DSC 2003 Working Papers
(Draft Versions)
http://www.ci.tuwien.ac.at/Conferences/DSC-2003/

Visualizing the Independence Problem Using Extended Association and Mosaic Plots

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

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

Abstract

Two visualization techniques for the independence problem in 2-way contingency tables—assocation and mosaic plots—are extended in two directions:

1. The visualization is enhanced by displaying the significance of an appropriate test for independence and by using better color schemes.
2. The implementation in the R system is improved using a more modular design and allowing for more flexible specification of plotting parameters.

Furthermore, some enhancements for displays of various types of independence in general multi-way tables are suggested.[1]

1 Introduction

Statistical models built for the analysis of multivariate data quickly become complex with increasing dimensionality. One idea of visualization techniques is to use the human visual system to detect structures in the data that possibly are not obvious from solely numeric output (e.g., test statistics). The R package `vcd`—inspired by the Book "Visualizing Categorical Data" (Friendly 2000)—includes methods for the (mostly graphical) exploration of categorical data, such as:

[1] Please note that all of this is work in progress and neither theoretically nor computationally fully sound.

- fitting and graphing of discrete distributions,
- plots and tests for independence and symmetry problems,
- visualization techniques for log-linear models.

Here, we focus on the visualization of the independence problem, in particular in 2-way tables.

2 Tests for independence in 2-way tables

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 way to measure the discrepancy between observed and expected values are the Pearson residuals

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

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

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

because it 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 thinkable functionals $\lambda(\cdot)$ which lead to reasonable test statistics $\lambda(r_{ij})$, the sum of squares is just one of them. Another functional which is very helpful for identifying the cells which cause the potential dependence 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 level α (Mazanec and Strasser 2000, ch. 7). Thus, the culprits causing the dependence can be easily identified. The critical values c_α can be derived from the conditional permutation distribution

of the table (n_{ij}). Of course, the χ^2 test 3 can also be carried out as a conditional test instead of using the unconditional asymptotic distribution.

Other classical tests for the hypothesis of independence include Fisher's exact test, asymptotic tests of the odds ratio and the Mantel-Haenzel test.

3 Visualization techniques for 2-way tables

The two best known visualization techniques for independence in 2-way tables are association plots and mosaic plots. Both are suitable to bring out departure 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}.

3.1 Association plots

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 orientation and color of the corresponding rectangle.

Figure 1 shows the association plot for the table of home and away goals for all 306 games in the 1995/6 season of the German soccer league Bundesliga (Knorr-Held 1999). In particular, it can be seen that there are fewer games ending 1:0 and more ending 1:1 than would be expected under independence. But it is absolutely unclear whether this difference is significant or not.

3.2 Mosaic plots

Mosaic plots 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 2).

Sequential horizontal and vertical recursive splits are used to visualize the frequencies of more than 2 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 arbitrary independence hypotheses. In this extension, the sign of the residuals is coded by solid and dashed line rectangles respectively. Furthermore, residuals that exceed 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 (in an HSV color scheme). The heuristic idea behind this shading is that cells with a large residual might cause the maximum test (4)

```
Loading required package: grid
```

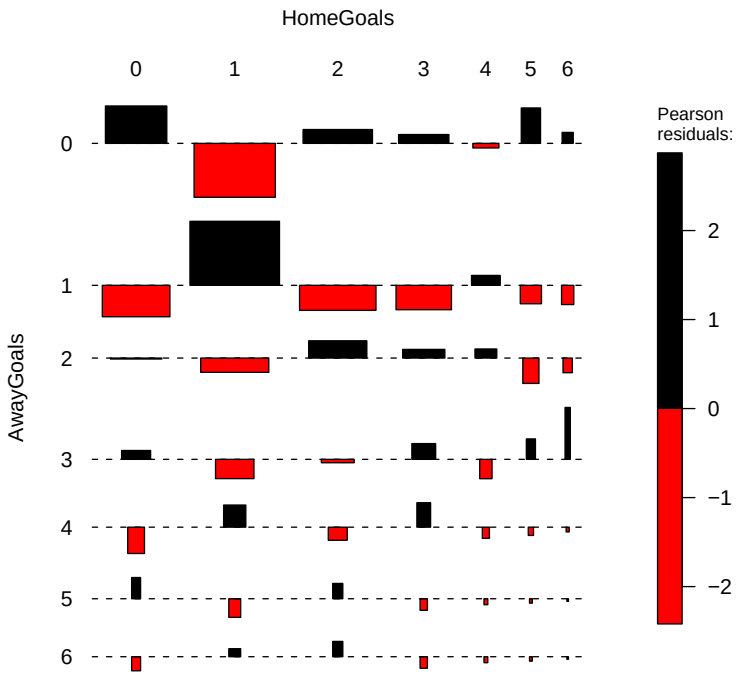

Figure 1: Association plot for the Bundesliga data.

