Residual-based Shadings for Visualizing (Conditional) Independence

Achim Zeileis, David Meyer, and Kurt Hornik

Residual-based shadings for enhancing mosaic and association plots to visualize independence models for contingency tables are extended in two directions: (a) perceptually uniform Hue-Chroma-Luminance (HCL) colors are used and (b) the result of an associated significance test is coded by the appearance of color in the visualization. For obtaining (a), a general strategy for deriving diverging palettes in the perceptually-based HCL space is suggested. As for (b), cut offs that control the appearance of color are computed in a data-driven way based on the conditional permutation distribution of maximum-type test statistics. The shadings are first established for the case of independence in 2-way tables and then extended to more general independence models for multi-way tables, including in particular conditional independence models.

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

1. INTRODUCTION

Relationships between categorical variables are typically analyzed based on the underlying contingency tables that can be explored for (in)dependence. Two standard methods from the statistical tool box are log-linear models (see, e.g., Agresti 2002) for modeling (in)dependence structures and mosaic plots (Hartigan and Kleiner 1981) for bringing them out graphically. Both methods can also be combined such that a certain mosaic plot visualizes a particular log-linear model by controlling splitting order and direction and shading of the tiles in the mosaic display (Friendly 1994, 1999; Theus and Lauer 1999; Hofmann 2001). In particular, Friendly (1994) suggested a shading strategy based on the residuals (typically, Pearson or deviance residuals) of the associated log-linear model that elevates the mosaic plot from a display for frequencies in a contingency table to a visualization technique that encompasses both observed frequencies and residuals. This shading allows for judging the quality of a model fit and spotting dependence patterns that have not been accounted for by the model. Other techniques for visualizing dependence in contingency tables such as association plots (Cohen 1980) can also be enhanced by using this residual-based shading.

In this paper, the shading of Friendly (1994) is extended in two directions: usage of perceptually-based HCL (Hue-Chroma-Luminance) colors and combination of visualization and significance testing. Friendly's shading is typically implemented in statistical software using color spaces such as HLS (Hue-Luminance-Saturation) or HSV (Hue-Saturation-Value) colors. The dimensions of both spaces are only poorly mapped to the perceptual dimensions of the human visual system (Brewer 1999; Ihaka 2003) which makes it more difficult to properly read and interpret the corresponding plots. For overcoming this problem, a general strategy for constructing diverging palettes in the perceptually-based HCL space (Ihaka 2003) is derived. To couple the visualization and the independence model of a contingency table more tightly than in previous approaches, the residual-based shading is extended such that appearance of color in the display is equivalent to significance of the associated independence test. This is achieved by using data-driven cut offs for the appearance of color computed from the conditional permutation distribution (Ernst 2004; Pesarin 2001) of maximum-type test statistics. The resulting residual-based shadings are illustrated both for mosaic and association plots using real world data sets.

The remainder of the paper is structured as follows: Section 2 gives a brief introduction to significance tests and visualization techniques for the independence problem in 2-way contingency tables. Based on this, Section 3 introduces the extended residual-based shadings using perceptually uniform HCL colors and combining visualization and significance testing. The results are generalized to multi-way tables in Section 4 with a more detailed discussion of adapting them to conditional independence problems. Section 5 summarizes the paper and gives some concluding remarks.

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 $i = 1, \ldots, I$ and $j = 1, \ldots, J$ and row and column sums $n_{i+} = \sum_i n_{ij}$ and $n_{+j} = \sum_j n_{ij}$, respectively. 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 estimated expected cell frequencies under H_0 are $\hat{n}_{ij} = n_{i+}n_{+j}/n_{++}$. As well-established in the stastical literature, a very closely related hypothesis is that of homogeneity which in particular leads to the same expected cell frequencies and is hence not discussed explicitely below. The probably best known and most used measure of discrepancy between observed and expected values are the Pearson residuals

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

The most convenient way to aggregate the $I \times J$ residuals to one test statistic is their sum of squares

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

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 first in statistics textbooks when addressing the independence problem in 2-way tables (see, e.g., Agresti 2002).

However, the sum of squares is not the only plausible way of capturing deviations from zero in the residuals. There are many other conceivable functionals $\lambda(\cdot)$ which lead to reasonable test statistics $\lambda([r_{ij}])$—and without further specification of a certain pattern of dependence no

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

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

functional $\lambda(\cdot)$ uniformly dominates all others in terms of power of the resulting test procedure. Therefore, the choice of the functional is usually also guided by the data analysis problem at hand: one functional which is particularly suitable for identifying the cells responsible for the 'dependence' (i.e., significant departure from independence), if any, is the maximum of the absolute values

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

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

Furthermore, an important reason for using the unconditional limiting distribution for the X^2 statistic from Equation 3 was the closed form result for the distribution. Recently, with the improving perfomance of computers, conditional inference (or permutation tests, conditioning on the observations)—carried out either by simulation or by computation of the (asymptotic) permutation distribution—have been receiving increasing attention (e.g., Ernst 2004; Pesarin 2001; Strasser and Weber 1999). For testing the independence hypothesis from Equation 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 the permutation distribution of test statistics of type $\lambda([r_{ij}])$.

Note, that virtually all ideas discussed in this paper also extend straightforwardly to the situation where other measures of discrepancy (such as, e.g., deviance residuals) are used instead of the Pearson residuals $[r_{ij}]$.

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 statistic from Equation 4 is $M = 1.87$ with $p = 0.0096$. Both p values have been computed from a sample of size $5,000$ from the permutation distribution under independence generated via sampling tables with the same row and column sums n_{i+} and n_{+j} using the Patefield (1981) algorithm and computing the respective statistic for each of these tables.

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 estimated 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 former 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 with areas 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 estimated 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 patients.

