Residual-based Shadings for Visualizing (Conditional) Independence

Achim Zeileis, David Meyer, and Kurt Hornik

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 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)

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.

2. INDEPENDENCE IN 2-WAY TABLES

In this section, the basic tools for testing and visualizing independence in 2-way tables are briefly reviewed. For illustration, a data set about treatment and improvement of patients with rheumatoid arthritis from Koch and Edwards (1988) is used. The data set is also discussed in Friendly (2000) and the subset of the 59 female patients from the study is given in Table 1.

2.1 Tests

To fix notations, we consider a 2-way contingency table with cell frequencies $\{n_{ij}\}$ for $i = 1, \dots, I$ and $j = 1, \dots, J$ 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

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

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

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

Table 1: Treatment and improvement among 59 patients with rheumatoid arthritis.

		Improved		
		None	Some	Marked
Treatment	Placebo	19	7	6
	Treated	6	5	16

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 an unconditional limiting χ^2 distribution with $(I-1)(J-1)$ degrees of freedom under the null hypothesis. This is the well-known χ^2 test which is typically introduced at the very beginning of the chapter about independence in 2-way tables in statistics textbooks (see e.g., Agresti 2002).

But the sum of squares is not the only plausible way of capturing deviations from 0 in the Pearson residuals. There are many other conceivable functionals $\lambda(\cdot)$ which lead to reasonable test statistics $\lambda(\{r_{ij}\})$: one which is particularly suitable for identifying the cells which cause the dependence (if any) is the maximum of the absolute values

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

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

Furthermore, an important reason for using the unconditional limiting distribution for the X^2 statistic (3) was the closed form result for the distribution. Recently, with the improving perfomance of computers, conditional inference (or permutation tests)—carried out either by simulation or by computation of the (asymptotic) permutation distribution—have been receiving increasing attention (Pesarin 2001; Strasser and Weber 1999). For testing the independence hypothesis (1), using a permutation test is particularly intuitive due to the permutation invariance (given row and column sums) of this problem. Consequently, all results in this paper are based on conditional inference performed by simulating from the permutation distribution of test statistics of type $\lambda(\{r_{ij}\})$.

Other measures of discrepancy (e.g., deviance residuals) could, of course, also be used instead of the the Pearson residuals $\{r_{ij}\}$. But as the ideas discussed here extend straightforwardly to that situation, we do not discuss it in detail.

For the arthritis data from Table 1, both tests indicate a clearly significant dependence of improvement on treatment: the sum-of-squares statistic from Equation (3) is $X^2 = 11.296$ with a p value of $p = 0.0032$, and the maximum statistion from Equation (4) is $M = 1.87$ with a p value of $p = 0.0096$. Both p values have been computed by drawing a sample of size 5000 from the permutation distribution under independence.

2.2 Visualizations

Two well-established visualization techniques for independence in 2-way tables are mosaic plots and association 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 latter focuses on the visualization of the Pearson residuals r_{ij} (under independence) while the latter primarily displays the observed frequencies n_{ij}.

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 as shown in the left panel of Figure 1. A rectangle corresponding to 100 percent of the observations is first split horizontally with respect to the treatment frequencies and then vertically with respect to the conditional improvement frequencies. This shows that there have been more placebo than treated patients with no improvement and vice versa for marked improvement. This strategy of splitting with respect to conditional frequencies given all previous variables can also directly be used for visualizing multi-way tables (see Hofmann 2003, for an overview of how to construct and read mosaic displays).

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 association plot for the arthritis data is shown in the right panel of Figure 1 which leads to the same interpretation as the mosaic plot: there are more placebo patients with no improvement and fewer with marked improvement than expected under independence—vice versa for the treated patiens.

