Residual-based Shadings in vcd

Achim Zeileis, David Meyer, and **Kurt Hornik**
Wirtschaftsuniversität Wien, Austria

Abstract

This vignette is a companion paper to Zeileis, Meyer, and Hornik (2005) which introduces several extensions to residual-based shadings for enhancing mosaic and association plots. The paper introduces (a) perceptually uniform Hue-Chroma-Luminance (HCL) palettes and (b) incorporates the result of an associated significance test into the shading. Here, we show how the examples can be easily reproduced using the **vcd** package.

Keywords: association plots, conditional inference, contingency tables, HCL colors, HSV colors, mosaic plots.

1. Introduction

In this vignette, we show how all empirical examples from Zeileis *et al.* (2005) can be reproduced in R(R Development Core Team 2006, http://www.R-project.org/), in particular using the package **vcd** (Meyer, Zeileis, and Hornik 2006a). Additionally, the pakcages **MASS** (see Venables and Ripley 2002), **grid** (see Murrell 2002) and **colorspace** (**?**) are employed. All are automatically loaded together with **vcd**:

```
> library("vcd")
> rseed <- 1071
```

Furthermore, we define a `rseed` which will be used as the random seed for making the results of the permutation tests (conditional inference) below exactly reproducible. In the following, we focus on the R code and output—for background information on the methods and the data sets, please consult Zeileis *et al.* (2005).

2. Arthritis data

First, we take a look at the association of treatment type and improvement in the `Arthritis` data. The data set can be loaded and brought into tabular form via:

```
> data("Arthritis", package = "vcd")
> (art <- xtabs(~Treatment + Improved, data = Arthritis, subset = Sex ==
+     "Female"))
```

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

Two basic explorative views of such a 2-way table are mosaic plots and association plots. They can be generated via `mosaic()` and `assoc()` from **vcd**, respectively. For technical documentation of these functions, please see Meyer, Zeileis, and Hornik (2006b). When no further arguments are supplied as in

```
> mosaic(art)
> assoc(art)
```

this yields the plain plots without any color shading, see Figure 1. Both indicate that there are more patients in the treatment group with marked improvement and less without improvement than would be expected under independence—and vice versa in the placebo group.

For 2-way tables, Zeileis *et al.* (2005) suggest to extend the shading of Friendly (1994) to also visualize the outcome of an independence test—either using the sum of squares of the Pearson residuals as the test statistic or their absolute maximum. Both statistics and their corresponding (approximate) permutation distribution can easily be computed using the function `coindep_test()`. Its arguments are a contingency table, a specification of margins used for conditioning (only for conditional independence models), a functional for aggregating the Pearson residuals (or alternatively the raw counts) and the number of permutations that should be drawn. The conditional table needs to be a 2-way table and the default is to compute the maximum statistic (absolute maximum of Pearson residuals). For the Arthritis data, both, the maximum test

```
> set.seed(rseed)
> (art_max <- coindep_test(art, n = 5000))

        Permutation test for conditional independence

data:  art
f(x) = 1.8696, p-value = 0.0096
```

and the sum-of-squares test, indicate a significant departure from independence.

```
> ss <- function(x) sum(x^2)
> set.seed(rseed)
> coindep_test(art, n = 5000, indepfun = ss)
```