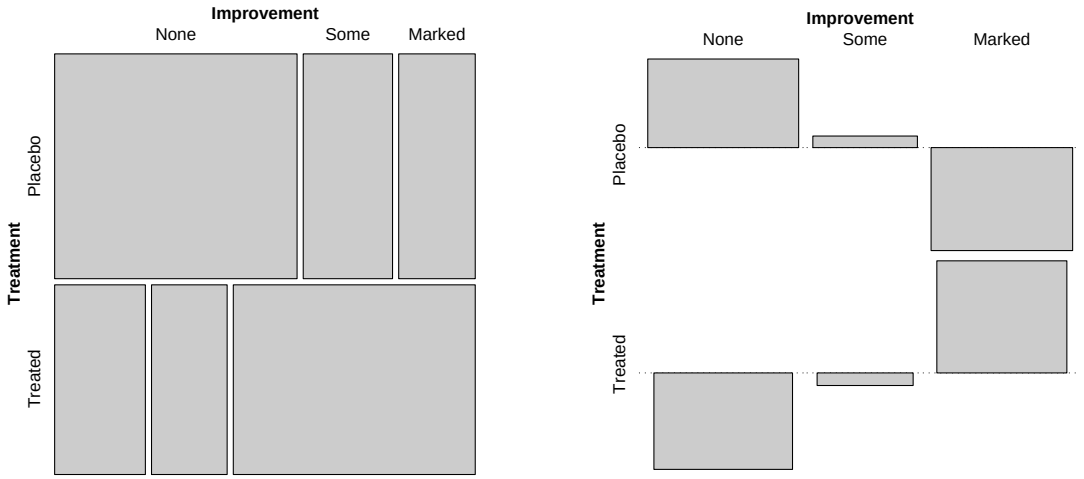

Figure 1: Classic mosaic and association plot for the arthritis data.

3. RESIDUAL-BASED SHADINGS

Colors are commonly used to enhance mosaic and association plots. To integrate a visualization of the 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, Zeileis, and Hornik 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 enhancing them from a plot for contingency tables to a visualization technique for log-linear models and their residuals—and thus also for independence problems including 2-way tables as the simplest 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 light blue and light red for positive and 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-linear 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 $\alpha = 0.0001$ levels. However, 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 color space. The implementation in the standard packages of the R system for statistical computing and graphics (R Development Core Team 2006) employs colors from HSV space. Both spaces are rather similar transformations of RGB (Red-Green-Blue) space (Brewer 1999; Poynton 2000) and are very common implementations of colors in many computer packages (Moretti and Lyons 2002) making the generation of the Friendly shading very simple. As both spaces can easily generate the same colors, we only discuss HSV space in the following.

The HSV space looks like a cone (see e.g., Wikipedia 2006, which also provides links to comparisons with other color spaces discussed in this paper) with black at its peak (zero value) and full color wheels for different saturations at the other end, around a white center (full value). Type and amount of color are controlled by hue and saturation, respectively. Typically, polar coordinates (h, s, v) rescaled for s and v to the interval $[0, 100]$ (or the unit interval) are used in this space, giving it the appearance of a 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, $h = 240$ (blue hue) for positive residuals. The absolute size of the residuals is then coded by the saturation s which is set to 0, 50 and 100, respectively, for small/medium/large residuals. The value is always fixed at $v = 100$. This is also depicted in the upper panel of Figure 2 which shows the saturation/value plane for the given hues $h = 0$ and $h = 240$. The full-color palette shows the colors used for the Friendly shading when the residuals are increasing from left to right. The reduced-color palette will be explained in Section 3.3.

Although this HSV-based shading is already very useful for enhancing mosaic and association plots and although HSV is a very commonly available implementation of color spaces, HSV color space 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 HSV dimensions are confounded, e.g., saturation is not uniform across different hues. A fully saturated blue $(240, 100, 100)$ is perceived to be much darker than a fully saturated red $(0, 100, 100)$ or green $(120, 100, 100)$. 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 graphics shaded with such colors harder to interpret. Finally, white is employed as the neutral color for small residuals in the Friendly shading, however typically grey is found to convey neutrality or un-interestingness much better than white (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' goes back until at least 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, 2004) 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 qualitative palettes 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