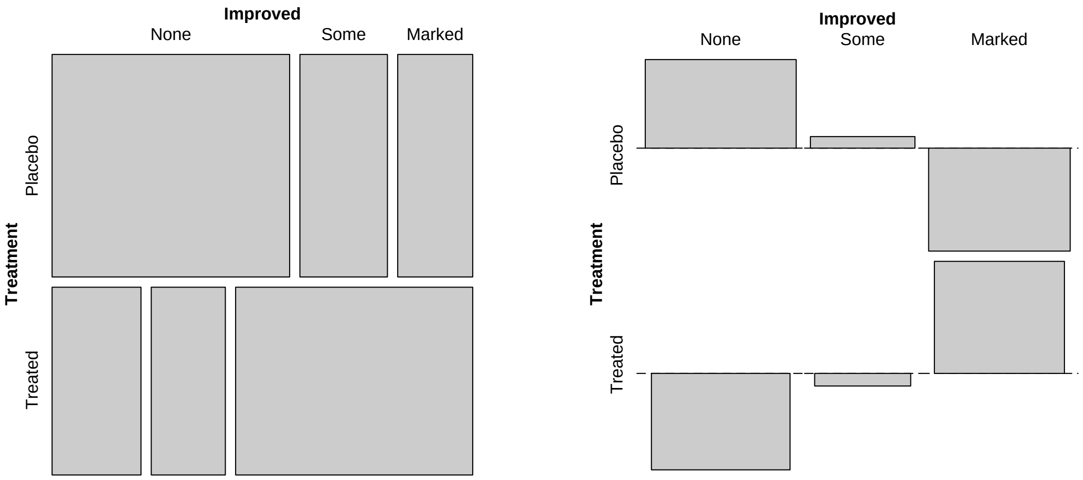

Figure 1: Classic mosaic and association plot for the Arthritis data.

3. RESIDUAL-BASED SHADINGS

Colors are commonly used to enhance mosaic and association plots. To integrate a visualization of the Pearson residuals $\{r_{ij}\}$ into the mosaic display—which in its 'raw' version only visualizes the observed frequencies $\{n_{ij}\}$—Friendly (1994) suggested a residual-based shading for the mosaic tiles that can also be applied to the rectangles in association plots (Meyer et al. 2003). In this section, we first briefly review the Friendly (1994) shading, before we suggest different colors and a combination of visualization and significance testing to extend these residual-based shadings.

3.1 Friendly Shading

The extensions of Friendly (1994) to mosaic plots provide a 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 thus also for independence problems including 2-way tables as the simplest special case.

The idea is to use a color coding for the mosaic tiles that visualizes the sign and absolute size of each residual r_{ij}: Cells corresponding to small residuals ($|r_{ij}| < 2$) are shaded white. Cells with medium sized residuals ($2 \leq |r_{ij}| < 4$) are shaded in light blue and light red for positive negative residuals respectively. Cells with large residuals ($|r_{ij}| \geq 4$) are shaded with a fully saturated blue and red respectively. Mosaic plots enhanced by this shading can thus also bring out departures from independence (or other log-liner models in multi-way tables) graphically and visualize patterns of dependence. The heuristic for choosing the cut offs 2 and 4 is that the Pearson residuals are approximately standard normal which implies that the highlighted cells are those with residuals *individually* significant at approximately the $\alpha = 0.05$ and $\alpha = 0.0001$ 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 addition to the shading of the rectangles themselves, the Friendly shading also encompasses a choice of line type and line color of the borders of the rectangles with similar ideas as described above. As both mosaic and association plots are area-proportional visualization techniques, we focus on area shadings and always use solid black borders throughout this paper, but the extensions suggested in the following could also be applied to control line type and color.

3.2 Colors

The way the (light) blue and red colors are chosen differs somewhat between various implementations of Friendly mosaic plots: In his original SAS implementation (see Friendly 2000), Michael Friendly uses colors from a palette based on HLS (Hue–Luminance–Saturation) color space. The implementation in the R system for statistical computing (R Development Core Team 2005, http://www.R-project.org/) employs colors from HSV (Hue–Saturation–Value) space. The latter is a very common implementation of colors in many computer packages and makes the generation of the Friendly shading very simple.

The HSV space originally looks like a regular cone with black at its peak (zero value) and full color wheels for different saturations at the other end, centered around a white (full value). Type and amount of color are controlled by hue and saturation respectively. Typically, polar coordinates that all range in the unit interval are used in this space, such that it looks like a regular cylinder. For generating colors in the Friendly shading, the following strategy is used: The hue h codes the sign of the residuals—$h = 0$ (red hue) is used for negative residuals, $h = 2/3$ (blue hue) for positive residuals. The absolute size of the residuals is then coded by the saturation s which is set to 0, 0.5 and 1 respectively for small/medium/large residuals respectively. The value is always fixed at $v = 1$. This is also depicted in the upper panel of Figure 2 which shows to slides from the HSV cylinder side to side. The plot shows the saturation/value plane for the given hues $h = 0$ and $h = 2/3$. The significant color palette shows the colors used for the Friendly shading when the residuals are increasing from left to right. The set of non-significant colors will be explained in the next section.