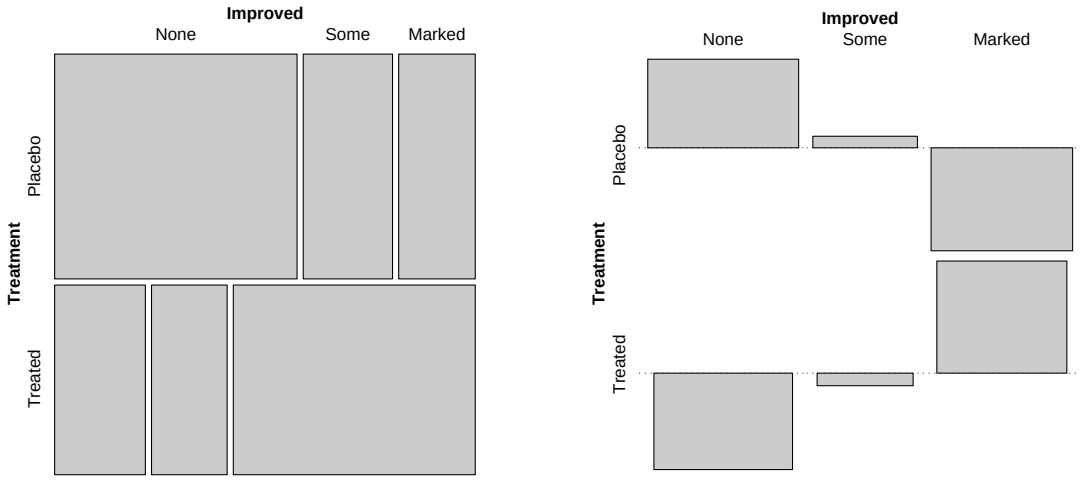

Figure 1: Classic mosaic and association plot for the arthritis data.

```
        Permutation test for conditional independence

data:  art
f(x) = 11.2962, p-value = 0.0032
```

Thus, it can be concluded that the treatment is effective and leads to significantly more improvement than the placebo. The classic views from Figure 1 and the inference above can also be combined, e.g., using the maximum shading that highlights the cells in an association or mosaic plot when the associated residuals exceed critical values of the maximum test (by default at levels 90% and 99%). To compare this shading (using either HSV or HCL colors) with the Friendly shading (using HSV colors), we generate all three versions of the mosaic plot:

```
> mosaic(art, gp = shading_Friendly(lty = 1, eps = NULL))
> mosaic(art, gp = shading_hsv, gp_args = list(interpolate = art_max$qdist(c(0.9,
+     0.99)), p.value = art_max$p.value))
> set.seed(rseed)
> mosaic(art, gp = shading_max, gp_args = list(n = 5000))
```

the results are shown in the upper row of Figure 2. The last plot could hae also been generated analogously to the second plot using `shading_hcl()` instead of `shading_hsv()`—`shading_max()` is simply a wrapper function which performs the inference and then visualizes it based on HCL colors.

3. Piston rings data

Instead of bringing out the result of the maximum test in the shading, we could also use a sum-of-squares shading that visualizes the result of the sum-of-squares test. As an illustration, we use the `pistonrings` data from the `HSAUR` (Everitt and Hothorn 2006) package giving the number of piston ring failurs in different legs of different compressors at an industry plant:

```
> data("pistonrings", package = "HSAUR")
> pistonrings

          leg
compressor North Centre South
        C1    17     17    12
        C2    11      9    13
        C3    11      8    19
        C4    14      7    28
```

Although there seems to be some slight association between the leg (especially center and South) and the compressor (especially numbers 1 and 4), there is no significant deviation from independence:

```
> set.seed(rseed)
> coindep_test(pistonrings, n = 5000)

        Permutation test for conditional independence

data:  pistonrings
f(x) = 1.7802, p-value = 0.112

> set.seed(rseed)
> (pring_ss <- coindep_test(pistonrings, n = 5000, indepfun = ss))
```