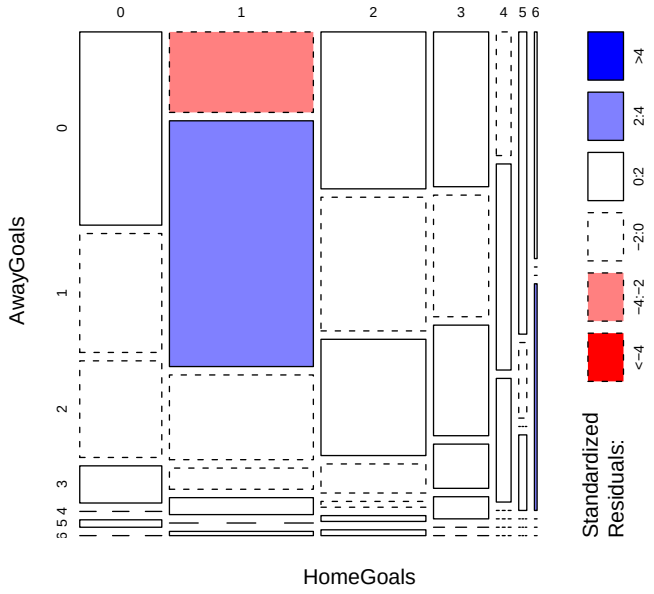

Figure 2: Friendly mosaic plot for the Bundesliga data.

to reject the hypothesis of independence. But in general it is unknown to which significance level α the values $c_\alpha = 2$ and 4 correspond.

Figure 2 shows the Friendly mosaic plot for the same Bundesliga data as in Figure 1. In this plot the residuals are visualized by the color shading of the cells. Again, the cells for the 1:0 and 1:1 results are highlighted, furthermore there seem to be too many 6:3 results as would be plausible under independence. Whereas the association plot in Figure 1 contains no information about the result of a test for independence, the shading in this plot conveys the impression that there might indeed be significant departure from independence (which would indicate that there is not only *no* home field advantage in the German Bundesliga but even an advantage for the away team). However, it is not clear if there is indeed evidence for this or not.

4 Extensions

In this section we suggest some improvements and extensions which can be applied to both visualization techniques introduced in Section 3. Firstly, the visual enhancements

- combining visualization of and testing for independence
- better color schemes

are addressed. Secondly, the implementation enhancements

- more modular implementation using `grid`
- more flexible plotting parameter specification

are described. None of these should be thought of as finished work but rather as a collection of ideas which still need thorough refinement and proper implementation.

4.1 Visualization enhancements

The extensions of Friendly (1994) to mosaic plots provide substantial improvement of the original mosaic plots and lifted them from a plot for contingency tables to a visualization technique for independence problems. However, it has two major drawbacks: Firstly, the significance level for the hard-coded critical values 2 and 4 is usually unknown. Secondly, the HSV colors for (light) blue and red were chosen due to the restrictions of the SAS software (in which the plots were first implemented), they 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 the critical values for the maximum statistic M from (4) can be derived from the conditional permutation distribution. Instead of the hard-coded values 2 and 4 the particular critical values, e.g., at the level $\alpha = 0.9$ and 0.95, for the table to be visualized could be used, thus highlighting exactly those

residuals who cause a (potential) dependency within the table. At the moment these critical values are derived in `vcd` by simulation of the underlying distribution.

As an alternative to the HSV colors the device-independent HCL (Hue-Chroma-Luminance) color space (Ihaka 2003) could be used. The resulting association plot (using a critical value for $\alpha = 0.95$) for the Bundesliga data can be seen in Figure 3. 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.