Although this HSV-based shading is already very useful for enhancing mosaic and association plots and although HSV is very commonly available implementation of color spaces, HSV colors in general and the Friendly shading in particular have a number of disadvantages. Most importantly, HSV colors are not perceptually uniform because the three HSV dimensions map only poorly to the three perceptual dimensions of the human visual system (Brewer 1999; Ihaka 2003). Consequently, the the HSV dimensions are confounded, e.g., saturation is not uniform across different hues. A fully saturated blue $(2/3, 1, 1)$ is much darker as a fully saturated red $(0, 1, 1)$ or green $(1/3, 1, 1)$. This makes it more difficult for the human eye to judge the size of shaded areas and can therefore lead to color-caused optical illusions when used in statistical graphs (Cleveland and McGill 1983). Furthermore, flashy fully saturated HSV colors are good for drawing attention to a plot, but hard to look at for a longer time (Ihaka 2003) which makes HSV-shaded graphics harder to interpret. Finally, white is employed as the neutral color for small residuals in the Friendly shading, but typically grey is found to convey neutrality or uninterestingness much better than white does

(Brewer 1999).

Alternative ways to choose colors have been available for a long time, but have been only slowly adopted for implementations of colors in computer packages in general and for shading in statistical graphs in particular. The idea of using perceptually based colors that are 'in harmony' go back to Munsell (1905) who introduced a color notation for balanced colors. Based on similar principles, Cynthia Brewer and co-workers suggested different types of palettes (qualitative/sequential/diverging) and provided the online tool **ColorBrewer.org** (Harrower and Brewer 2003) for selecting an appropriate palette for a specific problem. Furthermore, the Commission Internationale de l'Éclairage (CIE) introduced the two perceptually based color spaces CIELAB and CIELUV where the latter is typically preferred for emissive color technologies such as computer displays. Ihaka (2003) discusses how CIELUV colors can be used for choosing colors for statistical graphics such as barplots. By taking polar coordinates in CIELUV space, it is called HCL (Hue–Chroma–Luminance) space and qualitative palettes can easily be chosen by using a range of hues for fixed values of chroma and luminance. Such colors are always balanced towards the same grey and thus do not have the problem of varying saturations as for the HSV colors.

Here, we discuss how similar ideas can be used for deriving a diverging HCL palettes that provide a suitable translation of the ideas from the Friendly shading to perceptually uniform HCL colors. The HCL space looks like a double cone with black (zero luminance) at one end and white at the other (full luminance). In its middle, there is a full color wheel for different values of chroma (that controls the colorfulness). Unfortunately, the HCL space is not completely regular as the HSV space is and consequently its dimensions cannot be standardized to unit intervals: the hue ranges in $[0, 360]$ degrees and chroma and luminance both in $[0, 100]$ percent. But not all combinations (h, c, l) yield valid HCL colors and the admissable combinations of c and l vary across different hues h. But for the problem of constructing a diverging palette, this problem can easily be overcome as we just need two different hues (a 'negative' and a 'positive' hue) and hence choose two hues that correspond to similar shapes in the chroma/luminance plane. The lower panel of Figure 2 shows two such planes side to side for the hues $h = 0$ and $h = 260$. To obtain a sequence of colors with the same properties as the Friendly shading, the palette starts at a fully saturated red $(0, 100, 50)$, goes via a neutral color, ends at a fully saturated blue $(260, 100, 50)$, and uses linear interpolation in between. Instead of using white $(0, 0, 100)$ as the neutral color, a light grey $(0, 0, 90)$ is employed as motivated above.

This diverging palette (see the 'significant' colors in Figure 2, the 'non-significant' colors are again discussed in the following section) uses both chroma, i.e., the colorfulness, and luminance, i.e., the amount of grey, to code the absolute size of the quantity visualized, i.e., the Pearson residuals when applied to the independence problem. By changing the neutral color or by changing the maximum chroma respectively, this can be changed to using only chroma or luminance for this purpose, but using both seems to be a very effective way of visualization.

The advantages of using this HCL shading scheme instead of the HSV scheme is that it is based on a perceptually based color space that provides uniform saturations for different hues and is device inependent.