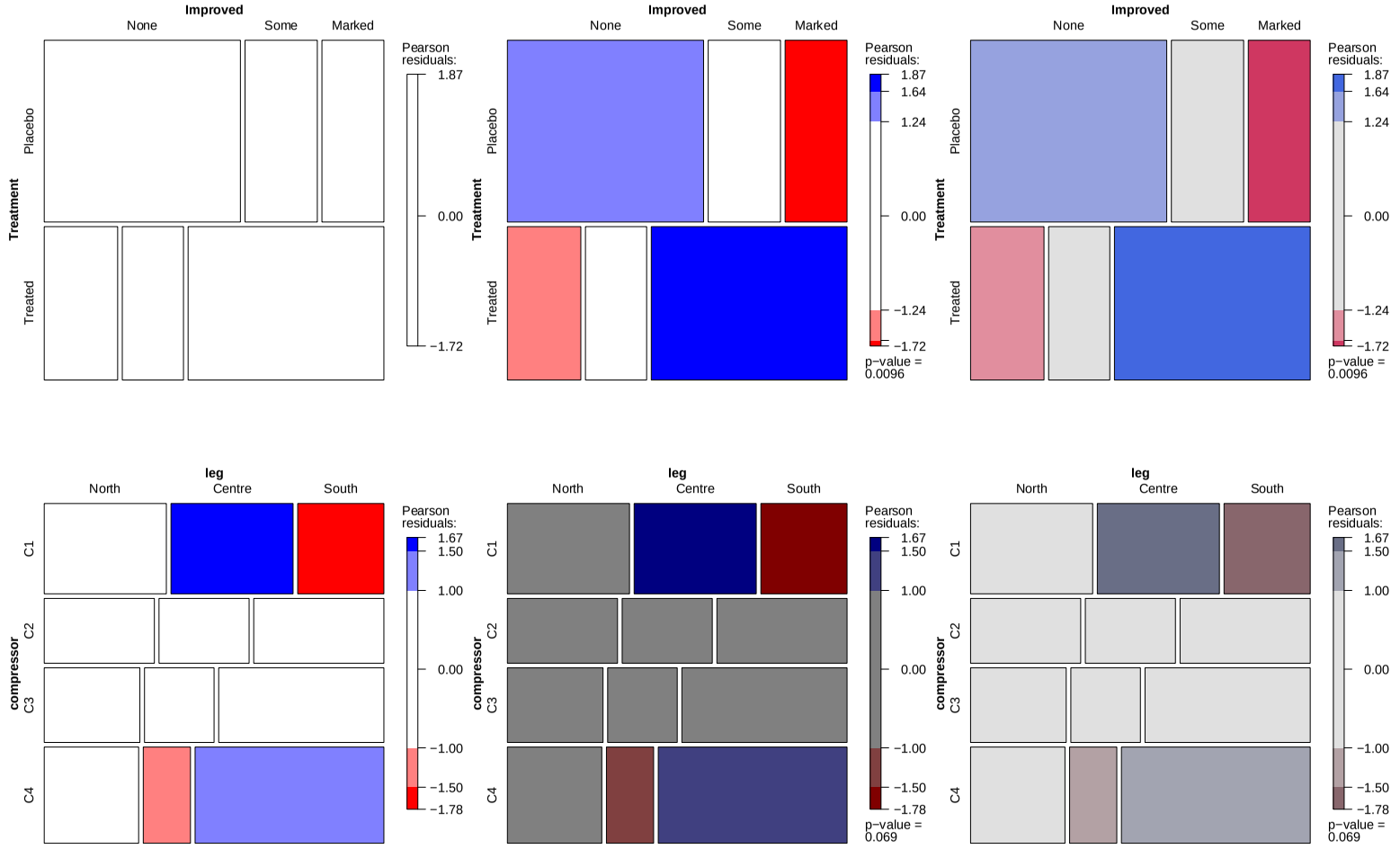

Figure 2: Upper row: Mosaic plot for the arthritis data with Friendly shading (left), HSV maximum shading (middle), HCL maximum shading (right). Lower row: Mosaic plot for the piston rings data with fixed user-defined cut offs 1 and 1.5 and Friendly shading (left), HSV sum-of-squares shading (middle), HCL sum-of-squares shading (right).

```
        Permutation test for conditional independence

data:  pistonrings
f(x) = 11.7223, p-value = 0.069
```

This can also be brought out graphically in a shaded mosaicplot by enhancing the Friendly shading (based on the user-defined cut-offs 1 and 1.5, here) to use a less colorful palette, either based on HSV or HCL colors:

```
> mosaic(pistonrings, gp = shading_Friendly(lty = 1, eps = NULL,
+     interpolate = c(1, 1.5)))
> mosaic(pistonrings, gp = shading_hsv, gp_args = list(p.value = pring_ss$p.value,
+     interpolate = c(1, 1.5)))
> mosaic(pistonrings, gp = shading_hcl, gp_args = list(p.value = pring_ss$p.value,
+     interpolate = c(1, 1.5)))
```

The resulting plots can be found in the lower row of Figure 2. The default in `shading_hcl()` and `shading_hsv()` is to use the asymptotical p value, hence we set it explicitely to the permtuation-based p value computed above.

