# Chapter 4
# Log-Linear Models

**Abstract** The classical log-linear models are introduced for two-way and multi-way contingency tables. Estimation theory, goodness-of-fit testing, and model selection procedures are discussed. Characteristic examples are worked out in `R` and interpreted. Log-linear models for three-dimensional tables are illustrated through mosaic plots. Graphical models are shortly discussed. Finally the collapsibility in multi-way tables, in connection to Simpson's paradox, is addressed.

**Keywords** Hierarchical log-linear models • Model fit and selection • Dissimilarity index • Graphical models • Simpson's paradox

## 4.1 Log-Linear Models for Two-way Tables

### *4.1.1 Model of Independence*

Independence (2.34) between the classification variables $X$ and $Y$ can equivalently be expressed in terms of the expected under independence cell frequencies $m_{ij}$ in a *log-linear model* form as

$$\log m_{ij} = \lambda + \lambda_i^X + \lambda_j^Y \ , \qquad i = 1, \ldots, I, \ j = 1, \ldots, J \ , \tag{4.1}$$

where $\lambda$ corresponds to the overall mean while $\lambda_i^X$, $\lambda_j^Y$ are the $i$th row and $j$th column main (or marginal) effects, respectively.

Model (4.1) could equivalently be expressed in terms of the expected under the assumed model probabilities $\pi_{ij}$. The usual choice is in terms of $m_{ij}$, because expected cell frequencies are common for the different sampling schemes while the underlying probability structure changes (see Sect. 2.2.1). For this, all log-linear models considered in the sequel will be expressed in terms of expected cell frequencies.

M. Kateri, *Contingency Table Analysis: Methods and Implementation Using R*, Statistics for Industry and Technology, DOI 10.1007/978-0-8176-4811-4_4,
© Springer Science+Business Media New York 2014

Interpretation is carried out in terms of the odds. For given column category $j$, under model (4.1), the odds of being in row $i_1$ instead of row $i_2$ ($i_1 \neq i_2$), $i_1, i_2 = 1, \ldots, I$, is

$$\frac{m_{i_1 j}}{m_{i_2 j}} = \frac{\exp(\lambda + \lambda_{i_1}^X + \lambda_j^Y)}{\exp(\lambda + \lambda_{i_2}^X + \lambda_j^Y)} = \exp(\lambda_{i_1}^X - \lambda_{i_2}^X) \ , \quad j = 1, \ldots, J \ , \tag{4.2}$$

independent of $j$. Similarly, for columns $j_1$ and $j_2$ ($j_1 \neq j_2, \ j_1, j_2 = 1, \ldots, J$),

$$\frac{m_{ij_1}}{m_{ij_2}} = \exp(\lambda_{j_1}^Y - \lambda_{j_2}^Y) \ , \quad i = 1, \ldots, I \ , \tag{4.3}$$

i.e., the odds of being in column $j_1$ instead of $j_2$ is determined only by the distance of the corresponding column main effect values and is independent of $i$. By (4.3), the conditional $j_1$ and $j_2$ column probabilities (within row $i$)

$$\frac{\mathsf{P}(Y = j_1 | X = i)}{\mathsf{P}(Y = j_2 | X = i)} = \exp(\lambda_{j_1}^Y - \lambda_{j_2}^Y) \ , \quad i = 1, \ldots, I \ ,$$

relate the same for all rows and this is true for any pair of columns $j_1$ and $j_2$. Thus, the conditional column distribution is the same for all rows, as should be for independent $X$ and $Y$.

Using (4.3), the expected under independence local odds ratios are

$$\theta_{ij}^L = \frac{m_{ij}/m_{i.j+1}}{m_{i+1.j}/m_{i+1.j+1}} = \frac{e^{\lambda_j^Y - \lambda_{j+1}^Y}}{e^{\lambda_j^Y - \lambda_{j+1}^Y}} = 1 \ , \quad i = 1, \ldots, I-1, \ j = 1, \ldots, J-1,$$

i.e., all equal to 1, as expected by (2.52).

The parameters in model (4.1) are $1 + I + J$ while we know that under independence the parameters are $(I-1) + (J-1)$. Hence, parameters in (4.1) are not uniquely determined unless constraints are imposed on the main effects. The traditionally used identifiability constraints are the sum to zero constraints:

$$\sum_{i=1}^{I} \lambda_i^X = \sum_{j=1}^{J} \lambda_j^Y = 0 \ . \tag{4.4}$$

Due to computational convenience, software applications replace (4.4) by the constraints that set a category effect to zero, usually the last ($\lambda_I^X = \lambda_J^Y = 0$) or the first ($\lambda_1^X = \lambda_1^Y = 0$).

The different set of constraints are equivalent and they affect only the reference point for physical interpretation. Thus, $\lambda_i^X$ compares the $i$th row category to the overall mean or to the first category, depending on whether model (4.1) is fitted under (4.4) or under $\lambda_1^X = 0$. The differences $\lambda_{i_1}^X - \lambda_{i_2}^X$ and $\lambda_{j_1}^Y - \lambda_{j_2}^Y$ are constraints invariant; thus, comparisons between categories are not affected by the identifiability constraints used.

Model (4.1) will be illustrated in Sect. 4.2.1, after we discuss technical matters on parameter estimation and model fit checking.

### 4.1.2 The Saturated Model

In case the classification variables $X$ and $Y$ are *not* independent, their interaction is significant and the corresponding *$XY$-interaction term* has to be added in the log-linear model expression, leading to the *saturated* model

$$\log m_{ij} = \lambda + \lambda_i^X + \lambda_j^Y + \lambda_{ij}^{XY} \ , \ \ i=1,\dots,I, \ j=1,\dots,J \ . \tag{4.5}$$

Identifiability constraints are also required for model (4.5). Under the sum to zero identifiability constraints, additional to (4.4) the following constraints hold for the interaction parameters:

$$\sum_{i=1}^{I} \lambda_{ij}^{XY} = \sum_{j=1}^{J} \lambda_{ij}^{XY} = 0 \ . \tag{4.6}$$

Analogous to model (4.1), the (4.4) and (4.6) constraints can be equivalently replaced by constraints equating the last (or first) row and column parameters to zero. For the interaction parameters this would be

$$\lambda_{Ij}^{XY} = \lambda_{iJ}^{XY} = 0, \ \ i=1,\dots I-1, \ \ j=1,\dots J-1$$

(or $\lambda_{1j}^{XY} = \lambda_{i1}^{XY} = 0$, for $i=2,\dots I$, $j=2,\dots J$).

The saturated model (4.5), under (4.4) and (4.6), has as many parameters as the number of cells, i.e., $IJ$. Thus, it does not impose any structure on the underlying association. It just reparametrizes the table's cells in an interpretational meaningful way. The local odds ratios are directly derived from the interaction parameters, since

$$\log \theta_{ij}^L = \lambda_{ij}^{XY} + \lambda_{i+1,j+1}^{XY} - \lambda_{i+1,j}^{XY} - \lambda_{i,j+1}^{XY} \ , \qquad i=1,\dots,I-1, \ \ j=1,\dots,J-1 \ . \tag{4.7}$$

For a simple $2 \times 2$ table and for the first category set to zero constraints ($\lambda_{11}^{XY} = \lambda_{12}^{XY} = \lambda_{21}^{XY} = 0$), it holds

$$\log \theta = \lambda_{22}^{XY} \ .$$

Evidently, the $\lambda^{XY}$ term indeed expresses the association between $X$ and $Y$. Furthermore, model (4.1) is derived by (4.7), setting

$$\lambda_{ij}^{XY} = 0 \ , \ \ i=1,\dots,I, \ j=1,\dots,J \ , \tag{4.8}$$

i.e., by eliminating the association between $X$ and $Y$. This means that (4.1) is *nested* in (4.7). We shall refer in detail to nested models in the context of log-linear models for multi-way tables in Sect. 4.4.

An example of the saturated model's implementation in practice is provided in Sect. 4.2.2.

Overall, log-linear models describe the way the involved categorical variables and their association (if significant) influence the count at each of the $IJ$ cells of the cross-classification of these variables. They are the discrete analogue of analysis of variance, where for each cell of the cross-classification, there is modeled the mean of a continuous variable, instead of a count. The analogy to classical analysis of variance is obvious once the log-linear model's parameters, subject to the sum to zero constraints (4.4) and (4.6), are identified in terms of expected cell frequencies:

$$\lambda = \frac{1}{IJ}\sum_{i,j}\log m_{ij} \tag{4.9}$$

$$\lambda_i^X = \frac{1}{J}\sum_{j}\log m_{ij} - \lambda\ , \quad i=1,\ldots,I\ , \tag{4.10}$$

$$\lambda_j^Y = \frac{1}{I}\sum_{i}\log m_{ij} - \lambda\ , \quad j=1,\ldots,J\ , \tag{4.11}$$

$$\lambda_{ij}^{XY} = \log m_{ij} - \lambda - \lambda_i^X - \lambda_j^Y\ ,\ \ i=1,\ldots,I,\ \ j=1,\ldots,J\ . \tag{4.12}$$

## 4.2 On Inference and Fit of Log-Linear Models

We have seen in Sect. 2.2.1 that the three common sampling schemes for contingency tables are inferential equivalent. For this, the ML estimates of the expected cell frequencies $m_{ij}$ under a log-linear model can be equivalently derived under any of these sampling assumption. For simplicity reasons, the Poisson log-likelihood function is usually considered. Assuming thus an independent Poisson distribution for each cell, $N_{ij} \sim \mathscr{P}(m_{ij})$, and upon observing a sample table $(n_{ij})_{I\times J}$, the Poisson log-likelihood *kernel* $\ell$ (ignoring the constants) is

$$\ell = \sum_{i,j}\left(n_{ij}\log m_{ij} - e^{\log m_{ij}}\right)\ . \tag{4.13}$$

Under a particular log-linear model assumption, substituting $\log m_{ij}$ in (4.13) by the model's formula, $\ell$ will be a function of the log-linear models parameters. Maximizing (4.13) with respect to these parameters, the sets of corresponding *likelihood equations* are derived. Their solution is the set of ML estimates of the parameters and consequently the ML estimates $\hat{m}_{ij}$ of the expected under this model cell frequencies.

Thus, for the independence model, substituting in (4.13) the $\log m_{ij}$ by (4.1) and maximizing with respect to $\lambda_i^X$ and $\lambda_j^Y$, the sets of *likelihood equations* are derived, respectively, as follows:

$$\hat{m}_{i+} = n_{i+}\ ,\ \ i = 1,\dots,I, \qquad \text{and} \qquad \hat{m}_{+j} = n_{+j}\ ,\ \ j = 1,\dots,J. \tag{4.14}$$

Their solution is the ML estimates of the expected cell frequencies $\hat{m}_{ij}$, provided in (2.35). The ML estimates of the $\lambda$ parameters in (4.1), under the sum to zero constraints (4.4), are

$$\hat{\lambda} = \frac{1}{I}\sum_s \log n_{s+} + \frac{1}{J}\sum_s \log n_{+s} - \log n \tag{4.15}$$

$$\hat{\lambda}_i^X = \log n_{i+} - \frac{1}{I}\sum_s \log n_{s+}\ ,\quad i = 1,\dots,I\ , \tag{4.16}$$

$$\hat{\lambda}_j^Y = \log n_{+j} - \frac{1}{J}\sum_s \log n_{+s}\ ,\quad j = 1,\dots,J\ , \tag{4.17}$$

and are obtained by (4.9)–(4.11), substituting the $m_{ij}$'s by the corresponding $\hat{m}_{ij}$'s.

The goodness of fit of a log-linear model is assessed asymptotically by the classical $X^2$ and $G^2$ test statistics, which are under the assumed model asymptotically $\mathscr{X}^2$ distributed with degrees of freedom ($df$) equal to the dimension of the sample space reduced by the number of the parameters estimated under the model. Note that the dimension of the sample space of a contingency table depends on the underlying sampling scheme. Thus, for an $I \times J$ table, for example, it is $IJ-1$ if the table is derived by a multinomial distribution (total $n$ is fixed), while it is $IJ$ when independent Poisson distributions are considered for each cell ($n$ is random). For this, the $\lambda$ of a log-linear model is a parameter only under Poisson sampling (counting for $n$). Consequently, the $df$ of the model are the same under both sampling schemes and the sampling schemes, given $n$ are inferentially equivalent.

For the independence model (4.1), the $X^2$ and $G^2$ tests are (2.36) and (2.37), respectively, with $\hat{m}_{ij}$ given by (2.35) or by (4.1), with the parameters being substituted by their ML estimates (4.15)–(4.17). The saturated model (4.5) fits the data perfectly ($X^2 = G^2 = 0$, $df = 0$).

