recursive splits are used to visualize the frequencies of more than two variables, each new variable conditional to the previously entered variables. A first extension by Friendly (1994) uses a color coding of the tiles to visualize deviations (residuals) from a given log-linear model fitted to the table, that is, from the expected frequencies under independence. This approach does not only work in 2-way tables but also in log-linear models fitted to multi-way tables. In this extension, positive and negative signs of the residuals are coded by rectangles with solid and dashed borders respectively. Furthermore, 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). In particular, this shading does not provide a visualization of the maximum test (4) as in general it is unknown to which significance level α the values $c_\alpha = 2$ and 4 correspond for the table being visualized.

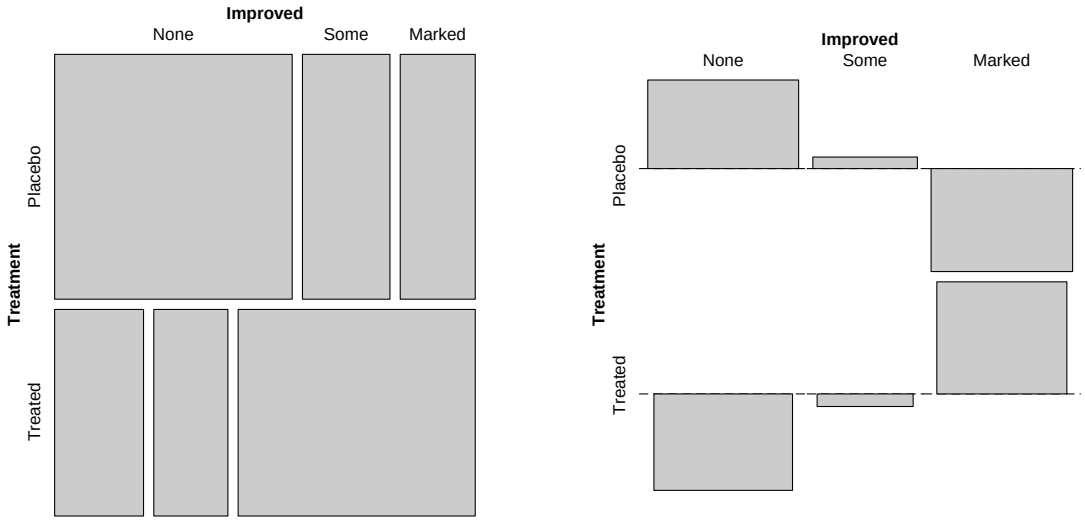

Figure 1: Classic mosaic and association plot for the Arthritis data.

3. RESIDUAL-BASED SHADINGS

remarks: Friendly 1994

3.1 Colors

currently: chosen based on HSV or HLS color spaces, hue (red and blue) codes sign of residuals, and saturation the absolute size in three steps, value currently unused (always set to full value)

disadvantages: - perceptually not uniform because the three perceptual dimensions (hue, lightness, saturation) are not properly mapped to the three dimensions of the color space and hence are confounded (Brewer 1999) in particular, saturation is not uniform across different hues.

- can lead to color-caused optical illusions in statistical graphs (**?**)

- flashy fully saturated HSV colors are good for drawing attention to a plot, but hard to look at for a longer time, hence mosaic displays using these colors are harder to interpret

- white is not very suitable as a neutral color, grey conveys neutrality or uninterestingness much better

Alternatives: - Munsell (1905) introduced color notation for balanced colors - Commission Internationale de l'Éclairage (CIE) introduced two perceptually based color spaces CIELAB and CIELUV where the latter is preferred for emissive color technologies such as computer displays. - **ColorBrewer.org** (Harrower and Brewer 2003) is an online tool for selcting different sets of

colors, based on similar ideas. It includes qualitative, sequential and diverging palettes. - Ihaka (2003) discusses how CIELUV colors can be used for choosing colors for statistical graphics such as barplots. He recommends to use polar coordinates in the CIELUV space—called HCL space, using hue, chroma and luminance—and suggests a strategy for choosing qualitative schemes based on these.