3.3 Significance

The shading scheme of Friendly (1994) was suggested to visualize the pattern of dependence in contingency tables, as discussed above, but the presence (or absence) of colors in a plot also always conveys an impression of interestingness (or uninterstingness respectively). That is, statisticians might be tempted to interpret the absence of color in a plot as a clue that there is no significant departure from independence. Or vice versa, colored cells would convey the impression that there is significant dependence. Currently, both is not true as can be seen in the left panel of Figure 3 which shows the mosaic display for the arthritis data with Friendly shading. Although there is significant dependence, no residual exceeds an absolute value of 2 and hence no cell is colored. Of course, it can be argued that the shading was not designed for this purpose and that different cut offs than 2 and 4 should be used here. But then again, it would be nice if these cut offs could be chosen automatically in a data-driven fashion which is what we do below.

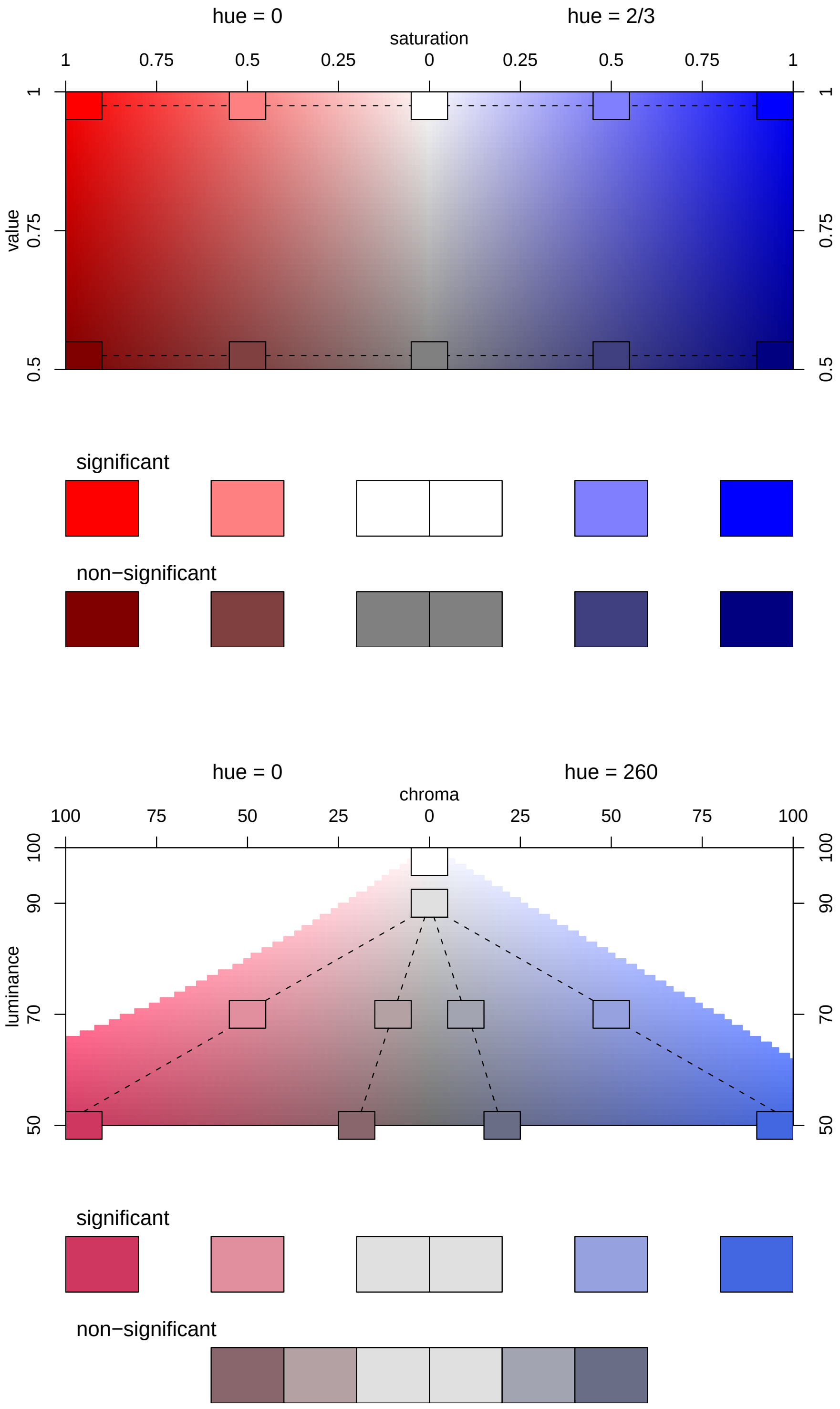

Figure 2: Extended shading in HSV (upper) and HCL space (lower).