The classical goodness-of-fit tests $X^2$ and $G^2$ are sensitive in sample size $n$, as already mentioned in Sect. 2.2.2. It is evident that for large $n$, they tend to reject even "good" models. For this, in the framework of log-linear models and in cases of large sample size $n$, a *dissimilarity index* is used that assesses the *practical* significance of the assumed model's lack of fit. This index $\hat{\Delta}$ is common in social sciences applications where also cross-tabulations of large sample sizes occur and is defined as

$$\hat{\Delta} = \frac{1}{2n} \sum_{i=1}^{I} \sum_{j=1}^{J} |n_{ij} - \hat{m}_{ij}| = \frac{1}{2} \sum_{i=1}^{I} \sum_{j=1}^{J} |p_{ij} - \hat{\pi}_{ij}| \tag{4.18}$$

The dissimilarity index $\hat{\Delta}$ ranges in the interval $[0,1]$ and expresses the percentage of observations that have to be moved to different cells in order to achieve a perfect fit. Thus, small values of $\hat{\Delta}$ are indicative of good fit with $\hat{\Delta} < 0.02$ or $< 0.03$ being the limit for a satisfying representation of the data by the assumed model. The sample index $\hat{\Delta}$ estimates the corresponding population index

$$\Delta = \frac{1}{2} \sum_{i=1}^{I} \sum_{j=1}^{J} |\pi_{ij} - \pi^{*}_{ij}| \ ,$$

which measures the dissimilarity between the population probability distribution $\pi = (\pi_{ij})$ and the probability distribution under the assumed model $\pi^{*} = (\pi^{*}_{ij})$. The approximate variance of the statistic $\hat{\Delta}$ and the associated confidence interval has been given by Kuha and Firth (2011). They also provide an updated review of literature on $\hat{\Delta}$, which has a long history.

In practice, log-linear models for two-way (and multi-way) contingency tables are fitted very easily in any software. In `R`, there are several options for getting log-linear models analysis. They can be fitted by `loglin` (of `stats`) or `loglm` (of the `MASS` package). Log-linear models will be fitted for Examples 2.4 and 2.3 by `loglm` in Sects. 4.2.1 and 4.2.2, respectively. However, the predominant approach is to analyze log-linear models in the *generalized linear model* (GLM) framework. Thus, Example 2.4 will be revisited in Sect. 5.4.1, after discussing the GLM and its connection to log-linear models.

### *4.2.1 Example 2.4 (Continued)*

The log-linear model of independence (4.1) will be fitted on Table 2.3 in `R`, by the `loglm` function of the package `MASS`. The parameter estimates derived by `loglm` are under the sum to zero constraints. The data can be either in matrix form or in a data frame.

The data of Table 2.3 are to be found in matrix `natfare`, constructed in Sect. 2.4.1.

After loading the `MASS` package, model (4.1) is then fitted by

```
> I.fit <- loglm( ~ WELFARE + DEGREE, data=natfare)
```

The model formula of the fitted model and the corresponding $G^2$ and $X^2$ goodness-of-fit tests is the standard output, obtained by

```
> I.fit
```

```
Call:
loglm(formula = ~ WELFARE + DEGREE, data = natfare)

Statistics::
                         X^2  df    P(> X^2)
Likelihood Ratio    10.36287   8   0.2404748
Pearson             10.52048   8   0.2303766
```

The goodness-of-fit tests above suggest not to reject the independence model. Thus we conclude that the respondents' belief about national funds for welfare does not depend significantly on their educational level. Recall that independence was visualized in the conditional barplot in Fig. 2.3, where the conditional distributions of educational levels within each category of opinion about welfare spending were similar.

Naturally, we derived the same Pearson's $X^2$ as in Sect. 2.2.6 by the classical `chisq()`. However, in the log-linear models framework, a more detailed interpretation can be extracted by the parameter estimates $\hat{\lambda}_i^X$ and $\hat{\lambda}_j^Y$ in means of (4.2) and (4.3), respectively. All items saved in object `I.fit` can be viewed by `names(I.fit)` and we verify that the parameters' ML estimates, satisfying the sum to zero constraints (4.4), are saved in `I.fit` under `$param`. They can be printed by

```
> I.fit$param
```

```
'(Intercept)'
[1] 3.923732

$WELFARE
too little  about right    too much
-0.2254279    0.1041910   0.1212369

$DEGREE
     LT HS           HS        JColg          BA         Grad
-0.1517607    1.1139057   -0.5802153   0.1528232   -0.5347529
```

Alternatively, they can be saved in new vectors, convenient for further use, like

`> L <- I.fit$param[1]     #` $\hat{\lambda}$

`> L.X <- I.fit$param[2:4] #` $(\hat{\lambda}_1^X, \hat{\lambda}_2^X, \hat{\lambda}_3^X)$

`> L.Y <- I.fit$param[5:9] #` $(\hat{\lambda}_1^y, \ldots, \hat{\lambda}_5^y)$

Thus, it is estimated that in year 2008, it was 1.4 times more probable a responder to believe that the national welfare spending was too much than that it was too little, independent of his educational level, since

$$\frac{\hat{m}_{3j}}{\hat{m}_{1j}} = \exp(\hat{\lambda}_3^X - \hat{\lambda}_1^X) = e^{0.1212-(-0.2254)} = e^{0.347} = 1.41 \ , \quad j = 1, \ldots, 5 \ ,$$

which is computed by

```
> exp(L.X [3]-L.X[1])
```

The ML estimates of the expected under independence cell frequencies are derived by

```
> fitted(I.fit)
```

```
                  DEGREE
WELFARE            LT HS        HS     JColg        BA      Grad
  too little    34.69319  123.0031  22.60314  47.04607  23.65445
  about right   48.23874  171.0283  31.42827  65.41466  32.89005
  too much      49.06806  173.9686  31.96859  66.53927  33.45550
```

The dissimilarity index $\hat{\Delta}$ can now be easily calculated as

```
> D <- sum(abs(natfare-fitted(I.fit)))/(2*sum(natfare))
```

and we find that $\hat{\Delta} = 0.038$, stating that 3.8 % of the observations have to be moved to achieve a perfect fit.

The Pearsonian residuals are given by

```
> residuals(I.fit)
```

```
                  DEGREE
WELFARE              LT HS         HS      JColg         BA       Grad
 too little      1.6724517 -0.6375826 -0.7794780  0.1386103 -0.1351891
 about right    -1.2226377 -0.3092454  0.2780712  0.3175824  1.3612691
 too much       -0.2973437  0.8277514  0.3555753 -0.4378190 -1.3419302
```

but there is no option in the `loglm` framework for getting the standardized residuals. The log-linear models can also be fitted in the GLM framework by `glm`, where the derived output is more informative (for example, the standard errors and significance of the parameters' ML estimates are also provided) and more options are available (standardized residuals calculation is one of them). This example is treated by `glm` in Sect. 5.4.1.

Function `loglm` applies also on a data frame. To construct the data frame for this example, the row and column factors, `WELFARE` and `DEGREE`, respectively, are defined and tied to the vector of observed frequencies `freq` in a data frame, named `nf.frame`, as shown below. The factors are defined for a frequency vector of length $IJ = 15$ that expands the cells of the table by rows.

```
> NI <- 3
> NJ <- 5
> row.lb <- c("too little","about right","too much")
> col.lb <- c("LT HS","HS", "JColg","BA", "Grad")
> WELFARE <- gl(NI,NJ,length=NI*NJ, labels=row.lb)
> DEGREE <- gl(NJ,1,length=NI*NJ, labels=col.lb)
> nt.frame <- data.frame(freq,WELFARE,DEGREE)
```

Then, the model is fitted as

```
> I.fit <- loglm( freq ~ WELFARE + DEGREE, data=nt.frame)
```

leading to the same output and options as described above.

### 4.2.2 Example 2.3 (Continued)

We have already seen in Sect. 2.2.3 that the independence hypothesis is rejected for the cross-classification in Table 2.2 of responders (in GSS2008) subject to their gender and confidence in banks and financial institutions. In the log-linear models framework, model (4.1) is fitted by

```
> I.fit <- loglm( ~ Gender + Conf, data=confinan)
```

giving the fit statistics that we already know from Sect. 2.2.3

```
> I.fit
```

```
Call:
loglm(formula = ~ Gender + Conf, data = confinan)

Statistics::
                       X^2  df     P(> X^2)
Likelihood Ratio  16.39847  2   .0002748643
Pearson           16.34136  2   .0002828258
```

Hence, the interaction between gender and confidence in banks is significant. The interaction between two variables $X$ and $Y$ is denoted in `R` by `X:Y`. Entering the term `Gender:Conf` in the model above, the saturated model is achieved

```
> sat.fit <- loglm( ~ Gender + Conf + Gender:Conf, data=confinan)
```

with $G^2 = X^2 = 0$ and $df = 0$ (perfect fit). Though no structure is imposed on the underlying probability table gaining in parsimony, the parameters are still informative for interpretational purposes. We get

```
> sat.fit$param
```

```
‘(Intercept)’
[1] 5.260226

$Gender
       males      females
-0.09028955   0.09028955

$Conf
  great deal    only some    hardly any
 -0.4147691    0.7337607    -0.3189915

$Gender.Conf
                  $Conf
     Gender   great deal    only some     hardly any
      males   -0.1701995   -0.009293922    0.1794934
    females    0.1701995    0.009293922   -0.1794934
```

In log-linear models, only the highest factor interaction parameters are interpreted. Thus, in presence of $\lambda^{XY}$, the main effects are not interpreted. Odds ratios can be calculated by (4.7) and corresponding conclusions can be expressed. Thus, based on the $\lambda^{XY}$ values of the output above, the odds of having hardly any instead of great confidence to banks is 2.01 times higher for men than for women, computed by

```
> L.XY <- sat.fit$param$Gender.Confinan
> 1/exp(L.XY[1,1]+L.XY[2,3]-L.XY[1,3]-L.XY[2,1])
[1] 2.012516
```

## 4.3 Log-Linear Models for Three-way Contingency Tables

Consider a three-way contingency table, cross-classifying the variables $X$, $Y$, and $Z$. In Sect. 3.2 we discussed on conditional and marginal distributions of such tables and their relations, preparing the field to introduce the various notions of independence in Sect. 3.4.

The hypothesis of *complete independence* of $X$, $Y$, and $Z$ (or mutual independence), defined by (3.16), is equivalently expressed in log-scale as

$$\log \pi_{ijk} = \log \pi_{i++} + \log \pi_{+j+} + \log \pi_{++k}, \quad i = 1,\ldots,I,\ j = 1,\ldots,J,\ k = 1,\ldots,K,$$

which indicates that the logarithmic model of complete independence is

$$\log m_{ijk} = \lambda + \lambda_i^X + \lambda_j^Y + \lambda_k^Z, \quad i = 1,\ldots,I,\ j = 1,\ldots,J,\ k = 1,\ldots,K, \tag{4.19}$$

with the main effect parameters $\lambda_i^X$, $\lambda_j^Y$, and $\lambda_k^Z$ satisfying identifiability constrains as the main effects of the log-linear models for two-way tables, i.e.,

$$\sum_{i=1}^{I} \lambda_i^X = \sum_{j=1}^{J} \lambda_j^Y = \sum_{k=1}^{K} \lambda_k^Z = 0 \quad \text{or} \quad \lambda_1^X = \lambda_1^Y = \lambda_1^Z = 0 \tag{4.20}$$

Analogously, hypothesis (3.17) of *joint independence* of $Y$ from $X$ and $Z$ is in log-scale equivalent to model

$$\log m_{ijk} = \lambda + \lambda_i^X + \lambda_j^Y + \lambda_k^Z + \lambda_{ik}^{XZ}, \quad \forall\, i,j,k. \tag{4.21}$$

Additionally to constraints (4.20), the parameters of model (4.21) satisfy the identifiability constraints

$$\sum_{i=1}^{I} \lambda_{ik}^{XZ} = \sum_{k=1}^{K} \lambda_{ik}^{XZ} = 0 \quad \text{or} \quad \lambda_{1k}^{XZ} = \lambda_{i1}^{XZ} = 0, \tag{4.22}$$

for all possible values of the non-summing subscript ($k$ or $i$).

The model of joint independence (4.21) involves only one two-factor interaction term, the $\lambda^{XZ}$, since $Y$ is joint independent from $X$ and $Z$, but $X$ and $Z$ can be dependent to each other. Obviously, on a three-way table two more models of joint independence can be defined, those having as single two-factor interaction the $\lambda^{XY}$ or the $\lambda^{YZ}$ term.