We discuss how similar ideas can be used for deriving a diverging HCL scheme that is suitable for visualizing residuals in association and mosaic displays.

HCL - irregular double cone, i.e., chroma/luminance plane looks rather different for different hues - choose two hues that look similar in the chroma/luminance plane to code the sign of the residuals. Also use red and blue at full chroma for large residuals and use a light gray (instead of a white at the top of the luminance scale) as the neutral color for small residuals, interpolate between. - for non-significant results, clearly reduce the amount of color, i.e. the maximum chroma used.

Advantages - based on a perceptually uniform color space - device independent

3.2 Significance

The extensions of Friendly (1994) to mosaic plots provide substantial improvement of the original mosaic plots and enhanced them from a plot for contingency tables to a visualization technique for log-linear models (and therefore also for certain independence problems). However, it has two major drawbacks: First, the significance level for the hard-coded critical values 2 and 4 is usually unknown. Second, the colors for (light) blue and red used in the Michael Friendly's SAS implementation and also in the base implementation in R are not device-independent and do not provide homogeneous saturations over different colors and copier proofness. We suggest some ways to overcome these drawbacks.

As pointed out in Section 2.1, the critical values for the maximum statistic M from (4) can be derived from the permutation distribution. Instead of the hard-coded values 2 and 4, the particular critical values, e.g., at the levels $\alpha = 0.1$ and 0.01, for the table to be visualized could be used. Hence, exactly those residuals causing the (potential) dependency within the table are highlighted. At the moment, these critical values are derived in `vcd` by simulation of the underlying distribution.

The implementation of Friendly mosaic plots in SAS uses (by default) colors in the HLS (Hue–Luminance–Saturation) color space with hues for blue and red, full saturation, and varying luminance for lighter colors. The current R implementation is based on colors in the HSV (Hue–Saturation–Value) space where decreasing the saturation from 1 towards 0 generates very similar colors as increasing the luminance for HLS colors from 0.5 towards 1. As an alternative to those HLS and HSV colors, the device-independent HCL (Hue–Chroma–Luminance) color space (Ihaka 2003) could be used which also provides homogeneous saturations over different colors and copier proofness. The resulting association plot (using critical values for $\alpha = 0.1$ and 0.01) for the Bundesliga data can be seen in Figure **??**. It indicates very clearly that there is no evidence for rejecting the hypothesis of independence in this table. Thus, there seems to be neither an advantage for the home team nor for the away team.

To illustrate what a 'postive' example looks like we give the association plot for the well-known Hair-Eye-Color data set (avilable in base R) which cross-tabulates the hair and eye color and gender of 592 statistics students: Figure **??** shows the association plot for the 328 female students. It is easy to see that the hypothesis of independence is rejected at 1% level by the maximum test (4) due to higher frequencies of students with hair and eye color black/brown and blond/blue and lower frequencies of blond/brown and brown/blue than would be plausible under independence. Furthermore, it can also be seen that the residual corresponding to the black/blue cell is significant at the 10% level.

We have seen that this approach performs very well when the test statistic which should be visualized is the maximum statistic (4). However, it is difficult to extend this approach to general tests for independence, in particular in the HCL space, where the ranges of chroma and luminance are not independent. In contrast the HSV space has the advantage that all three dimensions hue,