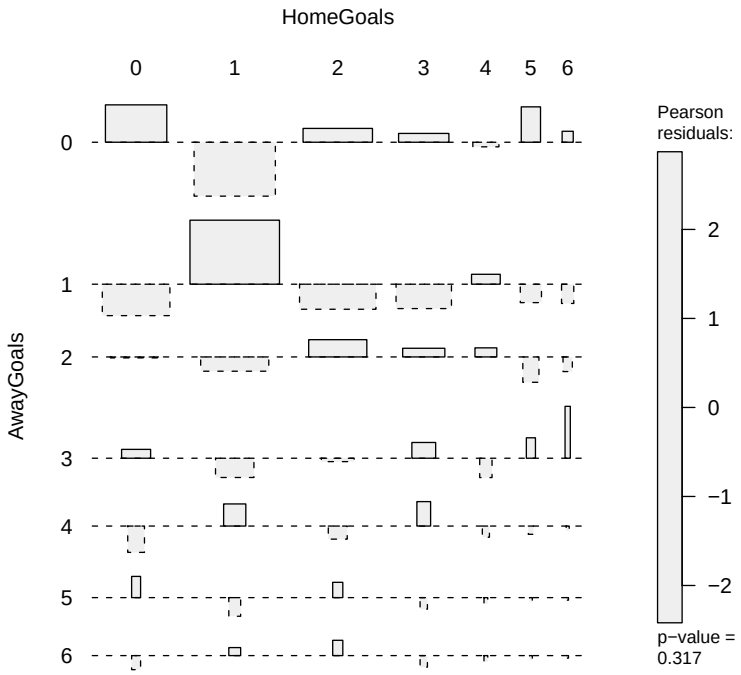

Figure 3: Extended association plots for the Bundesliga data.

To illustrate what a 'postive' example looks like we give the association plot for the famous Hair-Eye-Color data set cross-tabulating the hair and eye color of 592 statistics students. It is easy to see the hypothesis of independence is rejected at 95% level by the maximum test (4) due to higher frequencies of student with hair and eye color black/brown, blond/blue and red/green and lower frequencies of blond/brown and black/blue than would be plausible under independence.

We have seen that this approach performs very well when the test statistic which should be visualized is the maximum statistic (4) but it is difficult to extend this approach to general tests for independence, in particular in the HCL space, where

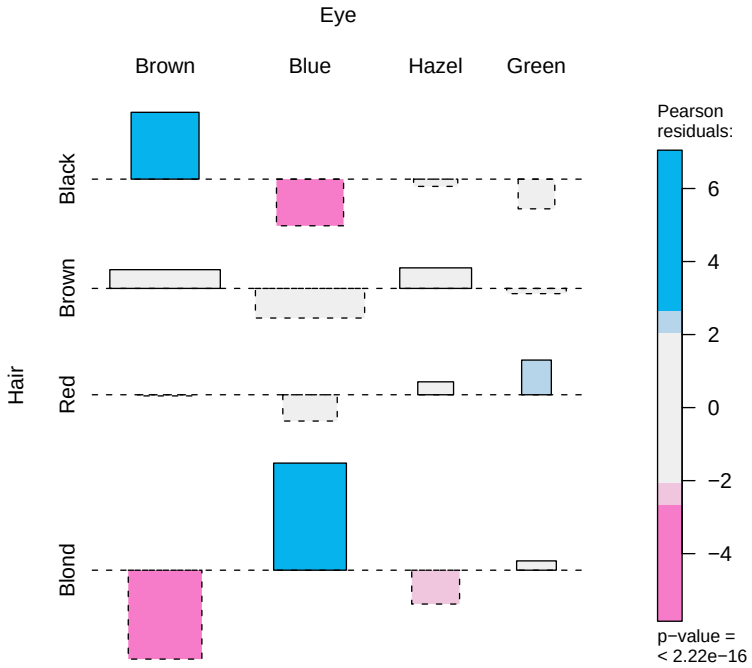

Figure 4: Extended association plots for the Hair-Eye-Color data.

the range of chroma and luminance are not independent. HSV space on the other hand 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 would 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 5. Again, it can be clearly seen that despite some large residuals there is no evidence against independence for the Bundesliga, but that there the null hypothesis has to be rejected for the Hair-Eye-Color data.

4.2 Implementation enhancements

The current implementation of association and mosaic plots in R suffers from two main disadvantages: Firstly, it is not easy to recycle the plots in conditioning plots or pairs plots (like mosaic matrices) as they have been implemented using R's base graphics engine. The new implementation was written from scratch in `grid` offering much more versatility amongst some other minor advantages and convenient improvements. Secondly, the graphics parameters of the rectangles in association and mosaic plots (like in almost all standard R plots), like color and line type, cannot be specified for each cell by the user. To overcome this, the user can specify in the current implementation in `vcd` either arrays of graphics parameters of the same dimensionality as the object that is plotted or a function which computes these graphics parameters based on the original table and its Pearson residuals. Functions are provided for the shading schemes described in the previous section.