If $X$ and $Y$ are independent *conditionally* on $Z$, then the underlying probabilities structure is captured in (3.18) as

$$\pi_{ijk} = \pi_{ij|k}\pi_{++k} = \pi_{i+|k}\pi_{+j|k}\pi_{++k} = \frac{\pi_{i+k}\pi_{+jk}}{\pi_{++k}},$$

which is equivalent to the log-linear model:

$$\log m_{ijk} = \lambda + \lambda_i^X + \lambda_j^Y + \lambda_k^Z + \lambda_{ik}^{XZ} + \lambda_{jk}^{YZ}, \quad \forall\, i,j,k. \tag{4.23}$$

The identifiability constraints of this model are (4.20), (4.22), and

$$\sum_{j=1}^{J} \lambda_{jk}^{YZ} = \sum_{k=1}^{K} \lambda_{jk}^{YZ} = 0 \quad \text{or} \quad \lambda_{1k}^{YZ} = \lambda_{j1}^{YZ} = 0, \tag{4.24}$$

for all possible values of the non-summing subscript ($k$ or $j$).

In model (4.23) are present two two-factor interaction terms (from the three possible for a three-way table). The missing interaction term, the $\lambda^{XY}$, is the one responsible for the physical interpretation of the model, signaling missing interaction, in the presence of the other variable. Thus, $X$ and $Y$ are conditionally independent, given $Z$. The model of conditional independence of $X$ and $Z$, given $Y$ (or of $Y$ and $Z$, given $X$) is defined analogously.

Naturally, the next model to be considered is the one having all three possible two-factor interactions. Thus, consider the model

$$\log m_{ijk} = \lambda + \lambda_i^X + \lambda_j^Y + \lambda_k^Z + \lambda_{ij}^{XY} + \lambda_{jk}^{YZ} + \lambda_{ik}^{XZ}, \quad \forall\, i,j,k. \tag{4.25}$$

Additional to (4.20), (4.22), and (4.24), constraints

$$\sum_{i=1}^{I} \lambda_{ij}^{XY} = \sum_{j=1}^{J} \lambda_{ij}^{XY} = 0 \quad \text{or} \quad \lambda_{1j}^{XY} = \lambda_{i1}^{XY} = 0, \tag{4.26}$$

for all possible values of the non-summing subscript ($j$ or $i$), are imposed on the parameters of this model.

It can be easily verified that under model (4.25), all *conditional odds ratios* of the $k$th $XY$ partial table for all pairs $(i,i')$, $(j,j')$ with $i < i'$ and $j < j'$

$$\frac{\pi_{ij|k}\pi_{i'j'|k}}{\pi_{i'j|k}\pi_{ij'|k}},\ i = 1,\ldots,I-1,\ i' = 2,\ldots,I,\ j = 1,\ldots,J-1,\ j' = 2,\ldots,J\ ,$$

are independent of $k$, $k = 1,\ldots,K$. Indeed, we have

$$\log\left(\frac{\pi_{ij|k}\pi_{i'j'|k}}{\pi_{i'j|k}\pi_{ij'|k}}\right) = \log\left(\frac{m_{ijk}m_{i'j'k}}{m_{i'jk}m_{ij'k}}\right) = \lambda_{ij}^{XY} + \lambda_{i'j'}^{XY} - \lambda_{i'j}^{XY} - \lambda_{ij'}^{XY}\ . \tag{4.27}$$

Hence, the $XY$ conditional association does not depend on $k$, i.e., is homogeneous across the levels of $Z$. Analogously it can be proved that also the $YZ$ and $XZ$ conditional associations are homogeneous across the levels of $X$ and $Y$, respectively. For this, model (4.25) is called the models of *homogeneous association*.

If we set $i' = i+1$ and $j' = j+1$ (without loss of generality), the conditional odds ratios above become the $\theta^{XY}_{ij(k)}$ local conditional odds ratios, defined in (3.4), and (4.27) leads to

$$\log \theta^{XY}_{ij(k)} = \lambda^{XY}_{ij} + \lambda^{XY}_{i+1.j+1} - \lambda^{XY}_{i+1.j} - \lambda^{XY}_{i.j+1} \text{ , } i = 1,\dots,I-1, \; j = 1,\dots,J-1, \tag{4.28}$$

independent of $k$. For the conditional odds ratios $\theta^{XZ}_{i(j)k}$ and $\theta^{YZ}_{(i)jk}$ hold analogous results.

Finally, the *saturated* model has an additional term, the three-factor interaction term $\lambda^{XYZ}$ that accounts for the more complex connection of all three variables:

$$\log m_{ijk} = \lambda + \lambda^X_i + \lambda^Y_j + \lambda^Z_k + \lambda^{XY}_{ij} + \lambda^{XZ}_{ik} + \lambda^{YZ}_{jk} + \lambda^{XYZ}_{ijk}, \quad \forall\, i,j,k. \tag{4.29}$$

All terms of saturated model satisfy identifiability constraints, of the type given above. Thus also for the three-factor interaction term it holds

$$\sum_{i=1}^{I} \lambda^{XYZ}_{ijk} = \sum_{j=1}^{J} \lambda^{XYZ}_{ijk} = \sum_{k=1}^{K} \lambda^{XYZ}_{ijk} = 0 \quad \text{or} \quad \lambda^{XYZ}_{1jk} = \lambda^{XYZ}_{i1k} = \lambda^{XYZ}_{ij1} = 0. \tag{4.30}$$

The parameters of the saturated model are in 1-1 correspondence with the $m_{ijk}$. Taking into consideration the appropriate constraints and solving simple equations we can express all $\lambda$ parameters as functions of the $m_{ijk}$'s, analogously to (4.9)–(4.12) for two-way tables.

All possible main effect and interaction terms that can appear in a three-way log-linear model are listed in Table 4.1, along with their number of them being "free," after the identifiability constraints consideration. All these "free" parameters sum to $IJK-1$, which is the dimension of the parameter space when the contingency table $\left(m_{ijk}\right)_{I\times J\times K}$ is multinomial distributed. The fixed term $\lambda$ is considered as a parameter only under the *Poisson* sampling scheme; in which case the number of possible "free" parameters is $IJK$ (in analogy to two-way contingency tables).

All the log-linear models considered so far are of a special type. In all of them, whenever a higher-order effect is in the model, then all possible lower-order effects involving the variables of this higher-order effect term are also in the model. Such models are called *hierarchical log-linear models* and are parsimoniously symbolized by the set of the highest-order terms (with respect to all variables) that define them uniquely. For instance, model $\log m_{ij} = \lambda + \lambda^X_i + \lambda^{XY}_{ij}$ for two-way tables is nonhierarchical, since it includes the term $\lambda^{XY}_{ij}$, without having the term $\lambda^Y_j$. Analogously, the absence of the term $\lambda^Z_k$ makes $\log m_{ijk} = \lambda + \lambda^X_i + \lambda^Y_j + \lambda^{XY}_{ij} + \lambda^{YZ}_{jk}$ nonhierarchical. The hierarchical log-linear models for three-way tables are given in Table 4.2, along with their notation.

**Table 4.1** Number of "free" parameters for each log-linear model term (main effect or interaction) applied on an $I \times J \times K$ contingency table, due to the identifiability constraints

| Term | Number of<br>parameters | Number of<br>"free" parameters | Identifiability<br>constraints |
|---|---|---|---|
| Main effects | | | |
| $\lambda_i^X$ | $I$ | $(I-1)$ | (4.20) for $\lambda_i^X$ |
| $\lambda_j^Y$ | $J$ | $(J-1)$ | (4.20) for $\lambda_i^Y$ |
| $\lambda_k^Z$ | $K$ | $(K-1)$ | (4.20) for $\lambda_i^Z$ |
| Two-factor interactions | | | |
| $\lambda_{ik}^{XZ}$ | $IK$ | $(I-1)(K-1)$ | (4.22) |
| $\lambda_{jk}^{YZ}$ | $JK$ | $(J-1)(K-1)$ | (4.24) |
| $\lambda_{ij}^{XY}$ | $IJ$ | $(I-1)(J-1)$ | (4.26) |
| Three-factor interaction | | | |
| $\lambda_{ijk}^{XYZ}$ | $IJK$ | $(I-1)(J-1)(K-1)$ | (4.30) |

**Table 4.2** Hierarchical three-way log-linear models

| Model | Description | $\log m_{ijk} =$ |
|---|---|---|
| $(X,Y,Z)$ | Independence of $X$, $Y$, $Z$ | $\lambda + \lambda_i^X + \lambda_j^Y + \lambda_k^Z$ |
| | Jointly independence of | |
| $(Y,XZ)$ | $Y$ from $X$ and $Z$ | $\lambda + \lambda_i^X + \lambda_j^Y + \lambda_k^Z + \lambda_{ik}^{XZ}$ |
| $(X,YZ)$ | $X$ from $Y$ and $Z$ | $\lambda + \lambda_i^X + \lambda_j^Y + \lambda_k^Z + \lambda_{jk}^{YZ}$ |
| $(Z,XY)$ | $Z$ from $X$ and $Y$ | $\lambda + \lambda_i^X + \lambda_j^Y + \lambda_k^Z + \lambda_{ij}^{XY}$ |
| | Conditional independence of | |
| $(XZ,YZ)$ | $X$ and $Y$, given $Z$ | $\lambda + \lambda_i^X + \lambda_j^Y + \lambda_k^Z + \lambda_{ik}^{XZ} + \lambda_{jk}^{YZ}$ |
| $(XY,XZ)$ | $Y$ and Z, given $X$ | $\lambda + \lambda_i^X + \lambda_j^Y + \lambda_k^Z + \lambda_{ij}^{XY} + \lambda_{ik}^{XZ}$ |
| $(XY,YZ)$ | $X$ and Z, given $Y$ | $\lambda + \lambda_i^X + \lambda_j^Y + \lambda_k^Z + \lambda_{ij}^{XY} + \lambda_{jk}^{YZ}$ |
| $(XY,XZ,YZ)$ | Homogeneous association | $\lambda + \lambda_i^X + \lambda_j^Y + \lambda_k^Z + \lambda_{ij}^{XY} + \lambda_{ik}^{XZ} + \lambda_{jk}^{YZ}$ |
| $(XYZ)$ | Saturated | $\lambda + \lambda_i^X + \lambda_j^Y + \lambda_k^Z + \lambda_{ij}^{XY} + \lambda_{ik}^{XZ} + \lambda_{jk}^{YZ} + \lambda_{ijk}^{XYZ}$ |

## 4.4 Hierarchical Log-Linear Models for Multi-way Tables

Log-linear models can be defined for contingency tables of dimension higher than three, in a similar manner as for three-way tables. Log-linear models for multi-way tables include higher-order interactions, up to interactions of order equal to the dimension of the table. The number of possible models increases with the dimension of the table, involving the procedure of deciding for the one appropriate to describe the underlying structure of association. In order to impose a structure on model building, especially helpful in model selection, log-linear modeling is usually restricted to the family of *hierarchical log-linear models*.

Furthermore, the presence of nonhierarchical interaction terms in a model causes interpretational inconveniences. For example, in a 4-way table, cross-classifying variables $X$, $Y$, $Z$, and $W$, how can we understand and explain that variable $X$ does not interact with $Y$ (absence of the $\lambda_{ij}^{XY}$ term from the model) but it interacts simultaneously with $Y, Z$, and $W$ (model includes the $\lambda_{ijk\ell}^{XYZW}$ term)? Even among the hierarchical log-linear models, the physical interpretation of the models becomes more involved as the dimension of the table increases. It is easier to understand and interpret a high-dimensional model by focusing on its missing terms. Missing interaction terms refer to variables that are conditional independent and conditional independence statements are easier to understand and express.

To clarify this, consider the hierarchical log-linear model $(XYZ, YW)$ applied on the 4-way table described above. The formula of this model would be

$$\log m_{ijk\ell} = \lambda + \lambda_i^X + \lambda_j^Y + \lambda_k^Z + \lambda_\ell^W + \lambda_{ij}^{XY} + \lambda_{ik}^{XZ} + \lambda_{jk}^{YZ} + \lambda_{j\ell}^{YW} + \lambda_{ijk}^{XYZ} \ .$$

Note that the missing two-factor interaction terms are $XW$ and $ZW$, while $W$ is associated to $Y$ and $X$, $Z$ are associated to each other and both to $Y$ (also in a three-factor interaction). This signals that $X$ and $W$ are conditionally independent, given $Y$ and $Z$. Indeed, the conditional $XW$ log local odds ratios

$$\log \theta_{i(jk)\ell}^{XW} = \log m_{i(jk)\ell} + \log m_{i+1(jk)\ell+1} - \log m_{i+1(jk)\ell} - \log m_{i(jk)\ell+1}$$

under the above model turn out to be

$$\log \theta_{i(jk)\ell}^{XW} = 0, \quad \forall i = 1, \ldots, I-1, \ \ell = 1, \ldots, L-1,$$

for all $j$ $(j = 1, \ldots, J)$ and $k$ $(k = 1, \ldots, K)$, fact that verifies the conditionally independence of $X$ and $W$, given $Y$, $Z$. In a symmetric manner, also $Z$ and $W$ are conditionally independent, given the other two.