4. Alzheimer and smoking

For illustrating that the same ideas can be employed for visualizing (conditional) independence in multi-way tables, Zeileis *et al.* (2005) use a 3-way and a 4-way table. The former is taken from a case-control study of smoking and Alzheimer's disease (stratified by gender). The data set is available in R in the package **coin** Hothorn, Hornik, van de Wiel, and Zeileis (2006).

```
> data("alzheimer", package = "coin")
> alz <- xtabs(~smoking + disease + gender, data = alzheimer)
> alz

, , gender = Female

       disease
smoking Alzheimer's Other dementias Other diagnoses
  None           91              55              80
  <10             7               7               3
  10-20          15              16              25
  >20            21               9               9

, , gender = Male

       disease
smoking Alzheimer's Other dementias Other diagnoses
  None           35              24              24
  <10             8               1               2
  10-20          15              17              22
  >20             6              35              11
```

To assess whether smoking behaviour and disease status are conditionally independent given gender, Zeileis *et al.* (2005) use three different types of test statistics: double maximum (maximum of maximum statistics in the two strata), maximum sum of squares (maximum of sum-of-squares statistics), and sum of squares (sum of sum-of-squares statistics). All three can be computed and assessed via permutation methods using the function `coindep_test()`:

```
> set.seed(rseed)
> coindep_test(alz, 3, n = 5000)

        Permutation test for conditional independence

data:  alz
f(x) = 3.348, p-value < 2.2e-16

> set.seed(rseed)
> coindep_test(alz, 3, n = 5000, indepfun = ss)

        Permutation test for conditional independence

data:  alz
f(x) = 35.8674, p-value < 2.2e-16

> set.seed(rseed)
> coindep_test(alz, 3, n = 5000, indepfun = ss, aggfun = sum)

        Permutation test for conditional independence

data:  alz
f(x) = 46.8285, p-value < 2.2e-16
```

The conditional mosaic plot in Figure 3 shows clearly that the association of smoking and disease is present only in the group of male patients. The double maximum shading employed 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

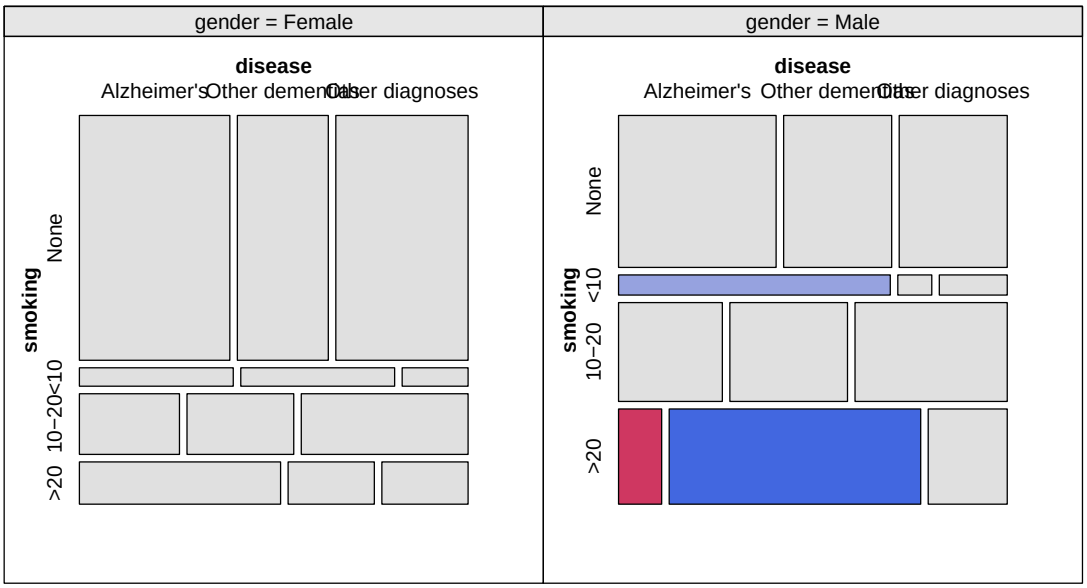

Figure 3: Conditional mosaic plot with double maximum shading for conditional independence of smoking and disease given gender.