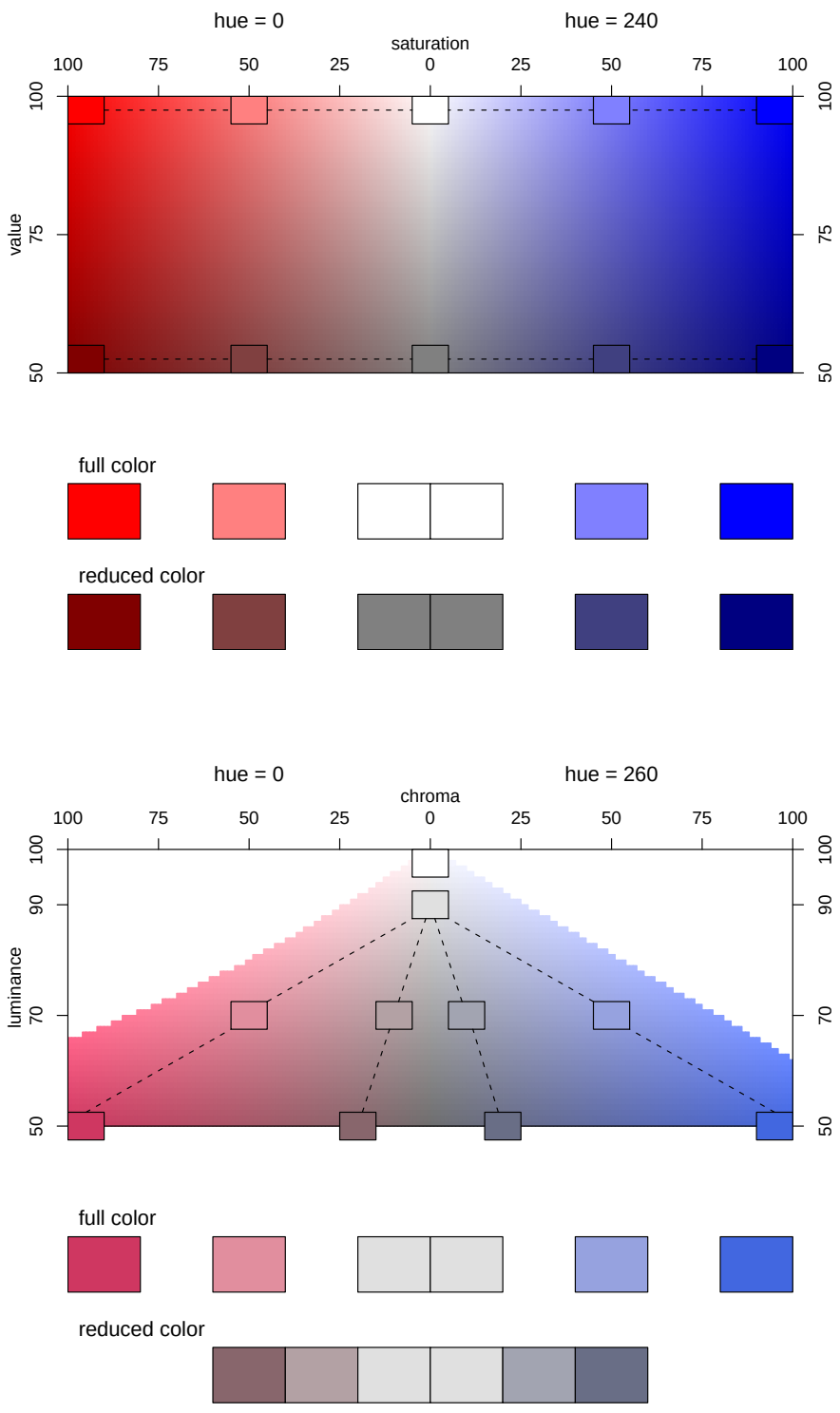

Figure 2: Residual-based shadings in HSV (upper) and HCL space (lower).

balanced towards the same grey and thus do not have the problem of varying saturations that the HSV colors have. In general, the HCL space offers much better support for selecting balanced palettes along simple paths through the space, which will be exploited below.

In the following, we discuss how ideas similar to those from Ihaka (2003) can be used for deriving 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 distorted 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 as regular as the HSV space: although its dimensions are usually also given by a hue ranging in $[0, 360]$ degrees and chroma and luminance ranging in $[0, 100]$ percent, not all combinations (h, c, l) yield valid HCL colors and the admissable combinations of c and l vary across different hues h. For the task 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 we can choose two hues that correspond to similar shapes in the chroma/luminance plane. The lower panel of Figure 2 shows two such planes side by side for the hues $h = 0$ and $h = 260$.[1] To obtain a sequence of colors with the similar 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. The diverging palette (see the full-color palette in Figure 2) 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 residuals r_{ij}—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 is a very effective way of visualization (i.e., yields palettes with more distinctive colors) and corresponds more closely with the properties of the Friendly shading.

Applying these palettes to the mosaic plot of the arthritis data yields the displays in the upper middle and right panels of Figure 3 (with data-driven cut offs as defined Section 3.3). This illustrates that especially the full color cells (in the 'marked' column) are less flashy and more balanced in the HCL shading as compared to the HSV shading.

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 un-interstingness, respectively). That is, viewers 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 are not true as can be seen in the upper left panel of Figure 3 which shows the mosaic display for the arthritis data with Friendly shading. Although there is significant dependence (as according to both the maximum and sum-of-squares tests), 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. However, in this situation, it would be nice if such cut offs could be chosen automatically in a data-driven way. Strategies for this are derived in the following.

The Friendly shading can be interpreted to be a visualization of the maximum statistic M from Equation 4 which always employs the critical values c_α 2 and 4. However, 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 upper middle and right panels 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

[1] The hue $h = 260$ is chosen rather than $h = 240$ because its chroma/luminance plane is most similar, as assessed by the area of the symmetric difference of the planes, to that of $h = 0$.

$\alpha = 0.01$ (using the diverging HSV and HCL palettes, respectively) derived from the permutation distribution of M for the arthritis data as described in Section 2.1. 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 violate the independence hypothesis 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 levels $\alpha = 0.1$ and $\alpha = 0.01$ are chosen because this leads to displays where fully colored cells are clearly significant ($p < 0.01$), cells without color are clearly non-significant ($p > 0.1$), and cells in between can be considered to be weakly significant ($0.01 \leq p \leq 0.1$). Of course, users could choose any other set of significance levels they feel comfortable with, e.g., only a single cut off 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, this typically results in too much color in the plot which in turn tends to conceal the important cells and over-emphasize the unimportant ones. Hence, a discrete shading with few colors is much easier to interpret.

This maximum shading is already very flexible and combines visualization and inference. However, it can only be applied when employing the maximum statistic because it is the only aggregation functional $\lambda(\cdot)$ where a single large residual $|r_{ij}|$ exceeding its critical value is equivalent to a significant value of the whole test statistic $\lambda([r_{ij}])$. Typically, applying the maximum statistic is feasible and also appropriate for exploratory analysis, but it would be desirable to also have a residual-based shading that can incorporate visualization of significance when the sum-of-squares statistic X^2 (or any other functional $\lambda(\cdot)$) is used. For the reasons discussed above, it is not possible to achieve this by shading individual cells differently but can only be realized by using different colors for the whole table. As outlined before, colorfulness is intuitively matched with interestingness, therefore a rather natural idea is to use the fully colored palette only when the corresponding test is significant and to use a less colorful palette if not. For the HCL scheme, the amount of color can conveniently be controlled by varying the maximum chroma value used. For the full colors, the maximum chroma was set to 100 as shown in Figure 2 and is decreased to 20 for the reduced-color palette. This palette still codes the absolute size of the residuals by luminance (i.e., the amount of grey), uses the same neutral grey for small 'un-interesting' residuals, codes positive and negative residuals by different hues, but gives less emphasis to the pattern by making the plot less colorful. A similar effect can be obtained in the HSV space if the value is reduced from 100 to 50 for reduced-color palettes (see Figure 2). As for the full colors, the dimensions used for creating this palette are confounded and hence the HCL scheme is clearly preferable.