For a higher-order example, let the variables $X_1, \ldots, X_7$ be cross-classified to form a $I_1 \times I_2 \times \ldots \times I_7$ contingency table. Then, model $(X_1X_2,\ X_1X_5,\ X_3X_4X_5,\ X_5X_6X_7)$ equates $\log m_{i_1 i_2 \ldots i_7}$ to the sum of the fixed term, $\lambda$, plus the sum of the seven main effects $\lambda_{i_k}^{X_k}$, $k = 1, \ldots, 7$, plus the sum of the eight two-factor interactions $\lambda_{i_k i_\ell}^{X_k X_\ell}$ from the 21 possible (the terms corresponding to the pairs $(k, \ell) =$ (1,2), (1,5), (3,4), (3,5), (4,5), (5,6), (5,7), (6,7) are in the model), plus the three-factor interactions terms $\lambda_{i_3 i_4 i_5}^{X_3 X_4 X_5}$ and $\lambda_{i_5 i_6 i_7}^{X_5 X_6 X_7}$. Observing the terms not included in the model, we can see that variables $X_1$, $X_2$ are jointly independent from $X_3$, $X_4$, conditional on $X_5$, $X_6$, $X_7$.

## 4.5 Maximum Likelihood Estimation for Log-Linear Models

For multi-way tables, the ML estimation procedure for a log-linear model $\mathscr{M}$ is analogous to the procedure followed in Sect. 4.2 for the two-way independence model (4.1). The log-likelihood function is of the (4.13) form, with the subscripts and the indices in the sum appropriately adjusted. Thus, for an $I_1 \times I_2 \times \ldots \times I_s$ table, cross-classifying variables $X_1, X_2 \ldots, X_s$, the kernel of the log-likelihood is

$$\ell(\boldsymbol{\lambda}) = \sum_{i_1, \ldots, i_s} \left( n_{i_1, \ldots, i_s} \log(m_{i_1, \ldots, i_s}) - e^{\log(m_{i_1, \ldots, i_s})} \right) , \tag{4.31}$$

where $m_{i_1,\dots,i_s}$ are the expected frequencies under the assumed model $\mathscr{M}$ and $\boldsymbol{\lambda}$ the vector of all its parameters. It is then maximized with respect to every parameter in $\boldsymbol{\lambda}$ and the set of the associate likelihood equations is derived.

For the three-way hierarchical log-linear model $(XZ, YZ)$, for example, the parameter vector $\boldsymbol{\lambda}$ (4.31) by (4.23) becomes

$$\ell(\boldsymbol{\lambda}) = \sum_{i_1,\dots,i_s} \left( n_{i_1,\dots,i_s}(\lambda + \lambda_i^X + \lambda_j^Y + \lambda_k^Z + \lambda_{ik}^{XZ} + \lambda_{jk}^{YZ}) - e^{\lambda + \lambda_i^X + \lambda_j^Y + \lambda_k^Z + \lambda_{ik}^{XZ} + \lambda_{jk}^{YZ}} \right) .$$

Then, solving $\frac{\partial \ell(\boldsymbol{\lambda})}{\partial \lambda_i^X} = 0$ leads to

$$\hat{m}_{i++} = n_{i++}, \quad i = 1, \dots, I ,$$

which are the likelihood equations corresponding to the $X$ main effect parameters. Analogously, with respect to the $XZ$ interaction parameters, $\frac{\partial \ell(\boldsymbol{\lambda})}{\partial \lambda_{ik}^{XZ}} = 0$ leads to

$$\hat{m}_{i+k} = n_{i+k}, \quad i = 1, \dots, I, \quad k = 1, \dots, K .$$

The remaining sets of likelihood equations are $\hat{m}_{+j+} = n_{+j+}$ $(j = 1, \dots, J)$ and $\hat{m}_{++k} = n_{++k}$ $(k = 1, \dots, K)$, for the $Y$ and $Z$ main effects, respectively, and $\hat{m}_{+jk} = n_{+jk}$ (for all $j,\ k$), corresponding to the $YZ$ interaction.

In general, log-linear models oppose some nice properties regarding their likelihood-based inference. It has been proved that the minimal sufficient statistics of a model $\mathscr{M}$ is the set of sample marginals, corresponding to the highest-order terms in the model, with respect to each variable. Thus, for $(XZ, YZ)$, the sufficient statistics are $(n_{i+k},\ n_{+jk})$, for all $i,\ j,\ k$, while for $(X, YZ)$, they would be $(n_{i++},\ n_{+jk})$, for all $i,\ j,\ k$. The likelihood equations of the model are then equating the sufficient statistics to their corresponding expecting values under $\mathscr{M}$ (Birch 1963).

The ML estimates under the independence model (4.1) are derived in closed-form expression but this is not the case in general. For most log-linear models for higher-dimensional tables, the likelihood equations do not lead to closed-form expressions for the ML estimates and have to be solved iteratively. The first algorithm applied for this was the *iterative proportional fitting* (IPF) algorithm. Predominant is now the *Newton–Raphson* (NR) algorithm, which will be presented in the context of the GLMs (Sect. 5.3.1).

Log-linear models for which closed-form MLEs exist are the *decomposable* models. The joint probability of a decomposable model can be factorized in a closed form in terms of marginal probabilities. This factorization is due to Goodman (1970, 1971c) while the term decomposable was introduced by Andersen (1974). Decomposable log-linear models received special attention in the 1970s and are treated in detail in Bishop et al. (1975, Sect. 3.4). They exhibit nice properties, connected also to *graphical log-linear models* (see Sect. 4.7.2).

## 4.6 Model Fit and Selection

The classical goodness-of-fit statistics to evaluate the fit of a multi-way log-linear model $\mathscr{M}$ are Pearson's $X^2$ and the LR statistic $G^2$, defined for an $I_1 \times I_2 \times \ldots \times I_s$ table as

$$X^2 = \sum_{i_1,\ldots,i_s} \frac{(n_{i_1,\ldots,i_s} - \hat{m}_{i_1,\ldots,i_s})^2}{\hat{m}_{i_1,\ldots,i_s}}, \tag{4.32}$$

$$G^2 = 2 \sum_{i_1,\ldots,i_s} n_{i_1,\ldots,i_s} \log(\frac{n_{i_1,\ldots,i_s}}{\hat{m}_{i_1,\ldots,i_s}}). \tag{4.33}$$

The asymptotic distribution for $X^2$ and $G^2$ under model $\mathscr{M}$ is $\mathscr{X}^2_{d-d_0}$, where $d = \prod_{k=1}^{s} I_k - 1$ is the total number of "free" cells of the table under consideration under the multinomial sampling scheme, $d_0$ the number of "free" parameters of the assumed model $\mathscr{M}$ (overall $\lambda$ is not considered as a parameter), and $\hat{m}_{i_1,\ldots,i_s}$ the ML estimate of the expected under $\mathscr{M}$ frequency for cell $(i_1,\ldots,i_s)$.

The residual degrees of freedom $df = d - d_0$ of the hierarchical log-linear models for three-way tables are given in Table 4.3. In this case $d = IJK - 1$ and $d_0$ is calculated by adding the number of "free" parameters for the terms in model from Table 4.1.

Evaluation of the model fit to the data includes also inspection of the residuals. The types of residuals discussed in Sect. 2.2.4 for two-way tables apply also to tables of higher dimension. The dissimilarity index $\hat{\Delta}$ in (4.18) is also defined for multi-way tables. It does not share the nice properties of $G^2$ but its relative reduction between models $\mathscr{M}_1$ and $\mathscr{M}_2$ can be used to compare practically the models, even if they are not nested.

The number of possible log-linear models increases with the dimension of the table, corresponding to different types of dependencies among the classification

**Table 4.3** Hierarchical three-way log-linear models and their residual $df$

| Model | Formula | $df$ |
|---|---|---|
| $(X,Y,Z)$ | (4.19) | $IJK-I-J-K+2$ |
| $(Y,XZ)$ | (4.21) | $(J-1)(IK-1)$ |
| $(X,YZ)$ | | $(I-1)(JK-1)$ |
| $(Z,XY)$ | | $(K-1)(IJ-1)$ |
| $(XZ,YZ)$ | (4.23) | $K(I-1)(J-1)$ |
| $(XY,XZ)$ | | $I(J-1)(K-1)$ |
| $(XY,YZ)$ | | $J(I-1)(K-1)$ |
| $(XY,XZ,YZ)$ | (4.25) | $(I-1)(J-1)(K-1)$ |
| $(XYZ)$ | (4.29) | 0 |

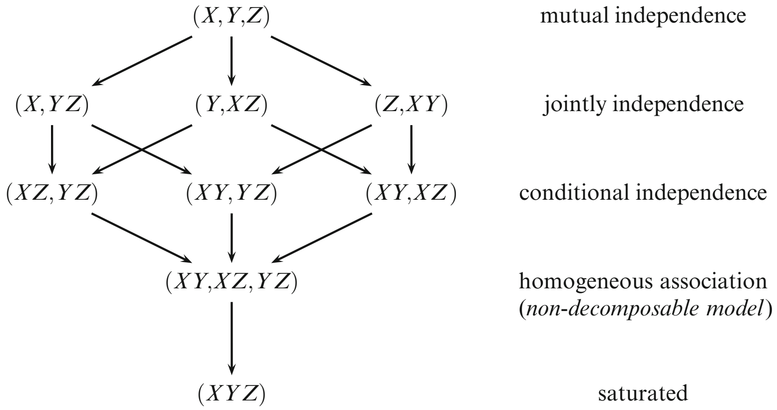


**Fig. 4.1** Sequences of nested models for three-way tables, from the saturated $(XYZ)$ to the model of mutual independence $(X,Y,Z)$

variables. Thus, model selection becomes a basic issue as the dimension of the table rises. The model selection procedure is based on the concept of “nested” models. In general, a model $\mathscr{M}_1$ is *nested* in model $\mathscr{M}_2$, denoted as $\mathscr{M}_1 \subset \mathscr{M}_2$, if $\mathscr{M}_1$ is derived from $\mathscr{M}_2$ by eliminating some of $\mathscr{M}_2$’s parameters. Thus $\mathscr{M}_2$ contains all the terms of $\mathscr{M}_1$ plus at least one more not present in $\mathscr{M}_1$.

Nested models are compared by conditional testing. Model $\mathscr{M}_1$ is more parsimonious than $\mathscr{M}_2$, but for this $G^2(\mathscr{M}_1) \geq G^2(\mathscr{M}_2)$. Given that model $\mathscr{M}_2$ holds, the adequacy of $\mathscr{M}_1$ is tested by

$$G^2(\mathscr{M}_1|\mathscr{M}_2) = G^2(\mathscr{M}_1) - G^2(\mathscr{M}_2)\ , \tag{4.34}$$

which under $\mathscr{M}_1$ is asymptotically $\mathscr{X}^2_{df(\mathscr{M}_1)-df(\mathscr{M}_2)}$ distributed.

The possible sequences of nested models for three-way tables are illustrated in Fig. 4.1. Conditional tests of the type (4.34) can be performed between models connected with arrows, not necessarily directly (see also Tutz 2012, Sect. 12.4).

The log-linear model selection procedure consists of a sequential search between hierarchical nested model, starting from the saturated model and removing terms (one at a time) by conditional testing the significance of the term removed. The process stops and decides for the model for which the next term to be removed leads to a significant increase of the test statistic (4.34). For each level of interaction, say $k$-factor interactions, the order the interaction terms are removed from the model is the order of their significance, less significant being removed first. The procedure described is a *step algorithm* of *backward* elimination. Alternatively, *forward* elimination algorithms start from the model of complete independence and continue to add terms, as long as they improve significantly the fit, according to the conditional test (4.34).

However, we should not let an algorithm decide blindly for the model. Sometimes, the nature of the problem or experimental conditions dictate the presence of nonsignificant terms in the model. For example, suppose in a survey responders are cross-classified according to their educational level ($X_1$), marital status ($X_2$), gender ($X_3$), and age in categories ($X_4$). From the experimental design it is controlled over gender and age, in the sense that the number of males and females participating in the survey is prespecified for each of the $K$ age categories. This means that the table marginals $n_{++i_3i_4}$ for $i_3 = 1,2$ and $i_4 = 1,\dots,K$ are set fixed by the design. If the $X_3X_4$ interaction term is not found to be significant by the selection algorithm and is thus not included in the model, then the corresponding likelihood equations, $m_{++i_3i_4} = n_{++i_3i_4}$ ($i_3 = 1,2$, $i_4 = 1,\dots,K$), are missing. Consequently, the number of males and females assigned by the adopted model to each age group will not agree with the known prespecified numbers. Thus, the $X_3X_4$ interaction term should be included in the model, even if it is nonsignificant. In this case, the $\lambda_{i_3i_4}^{X_3X_4}$ terms signal the underlying product multinomial sampling design and not the physical significance of this interaction.