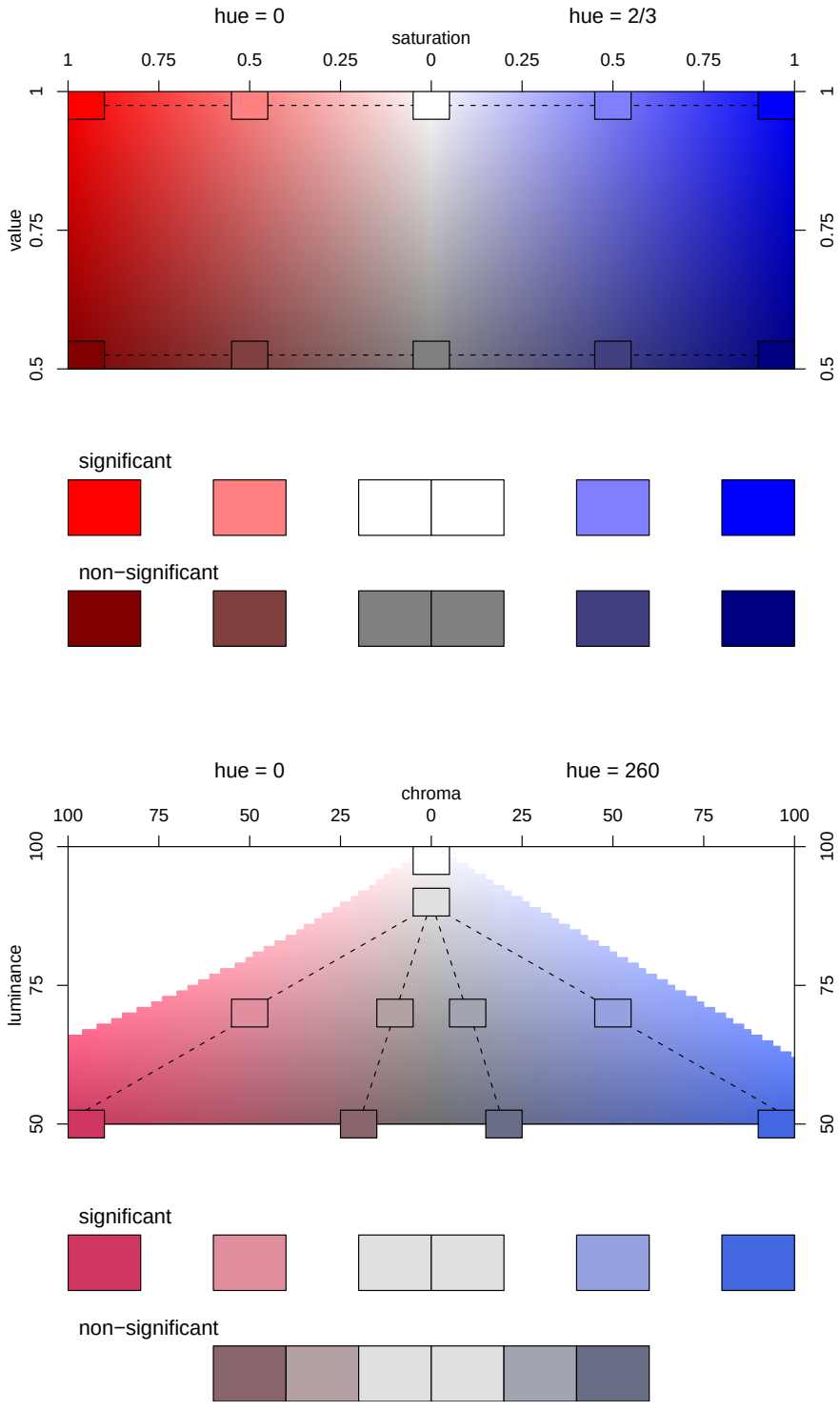

Figure 2: Extended shading in HSV and HCL space.

saturation and value have the range $[0, 1]$ and can be varied independently. Therefore, a different approach than the one described above could be to use the hue for the sign of the residuals as before, the saturation for the absolute size of the residuals, and the value as an indicator for the significance of some test statistic (only using the full value when the overall test for independence rejects the null hypothesis). The resulting mosaic plots under this paradigm are depicted in Figure **??**. Again, it can clearly be seen that despite some large residuals there is no evidence against independence for the Bundesliga data, but that the null hypothesis has to be rejected for the Hair-Eye-Color data.