To see such a sum-of-squares shading in practice, we employ a data set on the number of piston ring failures in three legs (north/center/south) in four different steam-driven compressors at an Imperial Chemical Industries plant. The contingency table is given in Haberman (1973), re-analyzed by Everitt and Hothorn (2006) and displayed in the lower row of Figure 3. Neither the sum-of-squares test ($X^2 = 11.722$, $p = 0.069$) nor the maximum test ($M = 1.78$, $p = 0.112$) find evidence for a departure from independence in the data (at the 5% level). However, as argued by Everitt and Hothorn (2006), a closer look at the size of the residuals might be interesting. To do so, we choose two fixed cut offs within the range of residuals, 1 and 1.5, showing that more failures (than expected under independence) in the center leg and less failures in the south leg were observed for compressor 1 and vice versa for compressor 4. This is brought out clearly by all three shadings in the lower row of Figure 3. However, less emphasis is given to this pattern by the sum-of-squares shadings in the middle and right panel because the reduced color palettes are used due to non-significance of the associated test (at 5% level). Comparing the HSV-based and HCL-based version shows that the latter are less flashy and more balanced.

Other strategies for constructing palettes for general aggregation functionals $\lambda(\cdot)$ (including the sum of squares) that are similar in spirit to the maximum shading are conceivable. Instead of using a reduced-color palette for non-significant tests, a no-color palette with only a light grey could be employed. This could also be seen as always using cut offs outside the range of residuals for non-significant tests. Analogously, for significant tests, cut offs that are always inside the range

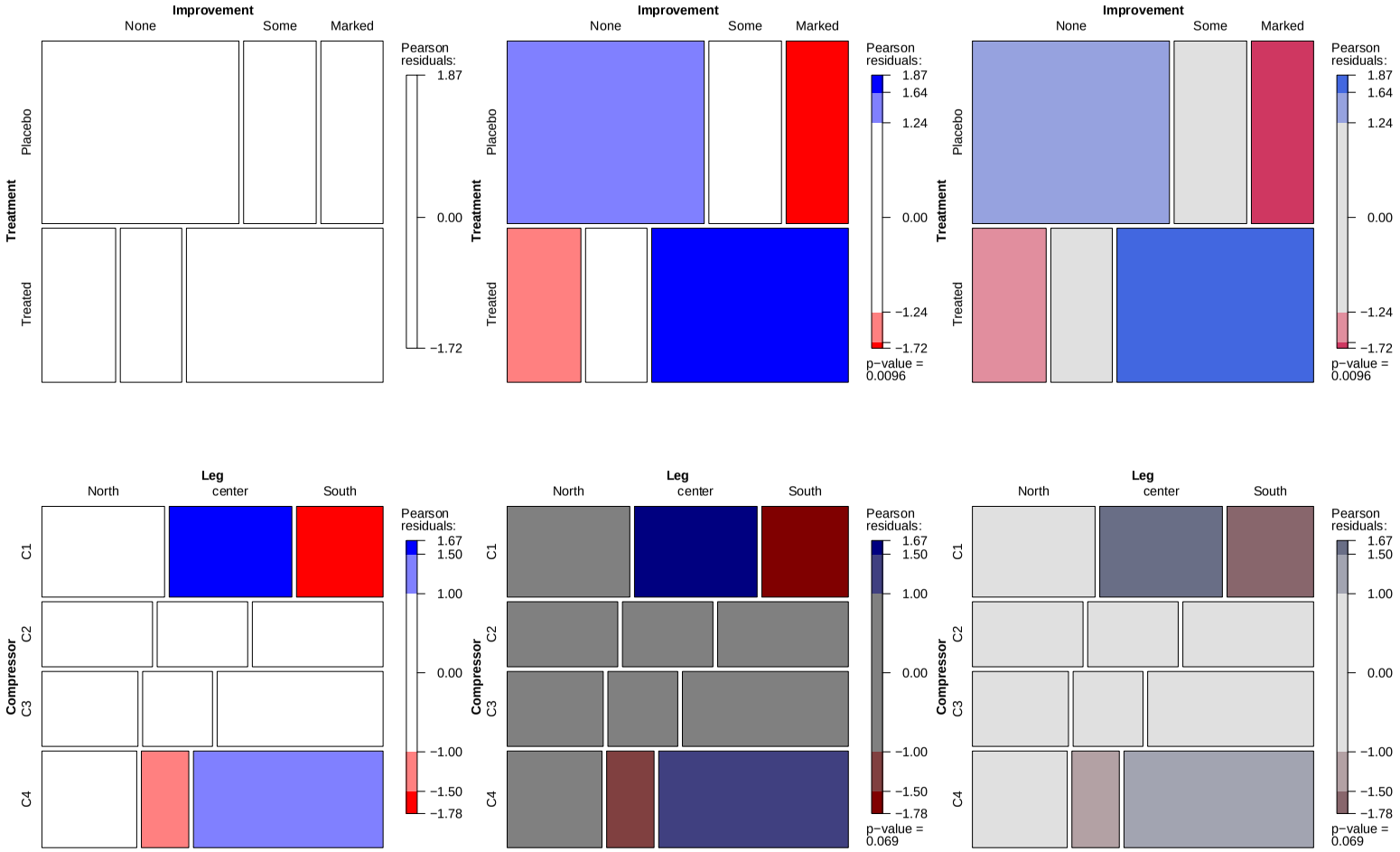

Figure 3: Upper row: Mosaic plot for the arthritis data with Friendly shading (left), HSV maximum shading (middle), HCL maximum shading (right). Lower row: Mosaic plot for the piston rings data with fixed user-defined cut offs 1 and 1.5 and Friendly shading (left), HSV sum-of-squares shading (middle), HCL sum-of-squares shading (right).

of residuals could be chosen. The latter could be determined by subject-matter knowledge, as a fraction of the associated critical value, or as certain quantiles of the absolute residuals.