We have already mentioned the crucial role the concept of conditional independence plays in understanding and recording structures of associations in multi-way contingency tables. An important application of the above described model selection procedure is for testing for conditional independence structures, which is exposed next for three-way tables.

### 4.6.1 Conditional Test of Conditional Independence

In the context of a $I \times J \times K$ contingency table with classification variables $X$, $Y$, and $Z$, if the model of homogeneous association $(XY,XZ,YZ)$ fits the data well, we can test for conditional independence between any two of them, given the third. This test will be *conditional* on homogeneous association. For example, the test of

$$H_0 : X, Y \text{ are independent, conditional on } Z \quad \text{vs.} \quad H_1 : \text{not } H_0$$

can be expressed as

$$H_0 : \text{model } (XZ,YZ) \quad \text{vs.} \quad H_1 : \text{model } (XY,XZ,YZ)\ ,$$

since we already know that the underlying association is homogeneous. The $H_0$ and $H_1$ models are nested; thus, the associated test can be based on the difference

$$G^2(XZ,YZ) - G^2(XY,XZ,YZ) \tag{4.35}$$

which, under $H_0$ and given that model $(XY,XZ,YZ)$ holds, is asymptotically distributed as $\chi^2_{(I-1)(J-1)}$, since $df_{(XZ,YZ)} - df_{(XY,XZ,YZ)} = (I-1)(J-1)$.

For stratified $2 \times 2$ contingency tables, the conditional test (4.35) applied on the $2 \times 2 \times K$ table has $df = 1$ and is analogue to the Mantel–Haenszel test (3.9). This provides an intuitive justification to the fact that the Mantel–Haenszel test works best when the partial associations across the stratification levels are similar (remarked in Sect. 3.3.1).

#### 4.6.1.1 Log-Linear Model Selection for Example 3.3

Reconsidering Example 3.3 of Sect. 3.3.4, if T, F, and C stand for the treatment outcome, the prognostic factor, and the clinic variables, then the homogeneous association log-linear model $(TF, TC, FC))$ and the model of treatment-factor conditional independence within clinic $(TC, FC))$ are of 5 and 6 $df$, respectively, and fitted in MASS as follows

```
: > hom.assoc<-loglm(~Treatment*Prognostic_Factor+
+            Prognostic_Factor*Clinic+Treatment*Clinic, data=dat)
> cond.ind.TF<-loglm(~Prognostic_Factor*Clinic+Treatment*Clinic)
```

The homogeneous association model is adequate, since $G^2(TF, TC, FC) = 7.950$ ($p$-value= 0.159) and $X^2(TF, TC, FC) = 7.894$ ($p$-value= 0.162), very close to the Breslow–Day–Tarone test statistic (3.15), which is equal to $BDT = 7.91$ ($df = 5$, $p$-value=0.161).

The conditional test (4.35) in this case is $G^2(TC, FC) - G^2(TF, TC, FC) = 34.845$, with associated $p$-value=3.570184e-09 ($df = 1$), computed as

```
> DG2 <- cond.ind.TF$deviance - hom.assoc$deviance
> p.value <- 1 - pchisq(DG2, 1)
```

while the corresponding difference in the $X^2$ statistics

```
> DX2 <- cond.ind.TF$pearson - hom.assoc$pearson
```

is $X^2(TC, FC) - X^2(TF, TC, FC) = 33.177$, also indicative of the inappropriateness of the conditional independence model considered (though not asymptotically $\mathscr{X}_1^2$ distributed). Thus the "treatment–prognostic factor" association is homogeneous across the clinics but conditional independence is rejected, based on the above $G^2$ conditional test. Recall that the Mantel–Haenszel test gave for this example $MH = 32.703$ ($df = 1$, $p$-value=1.074e-08), very close to the difference in $X^2$ statistics value above.

### *4.6.2 Log-Linear Model for Example 3.2*

Reconsider the $5 \times 7 \times 2$ contingency table of the example introduced in Sect. 3.2, which is already given in the R array `party.tab`. The three-way log-linear model that describes this data table best will be achieved by the backward stepwise algorithm. The stepwise model selection algorithms, forward or backward, presented in Sect. 4.6, are implemented in R by the `step` function. In `step` the contribution

**Table 4.4** Backward stepwise procedure of log-linear model selection for Example 3.2

```
Start: AIC=140
~D*P*G

                    Df        AIC          LRT      Pr(Chi)
- D:P:G             24        120.82       28.818   0.2271
<none>                        140.00

Step: AIC=120.82
~ D + P + G + D:P + D:G + P:G

                    Df        AIC          LRT      Pr(Chi)
- D:G               4         113.32        0.505   0.9730008
<none>                        120.82
- P:G               6         132.77       23.951   0.0005333 ***
- D:P               24        174.34      101.523   1.650e-11 ***

—
Signif. codes: 0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
Step: AIC=113.32
~ D + P + G + D:P + P:G

                    Df        AIC          LRT      Pr(Chi)
<none>                        113.32
- P:G               6         125.84       24.519    0.000419 ***
- D:P               24        167.41      102.091   1.319e-11 ***
—
Signif. codes: 0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
Call:
loglm(formula = ~ D + P + G + D:P + P:G,
data = party.tab, evaluate = FALSE)

Statistics:
                    X^2         df         P(> X^2)
Likelihood Ratio    29.32320    28         0.3962751
Pearson             29.17456    28         0.4037142
```

of a term is evaluated by the change its removal causes on Akaike's information criterion (AIC) value of the model. The saturated model is applied on `party.tab` and saved under `sat`. Then, with model `sat` as starting point, nonsignificant terms of this model are eliminated by the procedure `step`, as shown below. Recall that we work in library `MASS`.

```
> sat <- loglm(~ D*P*G, data=party.tab)
step(sat, direction="backward", test="Chisq")
```

The derived output is provided in Table 4.4.

Thus, according to the backward stepwise algorithm and based on conditional testings (4.34) between nested hierarchical log-linear models, the three-factor interaction is nonsignificant ($p$-value$= 0.227$). Further on, the two-factor interaction

**Table 4.5** Conditional testing between nested hierarchical log-linear models for Example 3.2

LR tests for hierarchical log-linear models

Model 1:
~ D + P + G
Model 2:
~ D + P + G + D:P
Model 3:
~ D + P + G + D:P + P:G
Model 4:
~ D + P + G + D:P + P:G + D:G
Model 5:
~ D * P * G

| | Deviance | df | Delta(Dev) | Delta(df) | P(> Delta(Dev) |
|---|---|---|---|---|---|
| Model 1 | 155.93392 | 58 | | | |
| Model 2 | 53.84259 | 34 | 102.0913317 | 24 | 0.00000 |
| Model 3 | 29.32320 | 28 | 24.5193932 | 6 | 0.00042 |
| Model 4 | 28.81808 | 24 | 0.5051182 | 4 | 0.97300 |
| Model 5 | 0.00000 | 0 | 28.8180773 | 24 | 0.22705 |
| Saturated | 0.00000 | 0 | 0.0000000 | 0 | 1.00000 |

$DG$ is also nonsignificant with $G^2_{(PG,DP)} - G^2_{(DG,PG,DP)} = 0.505$ and associated ($p$-value$=$ 0.973), based on the $\mathscr{X}^2_4$ approximation for the test statistic. The interaction terms $DP$ and $PG$ are both highly significant with $G^2_{(D,PG)} - G^2_{(PG,DP)} =$ 102.091 and $G^2_{(G,DP)} - G^2_{(DG,PG,DP)} = 24.519$, respectively. Thus, the backward elimination procedure concludes to the $(DP,PG)$ model, i.e., the responder's educational level ($D$) is conditional independent from his/her gender ($G$), given his/her party affiliation ($P$).

The procedure above provides at each stage the value of the AIC for the corresponding model as well. This criterion will be discussed in Sect. 5.3.2.

The successive conditional testings between nested hierarchical log-linear models, from the model of complete independence up to the saturated, adding terms according to their significance order, is summarized in the corresponding analysis of variance table, which is possible in `R` by function `anova`.

```
> I <- loglm(~D+P+G);  as_1 <- loglm(~D+P+G+D:P)
> as_2 <- loglm(~D+P+G+D:P+P:G); as_3 <- loglm(~D+P+G+D:P+P:G+D:G)
> anova(I, as_1, as_2, as_3, sat)
```

In the derived output (in Table 4.5), *deviance* coincides for log-linear models with the $G^2$ test statistic for the corresponding model (see Sect.5.3.2). The conditional $G^2$ values between successive nested models are in column `Delta(Dev)`, followed in next columns by the difference between their $df$ and the asymptotic $p$-value of the associated conditional test.

The mosaic plot of the observed frequencies for this example is provided in Fig. 3.2 (right). This mosaic plot can be enriched by displaying on it the residuals of each cell as well. Thus, the mosaic plot derived by

```
> mosaic(party.tab, gp = shading_Friendly,
         labeling= labeling_residuals)
```

is to be seen in Fig. 4.2 (left). It differs from Fig. 3.2 (right) in that the tiles are colored according to the value of the corresponding residuals for the model of complete independence. Negative significant residuals are red shaded while the positive significant are blue shaded, with the depth of the color strengthening for larger (in absolute value) residuals. We asked additional to label the tiles with the significant residual value, so red-shaded tiles are those with the negative residual values and blue with the positive ones. Cells with nonsignificant residuals are non-shaded (white) with red (dashed) frame for negative residuals and blue (solid) frame for positive ones. Thus, we observe that the highest positive residual corresponds to females with educational level less than high school, who are more political "independent" ("4") than expected under independence. The highest negative residual is for females with a bachelor degree, who are less political "independent" than expected under independence.

The residuals illustrated in the mosaic plot above were for the independence model (default). To refer to residuals of a different model, the output object of the assumed model has to take the position of the data matrix as input in `mosaic()`. Thus, the mosaic plot in Fig. 4.2 (left) can equivalently be obtained as

```
> mosaic(I, gp=shading_Friendly, labeling=labeling_residuals)
```

The residuals of the $(PG, DP)$ model are incorporated in the mosaic plot by

```
> mosaic(as_2, gp=shading_Friendly, labeling=labeling_residuals)
```

The derived plot is provided in Fig. 4.2 (right) and we can easily verify that the Pearsonian residuals for $(PG, DP)$ vary between $-1.58$ and $1.81$, without anyone being significant.

The residuals pictured so far are the Pearsonian residuals (default in `mosaic()`). Alternative option is the deviance residuals, controlled by the option `residuals=`. Thus, the deviance residuals for model $(D, P, G)$ are considered in the mosaic plot as

```
> mosaic(party.tab, gp = shading_Friendly, residuals="deviance",
         labeling= labeling_residuals)
```

For other type of residuals, they have to be calculated ahead and be read in `mosaic()`. This option will be illustrated in the context of GLMs for Example 2.4 in Sect. 5.4.1.