5 Multi-way tables

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:

- total independence: $\pi_{ijk\ldots} = \pi_{i++\ldots}\pi_{+j+\ldots}\pi_{++k\ldots}\ldots$
- conditional independence: $\pi_{ijk\ldots} = \pi_{i|k\ldots}\pi_{j|k\ldots}$
- joint independence: $\pi_{ijk\ldots} = \pi_{ij+\ldots}\pi_{++k\ldots}$

Classical non-graphical methods for these problems include the Chi-square 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.

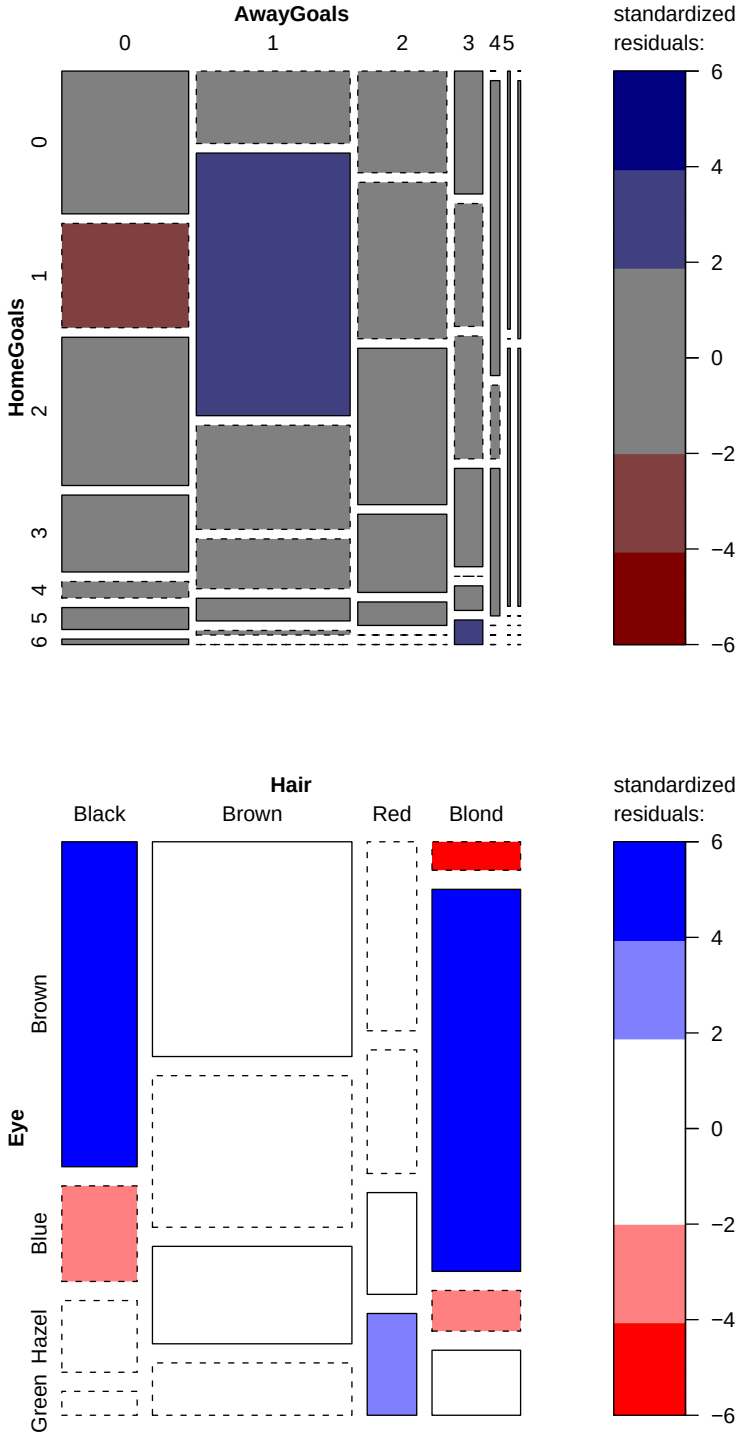

Figure 5: Extended mosaic plots for the Bundesliga and Hair-Eye-Color data.

Two natural ways to use the visualization techniques described in the previous sections would be to use (Trellis-like) conditioned 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). In this data the question whether there is sex discrimination at the UCB leads to the result that although women seem to be disadvantaged at the aggregated level there is no sex discrimination 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 conditioned assocation plot in Figure 6.