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.

```
> set.seed(rseed)
> cotabplot(~smoking + disease | gender, data = alz, panel = cotab_coindep,
+     n = 5000)
```

5. Corporal punishment of children

As a 4-way example, data 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 is used. 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, ≥40 years).

```
> data("Punishment", package = "vcd")
> pun <- xtabs(Freq ~ memory + attitude + age + education, data = Punishment)
> ftable(pun, row.vars = c("age", "education", "memory"))

                        attitude  no moderate
age   education  memory
15-24 elementary yes               1       21
                 no               26       93
      secondary  yes               2        5
                 no               23       45
      high       yes               2        1
                 no               26       19
25-39 elementary yes               3       41
                 no               46      119
      secondary  yes               8       20
                 no               52       84
      high       yes               6        4
                 no               24       26
40-   elementary yes              20      143
                 no              109      324
      secondary  yes               4       20
                 no               44       56
      high       yes               1        8
                 no               13       17
```

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. All three test statistics already used above confirm that memories and attitude are conditionally associated:

```
> set.seed(rseed)
> coindep_test(pun, 3:4, n = 5000)

        Permutation test for conditional independence

data:  pun
f(x) = 2.5725, p-value = 0.0056
```

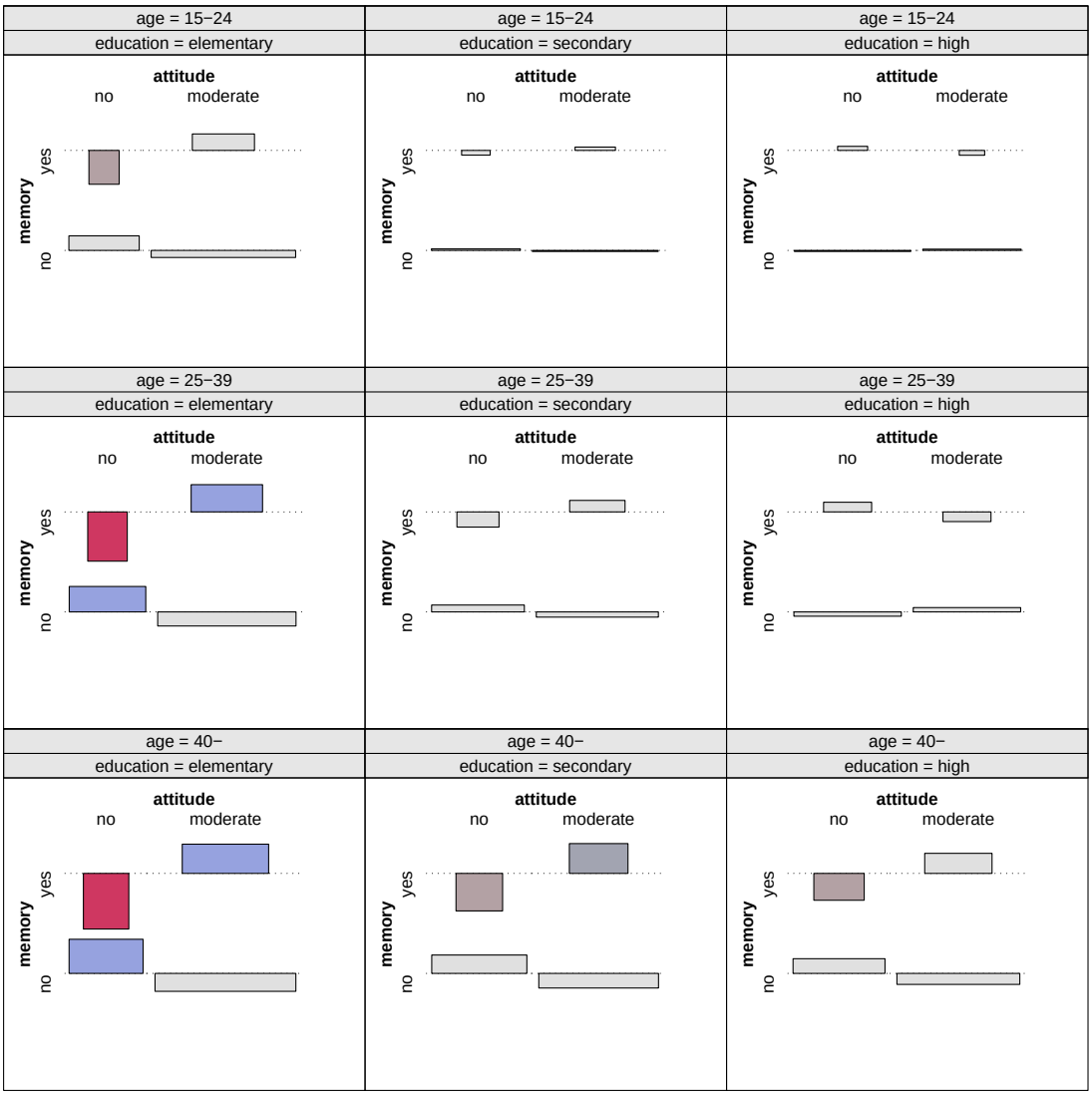

Figure 4: Conditional association plot with maximum sum-of-squares shading for conditional independence of memory and attitude given age and education.

```
> set.seed(rseed)
> coindep_test(pun, 3:4, n = 5000, indepfun = ss)

        Permutation test for conditional independence

data:  pun
f(x) = 11.6256, p-value = 0.0064

> set.seed(rseed)
> coindep_test(pun, 3:4, n = 5000, indepfun = ss, aggfun = sum)

        Permutation test for conditional independence
```