The ML estimates of the expected under the adopted model $(PG, DP)$ cell frequencies are saved in array `MLE` by

```
> MLE <- fitted(as_2)
```

In order to visualize the structure of association dictated by each model, the mosaic plots based on the ML estimates of the expected cell frequencies under characteristic models are provided in Fig. 4.3. In particular, the mosaic plot of the ML estimates under the complete independence model $(P, D, G)$ is in (a), while under $(DP, G)$ and $(DP, PG)$ in (b) and (c), respectively. For comparison reasons, in (d) is located the mosaic plot of the sample values, also given in Fig. 3.2 (right). Observe in (a) that under the complete independence all rectangular tiles

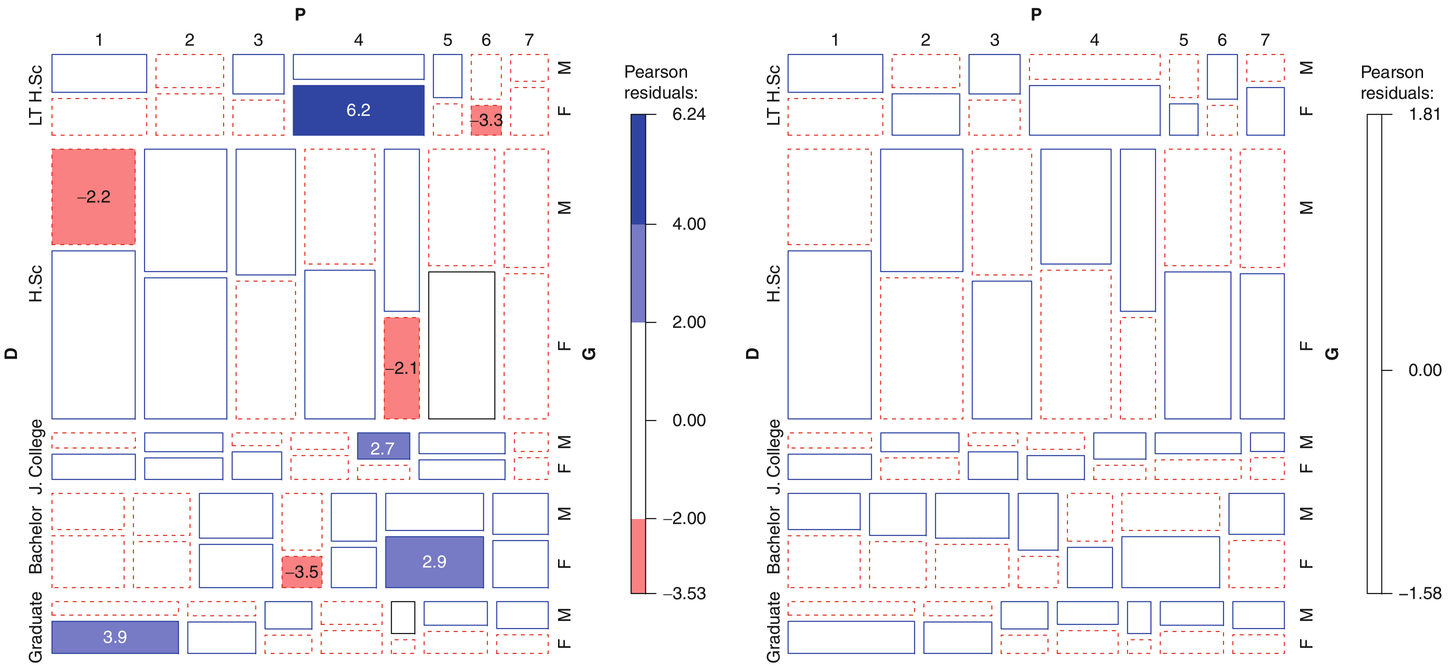


**Fig. 4.2** Mosaic plots for Example 3.2 (Table 3.2), with tiles shaded by the Pearsonian residuals for models $(P,D,G)$ (*left*) and $(DP,PG)$ (*right*)

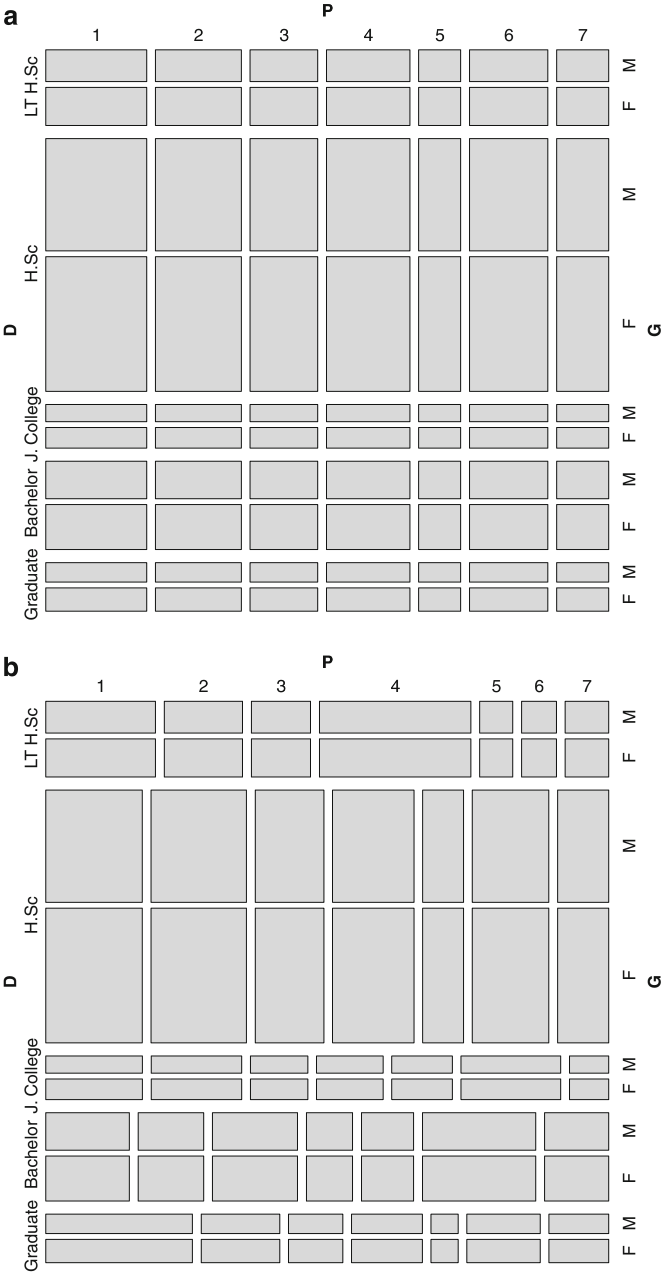


**Fig. 4.3** Mosaic plots of the ML estimates of the expected cell frequencies for Example 3.2 (Table 3.2) under models (**a**) $(P,D,G)$, (**b**) $(DP,G)$, (**c**) $(DP,PG)$ and of the observed cell frequencies in (**d**)

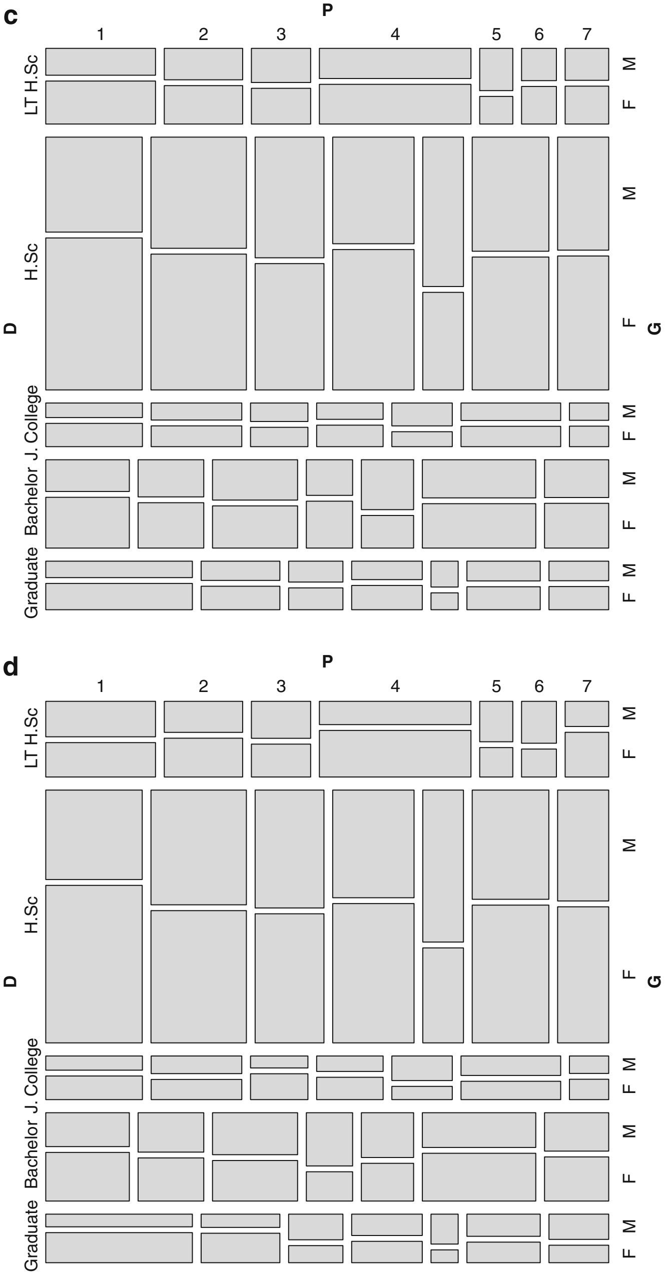


**Fig. 4.3** (continued)

are perfectly aligned. Adding the $DP$ interaction in (b), the alignment of the $DP$ rectangles is disturbed while the within $P$ by $G$ division of the rectangles remains aligned, since the $PG$ term is missing. This alignment is also lost in (c), which resembles closely to the mosaic plot of the observed frequencies in (d).

Mosaic plot in Fig. 4.3c is obtained by

```
> mosaic(MLE)
```

while replacing `MLE` with the array of estimates under $(P, D, G)$ or $(DP, G)$, plots (a) or (b) are derived, respectively.

## 4.7 Graphical Models

Log-linear models can also be defined as graphical models. Not all log-linear models are graphical, as we shall see next. Graphical models are useful whenever the detection of conditional independencies among the involved variables is of interest. They are a wide class of models whose conditional independence structure can be deduced by a graph. In the context of multi-way contingency tables, such graphs for log-linear models were introduced by Darroch et al. (1980), who called them *first order interaction graphs*. They are undirected graphs and in the context of graphical models they are known as *independence graphs* or *conditional independence graphs*. For reasons explained below, we shall use the term *conditional independence graphs*.

In case of high-dimensional contingency tables, graphical log-linear models provide guidance for possible collapsing over one or more classification variables without losing the relevant information. This dimension reduction problem is faced through the factorization criterion and model's decomposability.

To describe graphical models, one needs the basic notion of graph theory, the link between graph theory and probability models, and the group of models which are graphical, that is, whose conditionally independencies can be depicted by a graph.

Graphical models are not in the scope of this book but we shall introduce briefly the basic terminology on undirected graphs (Sect. 4.7.1) and the class of graphical log-linear models in order to connect them with classical log-linear models (Sect. 4.7.2) and use them in the discussion on dimension reduction of multi-way contingency tables by collapsing over one or more of the classification variables (Sect. 4.8).

### *4.7.1 Undirected Graphs*

An undirected graph consists of a finite set of nodes (or vertices) $V$ and a set of edges $E$, connecting some (or all) of the nodes in pairs. Consider, for example, a set of five notes $V = \{v_1, v_2, v_3, v_4, v_5\}$. Then, an undirected graph $\mathscr{G} = (V, E)$ consists

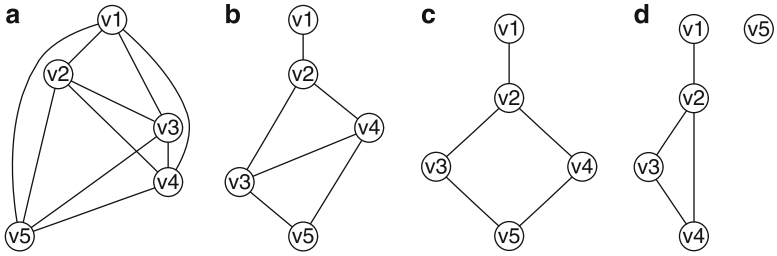


**Fig. 4.4** Undirected graphs $\mathscr{G} = (V, E)$ for $V = \{v_1, v_2, v_3, v_4, v_5\}$ and
(**a**) $E = \{\{v_1,v_2\},\{v_1,v_3\},\{v_1,v_4\},\{v_1,v_5\},\{v_2,v_3\},\{v_2,v_4\},\{v_2,v_5\},\{v_3,v_4\},\{v_3,v_5\},\{v_4,v_5\}\}$,
(**b**) $E = \{\{v_1,v_2\},\{v_2,v_3\},\{v_2,v_4\},\{v_3,v_4\},\{v_3,v_5\},\{v_4,v_5\}\}$,
(**c**) $E = \{\{v_1,v_2\},\{v_2,v_3\},\{v_2,v_4\},\{v_3,v_5\},\{v_4,v_5\}\}$,
(**d**) $E = \{\{v_1,v_2\},\{v_2,v_3\},\{v_2,v_4\},\{v_3,v_4\}\}$

of the five nodes in $V$ and up to ten edges (the elements of $E$) connecting pairwise the nodes. Possible graphs for this setup are provided in Fig. 4.4.