4. EXTENSIONS

To introduce the new residual-based shadings without too much overhead in Section 3, we have only considered the independence problem in 2-way tables. In this section, generalizations of these ideas to independence problems for multi-way tables are outlined.

Mosaic displays have been emphasized in the literature to be an excellent means of visualization for log-linear models (Friendly 1999; Theus and Lauer 1999); typical hypotheses of interest include complete, joint or conditional independence. For all of these hypotheses, tables of estimated expected values and residuals (again Pearson or deviance) can be computed and Friendly (1994, 1999) shows that his residual-based shading scheme can directly be applied to these more complex independence models. For inference, the most commonly used aggregation functional for the residuals is again the sum of squares yielding the associated Pearson or likelihood ratio statistic, respectively (Agresti 2002).

As these independence models for multi-way tables also provide the structure required for the residual-based HCL shadings derived in Section 3, both the maximum shading and the sum-of-squares shading can straightforwardly be applied. However, such independence models often additionally provide further structure that allows decomposition of the overall model into smaller independence problems which can be exploited both for choosing appropriate data-driven cut-offs in the shading and for selecting a suitable layout of mosaic or association plots. Specific strategies for the conditional independence problem in 3-way tables are derived in the following.

Association plots are not commonly used for contingency tables with more than two margins, although there is nothing in the definition that would prevent application in higher dimensions. However, as argued for the mosaic plots by Friendly (1999) and Theus and Lauer (1999), it becomes increasingly important to choose a good layout as the number of variables grows.

How the structure of the independence problem can be exploited for selecting a suitable layout and shading for mosaic and association plots, is exemplified with the conditional independence problem in 3-way tables. For a table $[n_{ijk}]$ with underlying theoretical probability distribution $[\pi_{ijk}]$, this can be formulated as

$$H_0 : \pi_{ij|k} \quad = \quad \pi_{i+|k}\pi_{+j|k}. \tag{5}$$

where $\pi_{ij|k}$ are the conditional probabilities given the stratum k with $k = 1, \ldots, K$. Under the assumption of conditional independence, we can again estimate expected frequencies $[\hat{n}_{ijk}]$ and the corresponding residuals $[r_{ijk}]$. To test the conditional independence hypothesis, usually the sum-of-squares statistic is used

$$\sum_{i,j,k} r_{ijk}^2 \quad = \quad \sum_k X_k^2, \tag{6}$$

which is simply the sum of the individual sum-of-squares statistics X_k^2 in each stratum k.

Alternatively, a maximum statistic similar to that from Equation 4 can be constructed

$$\max_{i,j,k} |r_{ijk}| \quad = \quad \max_k M_k. \tag{7}$$

As in the 2-way case, this allows for identification of the cells which are responsible for the deviation from conditional independence (if any). If it is not so much of interest in which *cell* but only in which *stratum* the deviation occurs, then it would be natural to use

$$\max_k \sum_{i,j} r_{ijk}^2 \quad = \quad \max_k X_k^2. \tag{8}$$

Given a critical value for this statistic, all strata k whose associated sum-of-squares statistics X_k^2 exceed the critical value are in conflict with the hypothesis of conditional independence.

All these statistics are of type

$$\lambda_{\text{agg}}(\lambda_{\text{indep}}([r_{ijk}])), \tag{9}$$

where λ_{indep} is a functional for assessing independence in stratum k and λ_{agg} is a functional for aggregating over the $k = 1, \ldots, K$ strata. If the maximum is used for the latter, then identification of the strata responsible for the non-independence is possible. If additionally λ_{indep} is the maximum, the corresponding cells can also be identified. Hence, the double maximum statistic from Equation 7 is the only functional allowing for detection of both the strata and the cells violating the conditional independence hypothesis.

However, the main purpose of the formulation of the different test statistics is not so much inference but their applicability to diagnostic plots via residual-based shadings. As already discussed, it is possible for all aggregation functionals to simply use either the full-color or the reduced-color shading for all cells in the contingency table—this strategy would have to be used for the sum-of-squares statistic from Equation 6. If λ_{agg} is the maximum as in Equation 8, then the full-color palette would only be used in those strata in conflict with the hypothesis of conditional independence whereas the reduced-color palette would be used for the remaining strata. Finally, if both λ_{agg} and λ_{indep} are the maximum, then the same strategy as in Section 3 can be pursued, i.e., only the full-color palette is used but with data-driven cut offs derived from the distribution of the double maximum statistic from Equation 7.

To arrange the shaded rectangles of the association or the mosaic plot, respectively, the most intuitive approach is to use the same conditioning in the display that was also used for conditioning in the model. For this situation, Friendly (1999) discusses a grouping similar to coplots (conditioning plots, see Cleveland 1993) that lead to trellis graphics (Becker, Cleveland, and Shyu 1996). Thus, a natural visualization of such an independence model would be a trellis-like coplot where each stratum k could be visualized by an association or mosaic display. This has also the advantage that only the conditional independence problem but not the conditioning distribution over $k = 1, \ldots, K$ is visualized which could obscure departures from conditional independence if