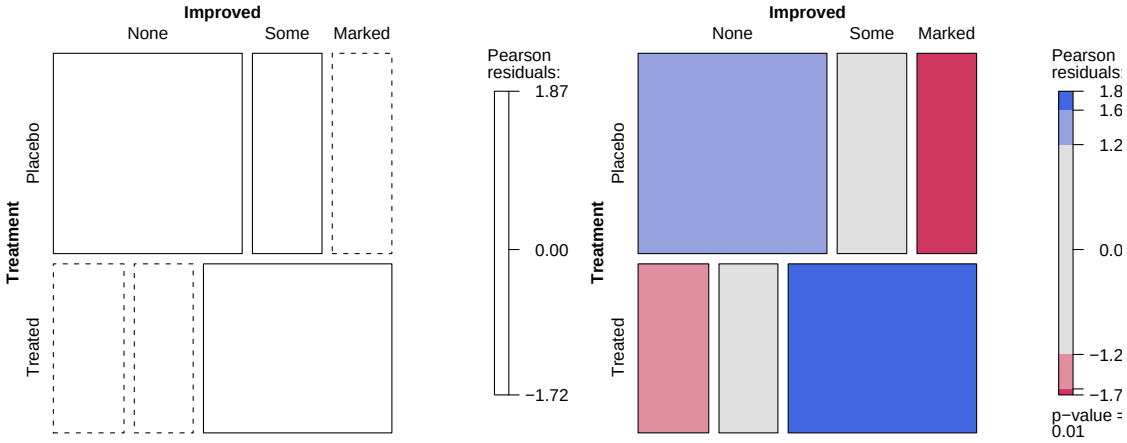

Figure 3: Mosaic plot for the Arthritis data with Friendly shading (left) and extended shading (right).

4. EXTENSIONS TO CONDITIONAL INDEPENDENCE

Friendly (1999) discusses grouping similar to coplots (conditioning plots) (Cleveland 1993) that lead to trellis graphics (Becker, R.A., Cleveland, W.S. and Shyu, M. "The Visual Design and Control of Trellis Display", Journal of Computational and Graphical Statistics).

Independence problems do not only occur in 2-way tables, although that is an important special case, but they are also important in tables of higher dimensionality and 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. Models of interest include the null hypotheses of conditional independence:

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

Classical non-graphical methods for these problems include the χ^2 test, Fisher's exact test, the Cochran-Mantel-Haenzel test (for $2 \times 2 \times K$-tables), and the analysis of log-linear models for more complex settings.

As an example, two natural ways to use the visualization techniques described in the previous sections would be to use (Trellis-like) conditioning plots or pairs plots (like mosaic matrices) to visualize these more complex patterns of independence.

Two ideas for the problem of conditional independence are briefly outlined here and 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 gender discrimination at the UCB leads to the result that although women seem to be disadvantaged at the aggregated level there is no gender discrimination conditioned on the department—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 assocation plot in Figure **??**.

Similarly, the same data can be visualized using a mosaic matrix where a conditional independence model is fitted in each plot (see Figure **??**).