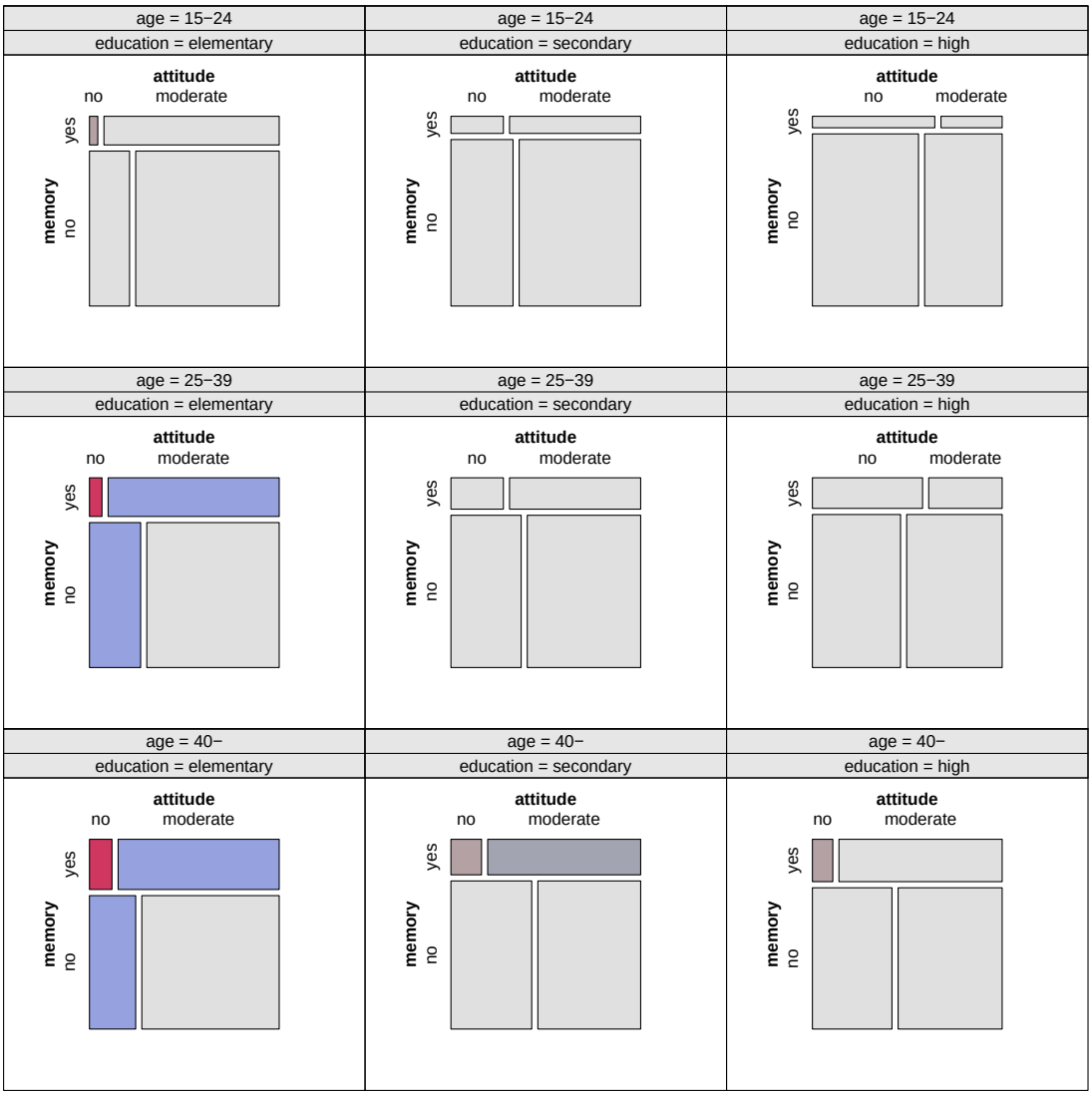

Figure 5: Conditional mosaic plot with maximum sum-of-squares shading for conditional independence of memory and attitude given age and education.

```
data:  pun
f(x) = 34.6041, p-value = 2e-04
```

Graphically, this dependence can be brought out using conditional association or mosaic plots as shown in Figure 4 and 5, respectively. 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. The figures employ the maximum sum-of-squares shading 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 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.

```
> set.seed(rseed)
> cotabplot(~memory + attitude | age + education, data = pun, panel = cotab_coindep,
+     n = 5000, type = "assoc", test = "maxchisq", interpolate = 1:2)

> set.seed(rseed)
> cotabplot(~memory + attitude | age + education, data = pun, panel = cotab_coindep,
+     n = 5000, type = "mosaic", test = "maxchisq", interpolate = 1:2)
```

References

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

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