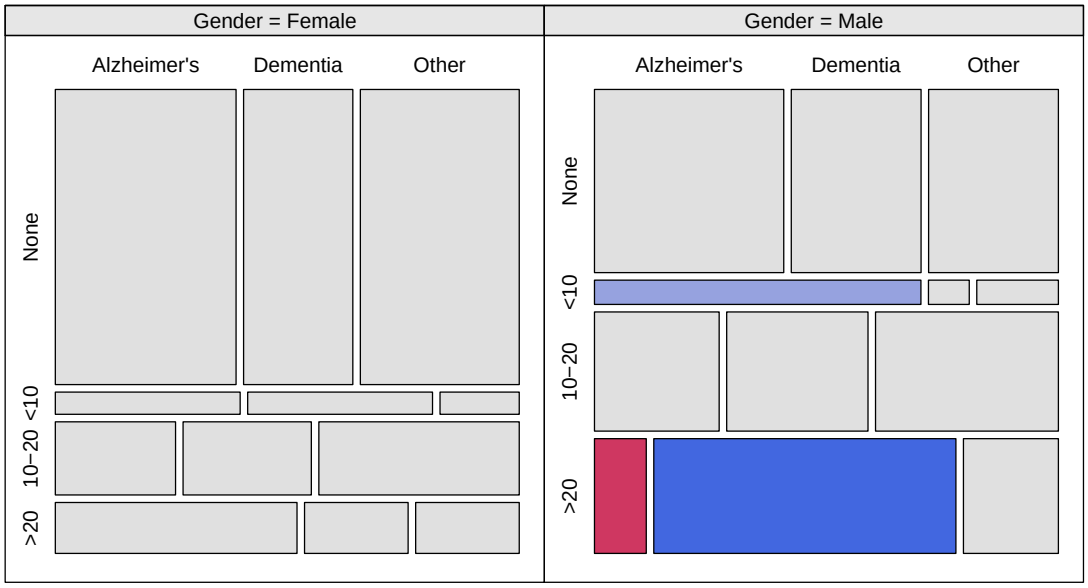

Figure 4: Conditional mosaic plot with double maximum shading for conditional independence of smoking and disease given gender.

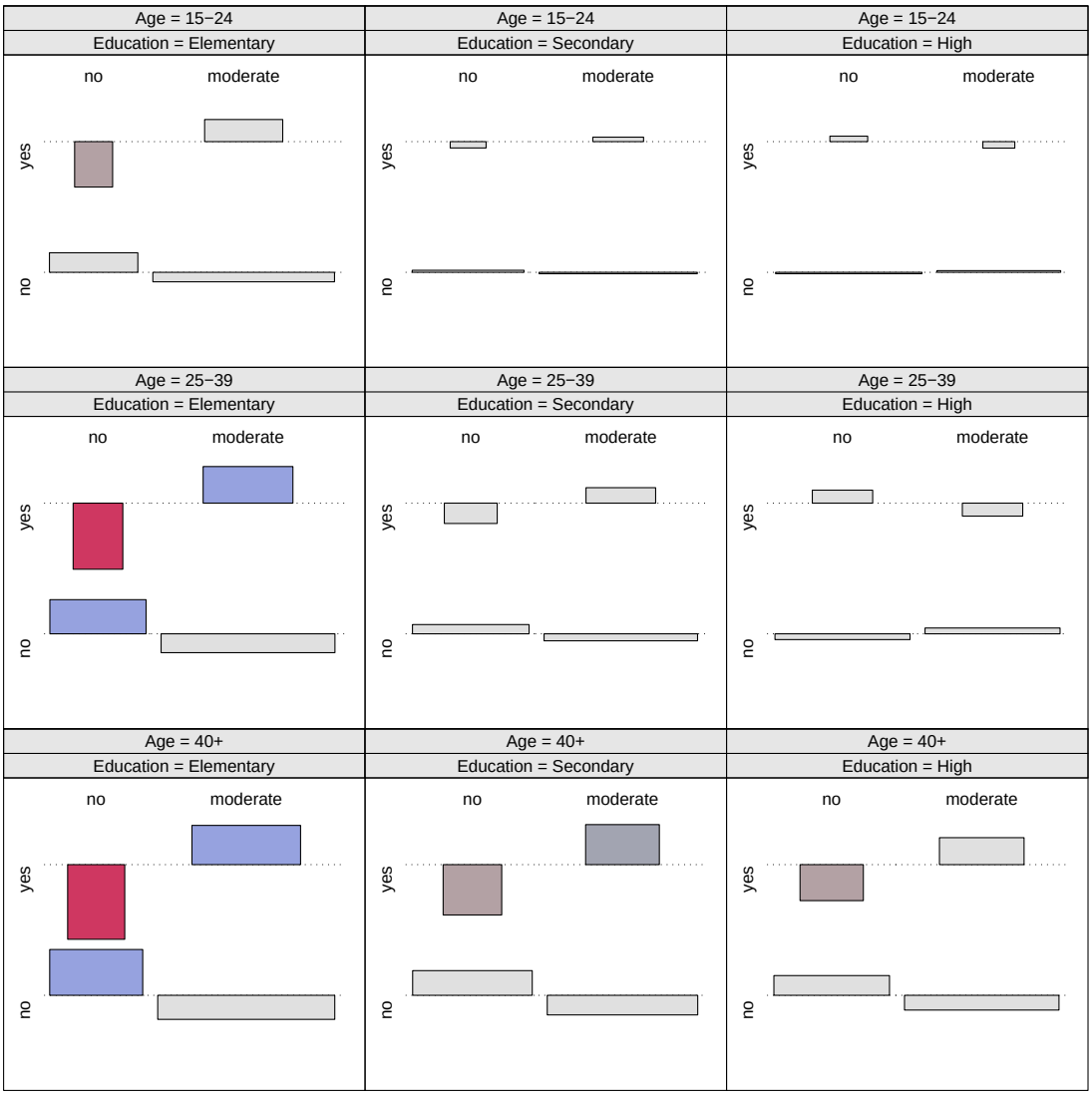

Figure 5: Conditional association plot with maximum sum-of-squares shading for conditional independence of memory and attitude given age and education.

the number of observations in each stratum n_{++k} are very different.