Next, we shall briefly refer to some terminology for undirected graphs. Two nodes $v_1,\ v_2 \in V$ are said to be *adjacent* in $\mathscr{G}$ if the edge $\{v_1, v_2\}$ belongs to $E$, so that they are connected by a line in the corresponding graph. A subset $A \subset V$ is *complete* if all pairs of nodes in $A$ are adjacent. A graph $\mathscr{G} = (V,\ E)$ is *complete* if its set of nodes $V$ is complete. A complete subset of nodes $C$ induces a complete subgraph of the graph of $\mathscr{G}$. If this subgraph becomes incomplete by the addition of a further node, then $C$ is maximally complete and is said to be a *clique*. A *path* is a sequence of distinct nodes $v_1, v_2, \ldots, v_k$ for which each successive pair of nodes are adjacent. Two nodes $v_i$ and $v_j$ are *connected* if there exists a path joining them. A *cycle* is a path with $v_1 = v_k$ and is said to be *chordless* if only the successive pairs of nodes in the cycle are adjacent. An edge of a cycle that connects to non-successive nodes in the cycle is characterized as a *chord*. A graph is *chordal* (or triangulated) if each of its cycles of four or more nodes has a chord. A subset of nodes $B$ *separates* two nodes $v_i$ and $v_j$ if every path joining them contains at least one node from $B$. A subset $B$ separates two subsets of $N$, $A$, and $C$, if it separates every pair of nodes $v_i \in A$ and $v_j \in C$.

## *4.7.2 Graphical Log-Linear Models*

Graphical models are a family of probability models, simplified through *conditional independencies* represented in graphs. Focusing on contingency tables, the family of graphical log-linear models is a subset of the hierarchical log-linear models that utilizes undirected graphs to represent conditional independencies.

The connection between graphical log-linear models for a $K$-way contingency table (with cross-classifying variables $X_1, \ldots,\ X_K$) and undirected graphs is achieved by assuming that (i) the set $V$ of a graph consists of $K$ nodes $(v_1, \ldots,\ v_K)$, one for each classification variable of the table, and (ii) the set of edges $E$ connects some (or all) of the nodes in pairs, indicating a lack of independence between the variables. There is a one-to-one correspondence between models and graphs. In particular, given an undirected graph, the corresponding graphical log-linear model is defined as the hierarchical log-linear model with generators the *cliques* of the graph. For this reason, graphical log-linear models are not always parsimonious models.

Thus, for a five-way table ($K = 5$), the graph provided in Fig. 4.4a is a *complete graph* and corresponds to the saturated model $(X_1X_2X_3X_4X_5)$, while the graphs of Fig. 4.4b–d correspond to the graphical log-linear models $(X_1X_2, X_2X_3X_4, X_3X_4X_5)$, $(X_1X_2, X_2X_3, X_2X_4, X_3X_5, X_4X_5)$, and $(X_2X_3X_4, X_1X_2, X_5)$, respectively. For instance, verify for model $(X_1X_2, X_2X_3X_4, X_3X_4X_5)$ that its three maximal interaction terms correspond to the cliques of the graph in Fig. 4.4b. Only log-linear models with this correspondence are graphical. Thus, the hierarchical log-linear model $(X_1X_2, X_2X_3, X_2X_4, X_3X_4, X_3X_5, X_4X_5)$ is not graphical. Such exclusions from the class of graphical log-linear models ensure the one-to-one correspondence between models and graphs mentioned above.

Conditional independence is the key concept for defining graphical log-linear models. Thus, a representative example of a non-graphical log-linear model is the model of homogeneous association for three-way tables, since it has no conditional independence interpretation. The set of conditional independencies involved in a graphical log-linear model are ruled by three Markov properties, whose description is out of the scope of this section. See Lauritzen (1996) or Højsgaard et al. (2012) for details.

Graphs of graphical log-linear models are interpreted in terms of their missing edges, which are indicative of the underlying conditional independence structure, justifying thus that they are referred to as *conditional independence graphs* (see also Agresti 2013). In particular, the variables corresponding to two nonadjacent nodes in a graph are *conditionally independent*, given the nodes (variables) in the paths connecting them.

Conditional independence is connected to separation of nodes' subsets. If subsets of nodes $A$ and $C$ are separated by subset $B$ in the graph, then, under the corresponding model, variables in $A$ are conditionally independent to variables in $B$, given $C$.

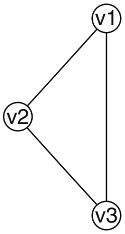
We have seen that an important subset of the hierarchical log-linear models are the *decomposable* models, which lead to MLEs of closed-form expression (see Sect. 4.5). Graphical decomposable log-linear models are graphical models with chordal graphs. The graphical models pictured in Fig. 4.4 are all decomposable except case (c).

Inference for graphical models is beyond the scope of this book. We shall only illustrate them briefly in the following section's examples, mostly to highlight their role in collapsing over one or more classification variables of a high-dimensional contingency table.

For constructing conditional independence graphs and fitting graphical models in `R` one can consult Højsgaard et al. (2012, Chaps. 1 and 2). For example, graphs (a) and (d) of Fig. 4.4 are derived in `gRbase` as shown below.

```
> library(gRbase)
> ag.a <- ug(~v1:v2:v3:v4:v5);  plot(ag.a)
> ag.d <- ug(~v1:v2+v2:v3:v4+v5;  plot(ag.d)
```

Often the association structure of a high-dimensional hierarchical log-linear model (not necessarily graphical) is visualized in terms of a graph, known as *association graphs*. Note however that there is not a one-to-one correspondence between log-linear models and association graphs. More than one log-linear models may have the same graph. For example, considering the graphs in Fig. 4.4 as association graphs of hierarchical log-linear models, (b) is also the graph for the model $(X_1X_2, X_2X_3, X_2X_4, X_3X_4, X_3X_5, X_4X_5)$, while (a) is also (among others) the conditional independence graph of the hierarchical log-linear model including all possible two-factor interactions and none of higher order. In general, a triangle subgraph



of a conditional independence graph expresses the association structure between $X_1$, $X_2$ and $X_3$ of a hierarchical log-linear model containing the corresponding three-factor interaction as well as of a model without this three-factor but all associated pairwise interactions.

The graphs presented so far are *undirected* graphs and are applicable when the classification variables are treated in a symmetric manner in terms of the underlying associations. In the case of one or more response variables, the association structures are visualized through the *directed acyclic graphs* and the *chain graphs*.

## 4.8 Collapsibility in Multi-way Tables

An intuitive way to treat multi-way tables is to reduce their dimension by collapsing over classification variables that are not of direct interest. This way, the collapsing variables are ignored, though they may correspond to covariates that influence the relationship among the variables of interest. Such variables are characterized as *confounding* variables and should be controlled (through conditioning on their levels) instead of ignored. Thus, the association structure among the variables of interest studied on the marginal table produced by collapsing over a confounding variable does not necessarily express their interrelationships but reflects possibly a confounded effect (that of the collapsing variable on the variables of interest).

Furthermore, collapsing over a confounding variable can falsify the structure of the underlying associations, since partial associations can differ substantially (even in direction) from the corresponding marginal ones, as already stated in the context

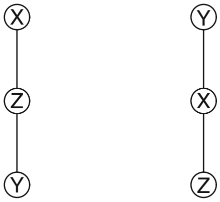


**Fig. 4.5** Conditional independence graphs for models $(XZ,YZ)$ (*left*) and $(XY,XZ)$ (*right*)

of $2 \times 2 \times K$ tables in Sect. 3.2.2. This phenomenon is known as *Simpson's paradox* (Yule 1912; Simpson 1951), which states that the association in a marginal table can be of different direction than conditional association at each corresponding partial table.

Hence, the dimension of a table should not be reduced without ensuring that confounding does not occur. Conditions under which collapsing is possible in three-way tables are exposed next, while a discussion on conditions for multi-way tables follows.

Consider a $I \times J \times K$ table, cross-classifying the variables $X$, $Y$, and $Z$ and suppose that we are interested in the $XZ$ association. The $XZ$ marginal and the $XZ$ conditional (given $Y$) local odds ratios coincide (and thus we could collapse over $Y$ without affecting the $XZ$ association), if either $X$, $Y$ are conditional independent, given $Z$, or $Y$, $Z$ are conditional independent, given $X$, i.e., if the underlying model is the $(XZ,YZ)$ or $(XY,XZ)$, respectively. These patterns of conditional independencies can easily be visualized in the conditional independence graphs of these models in terms of separated variables (see Fig. 4.5). Thus, under both models we can collapse over $Y$, since it is separated from $X$ $(Z)$ by $Z$ $(X)$ for model $(XZ,YZ)$ $(XY,XZ)$. With similar arguments we can verify in Fig. 4.5 (left) that for $(XZ,YZ)$ we could also collapse over $X$ but not over $Z$.

In general for multi-way contingency tables, conditions under which they can be collapsed are provided by Bishop et al. (1975, Chap. 2), who defined the classical parametric collapsibility. It is based on the condition that if a model for a multi-way tables partitions the classification variables into three mutually exclusive subsets $A$, $B$, and $C$, such that $B$ separates $A$ and $C$, then parameters relating variables in $A$ and variables in $A$ to variables in $B$ remain unchanged when collapsing over the variables in set $C$. This means that the association structure of a contingency table is not affected by collapsing over a variable (or a set of variables), only if it is conditionally independent to another variable (or set of variables) of the contingency table, conditioning on the rest of the variables. Since the concept of conditional independence is the fundamental kernel of graphical log-linear models (Sect. 4.7.2) and due to the "separation–conditional independence" connection, graphical models and the associated graphs are extremely useful in detecting patterns of conditional independencies and take decisions for collapsing, especially in high-dimensional contingency tables. For a discussion on the alternative approaches on collapsibility, see Sect. 4.9.4.

**Table 4.6** *DP* ML estimates of the expected under $(PG,DP)$ conditional and marginal local odds ratios for the data in Table 3.2

| $(G)$: males | Political party affiliation $(P)$ | | | | | | |
|---|---|---|---|---|---|---|---|
| Degree $(D)$ | 1 | 2 | 3 | 4 | 5 | 6 | 7 |
| 1: LT high school | 1.385 | 0.955 | 0.462 | 2.289 | 1.790 | 0.533 | |
| 2: High school | 0.948 | 0.878 | 0.980 | 1.818 | 0.876 | 0.591 | |
| 3: Junior college | 0.838 | 2.045 | 0.470 | 1.242 | 1.316 | 1.436 | |
| 4: Bachelor | 0.684 | 0.532 | 2.390 | 0.341 | 1.243 | 1.440 | |
| 5: Graduate | | | | | | | |

### 4.8.1 Collapsing for Example 3.2

Recall that for Example 3.2, model $(PG,DP)$ was selected. As expected due to (3.19) and the discussion above, since under $(PG,DP)$ variables $D$ and $G$ are conditionally independent given $P$, it holds

$$\hat{\theta}^{DP}_{ij(1)} = \hat{\theta}^{DP}_{ij(2)} = \hat{\theta}^{DP}_{ij}\ , \quad i = 1,\ldots,4,\ j = 1,\ldots,6,$$

and their common estimated expected values are provided in Table 4.6.

The estimates of the expected under $(PG,DP)$ conditional *DP* local odds ratios can be calculated in `R` following the procedure described in Sect. 3.2 for the corresponding observed local odds ratios just by replacing the `party.tab` by the `MLE` array. The *DP* partial fitted tables for male and female are respectively

```
> eDP1 <- MLE[„1];  eDP2 <- MLE[„2]
```

and the *DP* fitted marginal (over gender) table is

```
> eDPm <- margin.table(MLE, c(1,2))
```

The $4 \times 6$ table of fitted under $(PG,DP)$ conditional (for males) local odds ratios $\left(\hat{\theta}^{DP}_{ij(1)}\right)$ is then derived by

```
> eOR<-exp(t(matrix(as.vector(C%*%log(as.vector(t(eDP1)))),NJ-1)))
```

Replacing table `eDP1` by `eDP2` and `eDPm` in the command above, the conditional $\left(\hat{\theta}^{DP}_{ij(2)}\right)$ and the marginal $\left(\hat{\theta}^{DP}_{ij}\right)$ fitted tables are produced, respectively, which under $(PG,DP)$ coincide.

Alternatively, $(PG,DP)$ can be fitted as a graphical model in `gRim`. In Sect. 3.2.4 the data were stored in the array `part.tab`. In `gRim`, if the data are in a contingency table format, they need to be defined as table. Thus, the graphical model is fitted as follows.

```
> library(gRim)
> party <- as.table(party.tab)
> graph.PG.DP <- dmod(~P*G+D*P, data=party)
```

The conditional independence graph of the model, given in Fig. 4.6 (right), is derived by

```
> plot(graph.PG.DP)
```

Based on the graph, we observe that we could collapse over gender $(G)$.