The Friendly shading can be interpreted to be a visualization of the maximum statistic M (4) which always employs the critical values c_α of 2 and 4 respectively. But the problem is that it is not clear to which significance levels α these critical values correspond because the distribution of M depends on the underlying contingency table. The natural solution to this problem is to compute the critical values from the distribution of M in a data-driven way (i.e., for the table visualized) and use these instead of the hard-coded values 2 and 4. In the right panel of Figure 3 this is done for the arthritis data by employing the critical values 1.24 at level $\alpha = 0.1$ and 1.64 at level $\alpha = 0.01$ (and using the diverging HCL palette). By using these cut offs, the presence of color in the plot is equivalent to significance (of the maximum statistic M) at level $\alpha = 0.1$ and $\alpha = 0.01$ respectively and exactly the cells which 'cause' the dependence are highlighted. For the arthritis data, these are in particular the cells in the last column that signal that there are significantly more treated patients and fewer placebo patients with marked improvement than would be expected under independence.

The significance $alpha = 0.1$ and 0.01 are chosen because this leads to displays where fully colored cells are clearly significant (i.e., with a p value below 0.01), cells without color are clearly non-significant (i.e., with a p value above 0.1), and cells in between can be considered to be weakly significant (i.e., $0.1 \leq p < 0.01$). Of course, users could choose any other set of significance levels they feel comfortable with, e.g., only a single cutoff at $\alpha = 0.05$ or three cut offs at 0.1, 0.05 and 0.01 etc. Another option could be to use a continuous shading where the p value corresponding to a cell controls the interpolation between the neutral and the full color. However, our experience is that this typically results in too much color in a plot which in turn tends to conceal the important cells and over-emphasizes the unimportant ones. Hence, we stick to a shading with fewer colors that are easier to interpret.

This extended shading is already very flexible and combines visualization and inference, but it should only be applied when using the maximum statistic is appropriate. For the sum-of-squares statistic X^2 (and all other functionals $\lambda(\cdot)$), this shading cannot be used.

use pre-determined (or manually adjusted) cut offs, but visualize the significance by the amount of color in the plot, i.e., use only little color if the corresponding test result is non-significant.

For HCL, the amount of color can easily be controlled by varying the maximum chroma value. So far 100, now 20.

For HSV, a similar effect can be achieved by varying the value dimension which was not used up to now. Setting value to $v = 0.5$ a darker palette of colors can be obtained.

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, 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.

4. EXTENSIONS TO CONDITIONAL INDEPENDENCE

UCB admissions data (Bickel, Hammel, and O'Connell 1975) double max: 3.134, p-value = 0.0004 max Chisq: 17.248, p-value = 0.0006 sum Chisq: 19.938, p-value = 0.0034 simulated from 5000 repetitions

Friendly (1999) discusses grouping similar to coplots (conditioning plots) (Cleveland 1993) that lead to trellis graphics (Becker, Cleveland, and Shyu 1996).

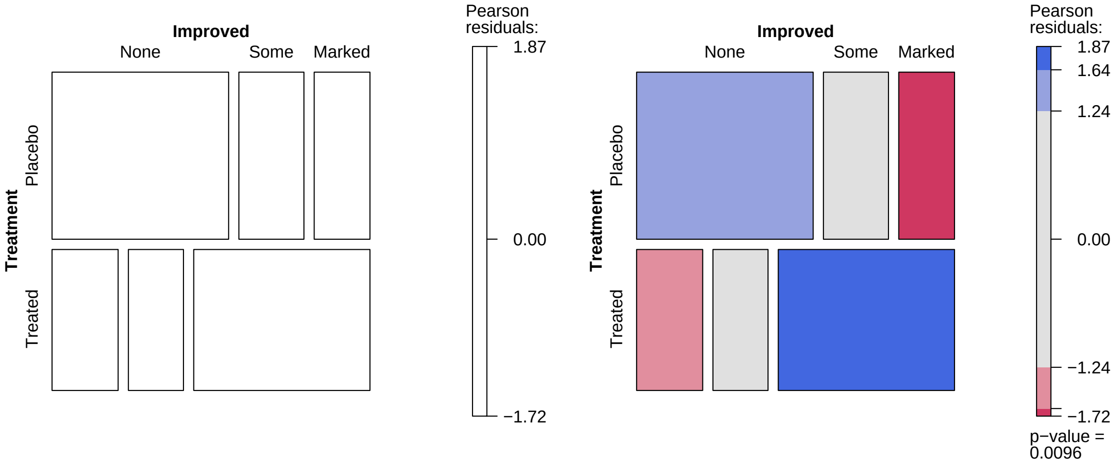

Figure 3: Mosaic plot for the Arthritis data with Friendly shading (left) and extended shading (right).