For illustration, a 3-way and a 4-way table are employed: The first is taken from a case-control study of smoking and Alzheimer's disease published in Salib and Hillier (1997) and re-analyzed using conditional inference techniques in Hothorn, Hornik, van de Wiel, and Zeileis (2006). It provides data on the smoking behaviour (no, <10, 10–20, >20 cigarettes per day), disease status (Alzheimer's, other dementias, other diagnoses) and gender. The question is whether smoking and disease status are conditionally independent given gender. All three statistics suggested above find evidence for departure from significance: the sum-of-squares statistic from Equation 6 is $\sum_k X_k^2 = 46.828$ ($p = 0$), the maximum sum-of-squares statistic from Equation 8 is $\max_k X_k^2 = 35.867$ ($p = 0$), and the double maximum statistic from Equation 7 is $\max_k M_k = 3.348$ ($p = 0$). The p values are again computed by drawing $5,000$ samples from the corresponding permutation distribution. The conditional mosaic plot in Figure 4 shows clearly that the association of smoking and disease is present only in the group of male patients. The double maximum shading employed

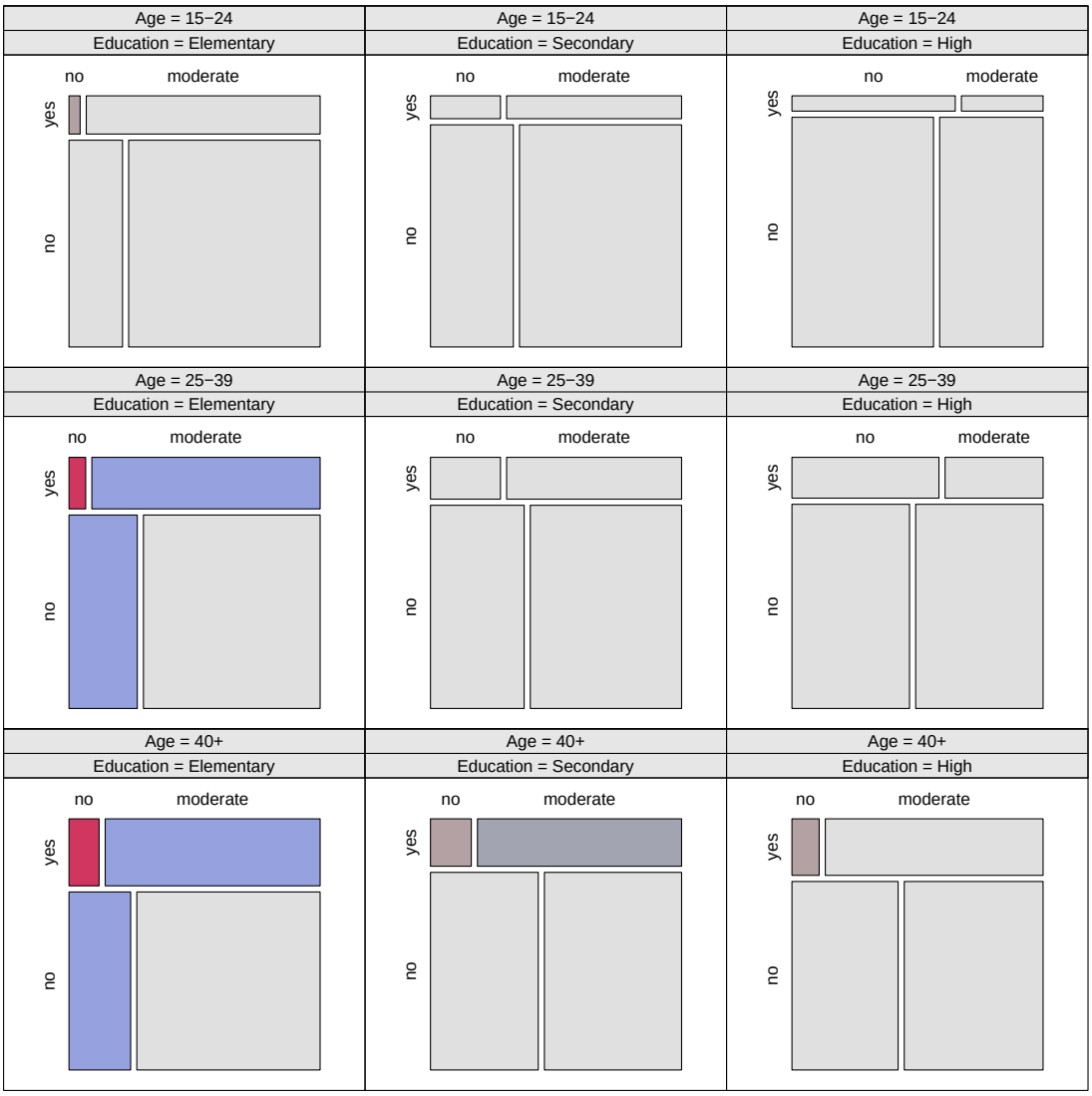

Figure 6: Conditional mosaic plot with maximum sum-of-squares shading for conditional independence of memory and attitude given age and education.

allows for identification of the male heavy smokers as the cells 'responsible' for the dependence: other dementias are more frequent and Alzheimer's disease less frequent in this group than expected under independence. Interestingly, there seems to be another large residual for the light smoker group (<10 cigarettes) and Alzheimer's disease—however, this is only significant at 10% and not at the 1% level as the other two cells.

As a 4-way example, we use the punishment data (Andersen 1991) from a study of the Gallup Institute in Denmark in 1979 about the attitude of a random sample of $1,456$ persons towards corporal punishment of children. The contingency table comprises four margins: memory of punishments as a child (yes/no), attitude as a binary variable (approval of "moderate" punishment or "no" approval), highest level of education (elementary/secondary/high), and age group (15–24, 25–39, $\geq$40 years). It is of interest whether there is an association between memories of corporal punishments as a child and attitude towards punishment of children as an adult, controlling for age and education. Figure 5 shows a conditional association plot of memory and attitude given age