Hothorn T, Hornik K, van de Wiel MA, Zeileis A (2006). "A Lego System for Conditional Inference." *The American Statistician*, **60**(3), 257–263. doi:10.1198/000313006X118430.

Meyer D, Zeileis A, Hornik K (2006a). **vcd***: Visualizing Categorical Data*. R package version 1.0-6.

Meyer D, Zeileis A, Hornik K (2006b). "The Strucplot Framework: Visualizing Multi-way Contingency Tables with **vcd**." *Journal of Statistical Software*, **17**(3), 1–48. URL http://www.jstatsoft.org/v17/i03/.

Murrell P (2002). "The **grid** Graphics Package." *R News*, **2**(2), 14–19. URL http://CRAN.R-project.org/doc/Rnews/.

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, URL http://www.R-project.org/.

Venables WN, Ripley BD (2002). *Modern Applied Statistics with S*. Springer-Verlag, New York, 4th edition. ISBN 0-387-95457-0, URL http://www.stats.ox.ac.uk/pub/MASS4/.

Zeileis A, Meyer D, Hornik K (2005). "Residual-based Shadings for Visualizing (Conditional) Independence." *Report 20*, Department of Statistics and Mathematics, Wirtschaftsuniversität Wien, Research Report Series. URL http://epub.wu-wien.ac.at/dyn/openURL?id=oai:epub.wu-wien.ac.at:epub-wu-01_871.