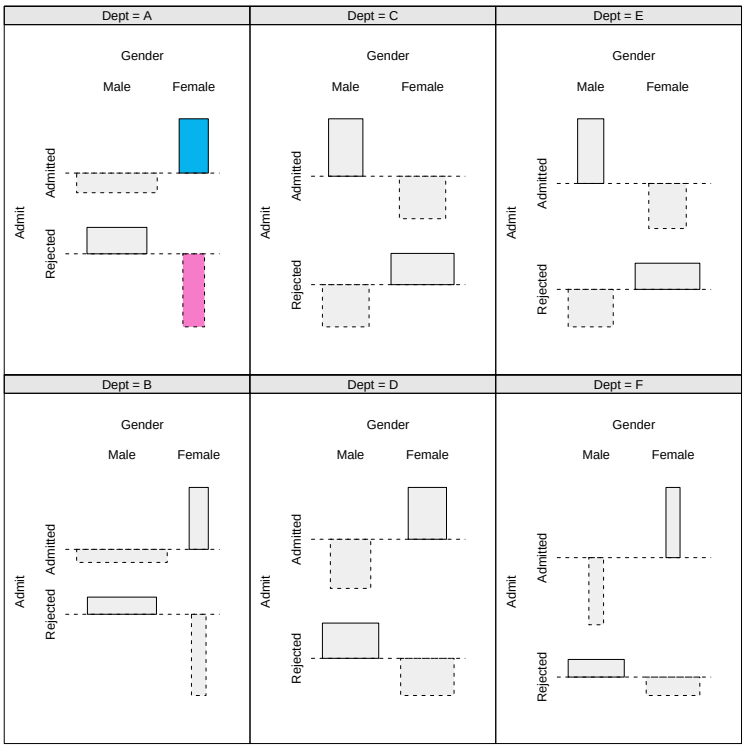

Figure 6: Extended conditioned association plots for the UCB Admissions data.

Similarly, the same data can be visualized using a mosaic matrix where a conditional independence model is fitted in each plot (see Figure 7).

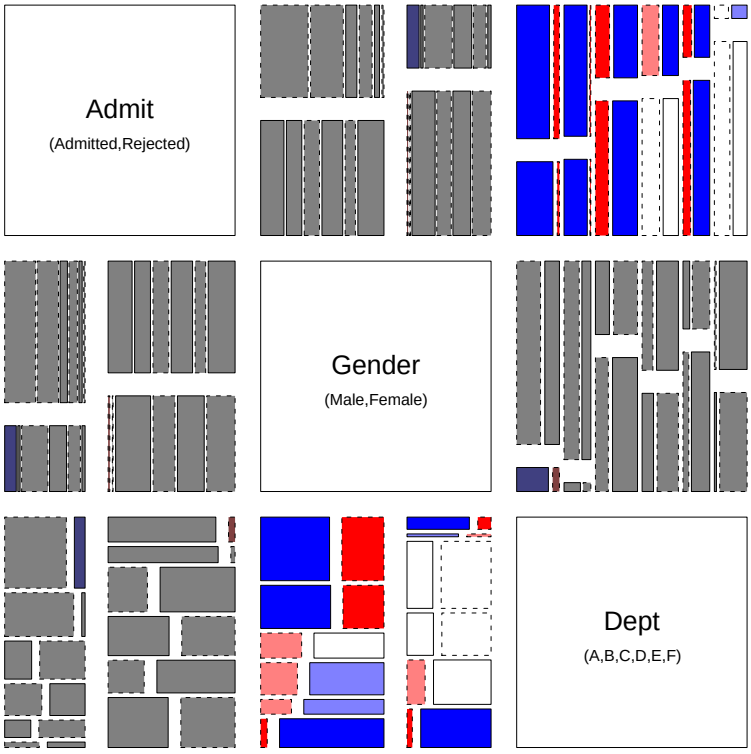

Figure 7: Extended conditioned association plots for the UCB Admissions data.

6 Conclusion

We suggest a set of enhancements for visualizing the independence problem in 2-way tables with some outlook to multi-way tables. The extensions aim at improving the visualization by displaying both the size of the residuals and the significance of a test for independence and by using better color schemes. Furthermore, a new implementation is outlined based on the graphics package `grid` which provides more modular design and more flexible specification of graphical parameters.

References

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

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

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

Ihaka, R. (2003), "Colour for Presentation Graphics," Unpublished Manuscript.

Knorr-Held, L. (1999), "Dynamic Rating of Sports Teams," Discussion Paper 98, SFB 386 "Statistical Analysis of Discrete Structures".

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