and education. This can be interpreted as a diagnostic residual plot for the associated log-linear conditional independence model. Alternatively, a conditional mosaic plot as in Figure 6 can be used for visualizing the contingency table and its associated residuals from the conditional independence model using the same shading. Both reveal an association between memories and attitude for the lowest education group (first column) and highest age group (last row): experienced violence seems to engender violence again as there are less adults that disapprove punishment in the group with memories of punishments than expected under independence. For the remaining four age-education groups, there seems to be no association: all residuals of the conditional independence model are very close to zero in these cells. All three tests agree again that there is significant departure from conditional independence in this table: the sum-of-squares statistic is $\sum_k X_k^2 = 34.604$ ($p = 0.0002$), the maximum sum-of-squares statistic is $\max_k X_k^2 = 11.626$ ($p = 0.0064$), and the double maximum statistic is $\max_k M_k = 2.573$ ($p = 0.0056$). The $\max_k X_k^2$ result is visualized by means of a maximum sum-of-squares shading in Figure 5 with user-defined cut offs 1 and 2, chosen to be within the range of the residuals. The full-color palette is used only for those strata associated with a sum-of-squares statistic X_k^2 significant at (overall) 5% level, the reduced-color palette is used otherwise. This highlights that the dependence pattern is significant only for the middle and high age group in the low education column. The other panels in the first column and last row also show a similar dependence pattern, however, it is not significant at 5% level and hence graphically down-weighted by using reduced color.

5. CONCLUSIONS

Various strategies for constructing residual-based shadings for visualizing (conditional) independence in contingency tables via mosaic and association plots are discussed. The shading of Friendly (1994) is extended in two directions: the use of perceptually uniform HCL colors and the combination of visualization and significance testing. To achieve the former, a general guideline for constructing diverging palettes in HCL space is introduced. The advantages of using this HCL shading scheme instead of an HSV scheme are that the colors from this perceptually-based color space provide uniform saturations over different hues and that the colorfulness in this shading can be controlled independently from the other two dimensions. To combine visualization and significance testing, two approaches are presented: The first approach, based on the maximum statistic, always uses a fully colored palette but relies on data-driven cut offs such that the presence of color is equivalent to significance of the associated maximum test. The second approach, also applicable to other statistics such as the sum of squares, uses pre-defined cut offs (e.g., 2 and 4) but codes the result of the associated significance test by using full colors only if it is significant and the same type of palette with a reduced amount of color otherwise. Both strategies can not only be applied to the simple independence problem in 2-way contingency tables, but also to arbitrary inpendence models fitted via log-linear models in higher dimensions. In addition, it might be possible to exploit the structure of a given independence problem to achieve better visualizations which is illustrated for the conditional independence problem. All significance tests are carried out using a conditional inference approach instead of relying on unconditional asymptotic results.

ACKNOWLEDGEMENTS

We are thankful to Paul Murrell and Ross Ihaka—for providing tools, insights and feedback that helped to derive the results presented in this paper—and to two anonymous referees and editor Luke Tierney—for valuable comments and suggestions that lead to a substantial improvement of the paper.

COMPUTATIONAL DETAILS

The results in this paper were obtained using R 2.3.1 (R Development Core Team 2006, `http://www.R-project.org/`) and the packages **vcd** 1.0-1 (Meyer, Zeileis, and Hornik 2006), **MASS** 7.2-27 (see Venables and Ripley 2002), **grid** 2.3.1 (see Murrell 2002) and **colorspace** 0.95 (Ihaka 2004). A vignette that demonstrates how all empirical examples can exactly be reproduced in R is provided in the package **vcd**, see `vignette("residual-shadings", package = "vcd")`.

REFERENCES

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

Andersen, E. B. (1991), *The Statistical Analysis of Categorical Data*, Berlin: Springer-Verlag, 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.

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.

Commission Internationale de l'Éclairage (2004), *Colorimetry*, Vienna, Austria: Publication CIE 15:2004, 3rd ed., ISBN 3-901-90633-9.

Ernst, M. D. (2004), "Permutation Methods: A Basis for Exact Inference," *Statistical Science*, 19, 676–685.

Everitt, B. S. and Hothorn, T. (2006), *A Handbook of Statistical Analyses Using R*, Boca Raton, Florida: Chapman & Hall/CRC.

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.

Haberman, S. J. (1973), "The Analysis of Residuals in Cross-classified Tables," *Biometrics*, 29, 205–220.

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. F., New York: Springer-Verlag, 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.

Hothorn, T., Hornik, K., van de Wiel, M. A., and Zeileis, A. (2006), "A Lego System for Conditional Inference," *The American Statistician*, 60, 257–263.

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.

— (2004), **colorspace***: Colorspace Manipulation*, R package version 0.95.

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-Verlag.

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.

— (2006), "The Strucplot Framework: Visualizing Multi-way Contingency Tables with **vcd**," *Journal of Statistical Software*, 17, 1–48.

Moretti, G. and Lyons, P. (2002), "Tools for the Selection of Colour Palettes," in *Proceedings of the New Zealand Symposium On Computer-Human Interaction (SIGCHI 2002)*, University of Waikato, New Zealand.

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

Murrell, P. (2002), "The **grid** Graphics Package," *R News*, 2, 14–19.

Patefield, W. M. (1981), "An Efficient Method of Generating $R \times C$ Tables with Given Row and Column Totals," *Applied Statistics*, 30, 91–97, Algorithm AS 159.

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

Poynton, C. (2000), "Frequently-Asked Questions About Color," URL `http://www.poynton.com/ColorFAQ.html`, accessed 2006-09-14.

R Development Core Team (2006), *R: A Language and Environment for Statistical Computing*, R Foundation for Statistical Computing, Vienna, Austria, ISBN 3-900051-00-3.

Salib, E. and Hillier, V. (1997), "A Case-Control Study of Smoking and Alzheimer's Disease," *International Journal of Geriatric Psychiatry*, 12, 295–300.

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.

Venables, W. N. and Ripley, B. D. (2002), *Modern Applied Statistics with S*, New York: Springer-Verlag, 4th ed., ISBN 0-387-95457-0.

Wikipedia (2006), "HSV Color Space — Wikipedia, The Free Encyclopedia," URL `http://en.wikipedia.org/w/index.php?title=HSV_color_space&oldid=74735552`, accessed 2006-09-14.