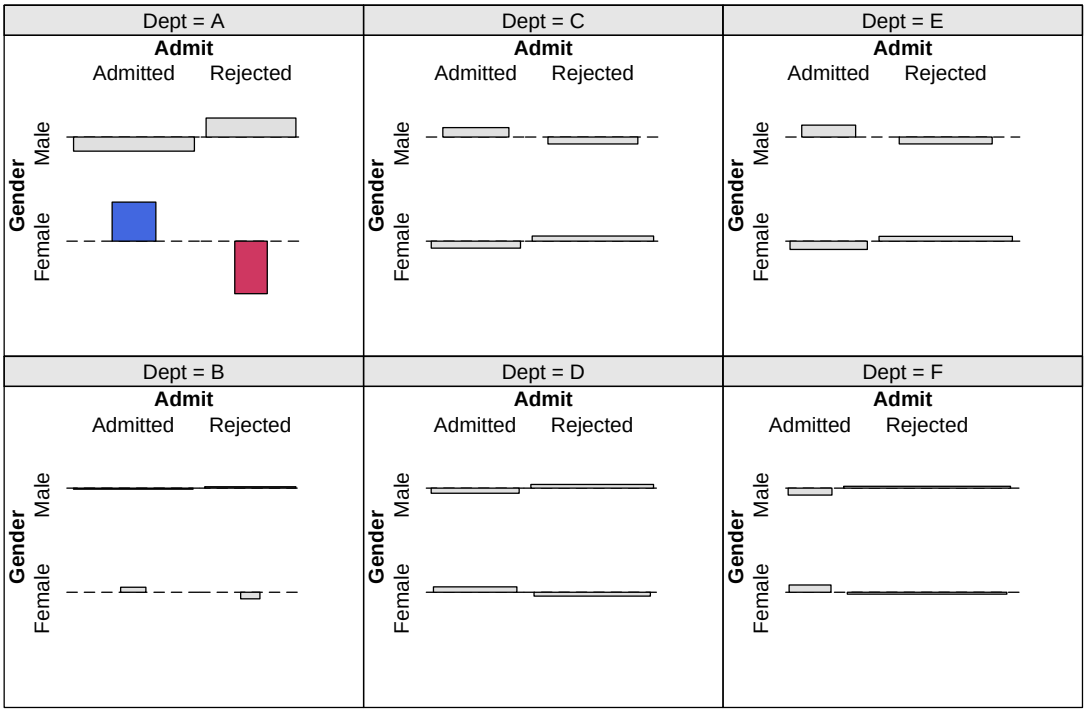

Figure 4: Conditional association plot for UCB admissions.

5. CONCLUSIONS

REFERENCES

Bickel, P., Hammel, E., and O'Connell, J. (1975), "Sex Bias in Graduate Admissions: Data from Berkeley," *Science*, 187.

Brewer, C. A. (1999), "Color Use Guidelines for Data Representation," in *Proceedings of the Section on Statistical Graphics, American Statistical Association*, Alexandria, VA, pp. 55–60.

Cleveland, W. S. (1993), *Visualizing Data*, Summit, New Jersey: Hobart Press.

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.

— (1999), "Extending Mosaic Displays: Marginal, Conditional, and Partial Views of Categorical Data," *Journal of Computational and Graphical Statistics*, 8, 373–395.

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

Harrower, M. A. and Brewer, C. A. (2003), "**ColorBrewer.org**: An Online Tool for Selecting Color Schemes for Maps," *The Cartographic Journal*, 40, 27–37.

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

Hofmann, H. (2001), "Generalized Odds Ratios for Visual Modelling," *Journal of Computational and Graphical Statistics*, 10, 1–13.

— (2003), "Constructing and Reading Mosaicplots," *Computational Statistics & Data Analysis*, 43, 565–580.

Ihaka, R. (2003), "Colour for Presentation Graphics," in *Proceedings of the 3rd International Workshop on Distributed Statistical Computing, Vienna, Austria*, eds. Hornik, K., Leisch, F., and Zeileis, A., ISSN 1609-395X.

Koch, G. and Edwards, S. (1988), "Clinical Efficiency Trials with Categorical Data," in *Biopharmaceutical Statistics for Drug Development*, ed. Peace, K. E., New York: Marcel Dekker, pp. 403–451.

Mazanec, J. A. and Strasser, H. (2000), *A Nonparametric Approach to Perceptions-Based Market Segmentation: Foundations*, Berlin: Springer.

Meyer, D., Zeileis, A., and Hornik, K. (2003), "Visualizing Independence Using Extended Association Plots," in *Proceedings of the 3rd International Workshop on Distributed Statistical Computing, Vienna, Austria*, eds. Hornik, K., Leisch, F., and Zeileis, A., ISSN 1609-395X.

Munsell, A. H. (1905), *A Color Notation*, Boston, Massachusetts: Munsell Color Company.

Theus, M. and Lauer, S. R. W. (1999), "Visualizing Loglinear Models," *Journal of Computational and Graphical Statistics*, 8, 396–412.