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 ($\pi_{ij\dots}$) with more than two dimensions. Models of interest include the null hypotheses of conditional independence:

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

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 **??**).

5. CONCLUSIONS

REFERENCES

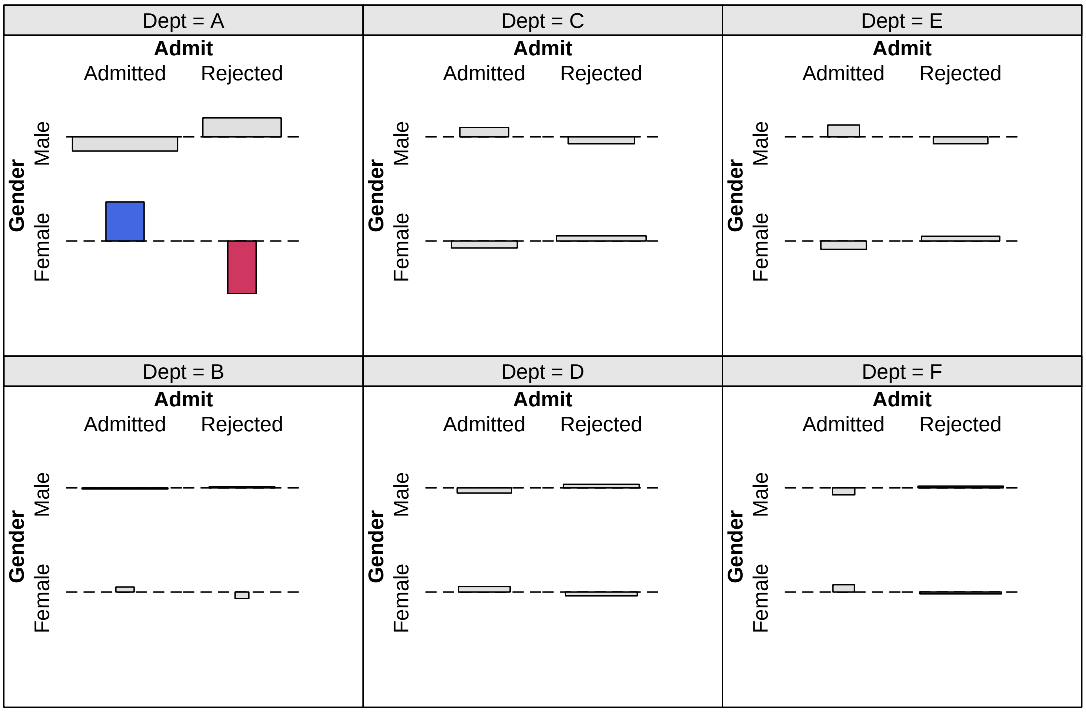

Figure 4: Conditional association plot for UCB admissions.

Agresti, A. (2002), *Categorical Data Analysis*, Hoboken, New Jersey: John Wiley & Sons, 2nd ed.

Becker, R. A., Cleveland, W. S., and Shyu, M.-J. (1996), "The Visual Design and Control of Trellis Display," *Journal of Computational and Graphical Statistics*, 5, 123–155.

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.

Cleveland, W. S. and McGill, R. (1983), "A Color-Caused Optical Illusion on a Statistical Graph," *The American Statistician*, 37, 101–105.

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.

Pesarin, F. (2001), *Multivariate Permutation Tests*, Chichester: John Wiley & Sons.

R Development Core Team (2005), *R: A Language and Environment for Statistical Computing*, R Foundation for Statistical Computing, Vienna, Austria, iSBN 3-900051-00-3.

Strasser, H. and Weber, C. (1999), "On the Asymptotic Theory of Permutation Statistics," *Mathematical Methods of Statistics*, 8, 220–250.

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