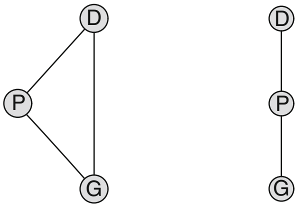


**Fig. 4.6** Conditional independence graphs for Example 3.2 (Table 3.2) for the saturated model $(DPG)$ (*left*) and for the graphical log-linear model $(DP, PG)$ (*right*)

**Table 4.7** Cross-classification of a sample of 2,228 responders according to their age, presence of depression, their gender ($G$), and whether they are living alone

Living alone ($L$): no

| Gender ($G$) | | | Gender ($G$) | | |
|---|---|---|---|---|---|
| Males | Depression ($D$) | | Females | Depression ($D$) | |
| Age ($A$) | No | Yes | Age ($A$) | No | Yes |
| $\leq 45$ | 283 | 16 | $\leq 45$ | 310 | 44 |
| $> 45$ | 270 | 13 | $> 45$ | 262 | 63 |

Living alone ($L$): yes

| Gender ($G$) | | | Gender ($G$) | | |
|---|---|---|---|---|---|
| Males | Depression ($D$) | | Females | Depression ($D$) | |
| Age ($A$) | No | Yes | Age ($A$) | No | Yes |
| $\leq 45$ | 212 | 34 | $\leq 45$ | 291 | 46 |
| $> 45$ | 113 | 63 | $> 45$ | 138 | 70 |

Analogously, collapsing over the educational level ($D$) is also possible but not over the party affiliation ($P$).

### 4.8.2 Example 4.1

Consider the $2 \times 2 \times 2 \times 2$ contingency table produced by cross-classifying a sample of 2,236 responders according to presence of depression ($D$), their gender ($G$), and whether they are living alone ($L$) and are aged above 45 ($A$), given in Table 4.7 (artificial data).

If we are interested in the association between depression and age, the marginal $AD$ sample odds ratio is

$$\hat{\theta}^{AD} = \frac{1,096 \cdot 209}{140 \cdot 783} = 2.09 \, ,$$

indicating that the odds of depression is 2.1 times higher for people over 45. But, looking at the conditional $AD$ sample odds ratio, for all possible combinations of $G$ and $L$, we get

$$\left(\hat{\theta}^{AD(LG)}\right) = \begin{pmatrix} 0.852 & 1.694 \\ 3.476 & 3.209 \end{pmatrix},$$

realizing that Simpson's paradox occurs. Indeed, we observe that for men the $AD$ association changes direction with respect to the living conditions (first column). In particular, for men not living alone, the odds of depression is 1.2 ($= 1/0.852$) times higher for people up to 45 than older while for men living alone it is 3.5 times higher for people older than 45.

Studying the underlying association structure, we proceed to log-linear model selection via the backward stepwise procedure, implemented in `R` as follows.

```
> freq<-c(283,270,16,13,310,262,44,63,212,113,34,63,291,138,46,70)
> names<-list(A=c("<45",">=45"), D=c("no","yes"), G=c("M","F"),
+             L=c("no","yes"))
> dat <- array(freq, c(2,2,2,2), dimnames=names)
> sat <- loglm(∼A*D*G*L, data=dat)
> step(sat, direction="backward" , test="Chisq")
```

The proposed model is the $(ADL,DGL)$ with $G^2 = 3.886$ and $p$-value=0.4216 (based on the $\mathscr{X}_4^2$ approximation), which is graphical.

In the graphical models framework, the saturated model is fitted in `GRim` as

```
> depression <- as.table(dat)
> graph.sat <- dmod(~A*D*G*L, data=depression)
```

while

```
> plot(graph.sat)
```

produces its conditional independence graph, pictured in Fig. 4.7 (left). Based on the saturated model, the backward model selection procedure

```
mod.sel <- backward(graph.sat)
```

```
. BACKWARD: type=decomposable search=all, criterion=aic(2.00), alpha=0.00
. Initial model:  is graphical=TRUE  is decomposable=TRUE
  change.AIC -4.1140  Edge deleted: G A
```

suggests to delete the GA edge (based on the AIC, see Sect. 5.3.2). This leads, as expected, to $(ADL,DGL)$. As a graphical model, it is fitted by

```
> graph.model <- dmod(~A*D*L+D*G*L, data=depression)
```

The derived output

```
> graph.model
```

```
Model: A dModel with 4 variables

graphical    : TRUE        decomposable    : TRUE
-2logL       : 10940.92    mdim            : 11      aic : 10962.92
ideviance    : 171.80      idf             : 7       bic : 11025.72
deviance     : 3.89        df              : 4
```

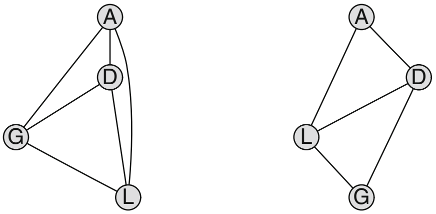


**Fig. 4.7** Conditional independence graphs for Example 4.1 (Table 4.7) for the saturated model $(ADGL)$ (*left*) and for the graphical log-linear model $(GLD, DLA)$ (*right*)

provides information on whether the fitted model is graphical and decomposable as well as on its goodness of fit. `-2logL` is minus twice the maximized log-likelihood and `mdim` the number of parameters in the model. `deviance` and `df` give the likelihood ratio statistic value and the associated degrees of freedom of the fitted model while `ideviance` is $G^2((A,D,G,L)|(ADL,DGL))$, that is, the likelihood ratio statistic for testing independence conditional on the fitted model. The degrees of freedom corresponding to this conditional test are `idf`. The term "deviance" and the other two criteria (AIC and BIC) will be introduced in Sect. 5.3.2.

The fact that $(GLD, DLA)$ is a graphical decomposable log-linear model can easily be verified also by its association graph, derived by

```
> plot(graph.model)
```

and provided in Fig. 4.7 (right). By this graph, we verify that Simpson's paradox may occur when collapsing over $L$. On the other hand, when marginalizing over $G$, the $AD$ and $AL$ association structures are still well estimated, i.e., Simpson's paradox does not occur. Also collapsing over $A$ is possible (the $DG$ and $GL$ association structures are not affected). Thus, the Simpson paradox we noticed for the $AD$ marginal table is due to the marginalization over $L$.

## 4.9 Overview and Further Reading

### *4.9.1 On Log-Linear Models Analysis*

The contribution of Birch (1963) in the analysis of multi-way tables was essential. He was the first who pointed out the equivalence of multinomial and Poisson log-linear models. Furthermore, his result that the ML estimates are the sole estimates that equate a log-linear model's sufficient statistics to their fitted values was a milestone for the log-linear models analysis. He generalized earlier work by Roy, Mitra, and Kastenbaum. It is fair to mention that the first who worked on the interaction structure for multi-way tables was Bartlett (1935), who considered the $2 \times 2 \times 2$ case. A review of these early results is provided by Goodman (1963b, 1964). Fundamental in the multi-way log-linear models establishment was the contribution of Cox, Darroch, Good, Goodman, Bishop, and Fienberg. Seminal to the theoretical development of the topic is the contribution of Haberman, who generalized Birch's results and provided a formal investigation of MLEs for log-linear models and their properties (see Haberman 1974a). He also developed Newton–Raphson iterative algorithms for their fit (Haberman 1973a), which gave estimates of the standard error of the MLEs as well, and made thus their asymptotic significance testing feasible. The algorithm applied by then was the IPF algorithm of Deming and Stephan (1940), adjusted for log-linear models by Bishop (1969), Fienberg (1970a), and Darroch and Ratcliff (1972). Ku and Kullback (1974) approached log-linear models by indicating the analogies to linear models for continuous variables. Lang (1996b) provided a detailed discussion on the comparison of multinomial and Poisson log-linear models. For an extended historical review on the development of the inferential methods for log-linear models we refer to Fienberg and Rinaldo (2007).

Zero frequencies of a contingency table need special consideration and are distinguished between two types, the *sampling* and the *structural zeros*. Sampling zeros refer to cells of low but positive probability that may lead to zero observed frequencies in a certain realization. They are thus *random zeros*, and the corresponding cells are included in the analysis leading to nonzero expected frequencies estimates. Sampling zeros need no special consideration, in general. Traditionally, it has been suggested either to add a small positive constant $\varepsilon$ only to the zero cells (see Grizzle et al. 1969) or to add it always (see Cox 1970b; Goodman 1970). Bishop (1969), Fienberg (1970b), and Goodman (1971b) dealt further with the problem of log-linear models' ML estimation in the presence of sampling zeros. Glonek et al. (1988) proved that for hierarchical log-linear models for multi-way contingency tables, the positivity of the sufficient statistics (i.e., corresponding marginal totals of the table) ensure the existence of the MLEs if and only if the model is decomposable. For non-decomposable models, they discuss the additional conditions required.

Tables with many sampling zeros (*sparse tables*) require special consideration, since the standard asymptotic theory does not apply and technical problems may arise in the estimation procedure. A contingency table with large number of cells and relative small total sample size will contain many zero cells and is called *sparse*. The basic asymptotic theory for testing nonparametric null hypotheses for multinomial data under sparseness assumption has been developed by Holst (1972) and Morris (1975). In case of sparse two-way tables, Mehta and Patel (1983) show that Fisher's exact test and Pearson's $X^2$ can lead to contradictory conclusions. Zelterman (1987) indicated that $X^2$ can show significant bias in testing independence for sparse tables and proposed a new statistic, $D^2$, which is also supported by Haberman (1988) in the context of null hypotheses defining unequal cell probabilities. Goodness-of-fit tests for sparse multinomials are reviewed and compared in Kim et al. (2009).

A class of test statistics for sparse tables with ordered categories are proposed by Burman (2004), which under certain conditions are asymptotically more powerful tests than Pearson's chi-square. Classes of goodness-of-fit tests under sparseness for multidimensional multinomial contingency tables are considered by Maydeu-Olivares and Joe (2005, 2006), based on low-order marginal proportions. Koehler (1986) and Dale (1986) studied the fit of log-linear models on sparse tables. Fienberg and Rinaldo (2012) studied ML estimation in log-linear models, conditions of their existence, and the role of the sampling zeros. An alternative approach to treat sparse tables is the Bayesian (Sect. 10.5). Sparseness is also met in high-dimensional data (see Sect. 10.6). On the other hand, structural zeros are cells of zero probability that must be excluded from the analysis and thus not estimated. Structural zeros will be faced in Sect. 5.5.

Statistical inference for categorical data is mostly asymptotic, based on large sample approximations. For log-linear models, Haberman (1977) provided conditions for the asymptotic normality of linear functions of the MLEs and for the asymptotic chi-squared distribution of Pearson's $X^2$ and the $G^2$ goodness-of-fit statistics. He further pointed out that they remain applicable even if individual cell frequencies are small, provided the sample size and the number of cells of the table are large. The analysis of small sample contingency tables is briefly reviewed in Sect. 10.4.

Friendly (1994) connected mosaic plots to log-linear models, visualizing on mosaic displays beyond the observed cell frequencies (by the area of the cell rectangular) also the residuals (through shadings of the cell areas). For more on visualizing log-linear models via mosaic plots, we refer to Theus and Lauer (1999). Zeileis et al. (2007) visualized on mosaic plots departures of independence in two-way tables and models of conditional independence for three-way tables through residual shadings that code also significance of associations.

Beyond MLEs, the broad class of best asymptotic normal (BAN) estimators has been developed for the multinomial distribution by Neyman (1949), which share optimal large sample properties. In this class belong the *weighted least squares* (WLS) estimators, which are simpler to compute than the MLEs. The basic reference on WLS estimation for categorical data models is Grizzle et al. (1969).

Early contributions on treating misclassification of categorical data are by Bross (1954), facing the problem in 2 × 2 tables, and by Mote and Anderson (1965), considering its effect on $X^2$ tests. Espeland and Odoroff (1985) proposed a log-linear model for misclassified categorical data, fitted by the EM algorithm, generalizing earlier results by Chen (1979). A review on methods of categorical data analysis subject to misclassification is provided by van den Hout and van der Heijden (2002) while Buonaccorsi (2010, Chap. 2) treats two-way tables under misclassification extensively.

### 4.9.2 Residual Analysis: Outlier Detection

Residuals for two-way tables were introduced by Anscombe and Tukey (1963), who proposed graphical and analytical procedures to analyze the residuals. Later on, Cox and Snell (1968) defined residuals in a more general setup and studied their asymptotic properties. They did not deal with contingency tables but discussed problems concerning Poisson and binomial distributed samples. Haberman (1973b) developed methods of residual analysis for log-linear models in two-way tables, complete and incomplete. In particular, he considered the models of independence and quasi-independence, supporting the use of the standardized residuals over the Pearsonian. Pearsonian and standardized residuals were compared in terms of the type I error rates of post hoc cellwise tests for two-way tables under independence and homogeneity models by MacDonald and Gardner (2000) and García-Pérez and Núñez-Antón (2003). The conclusions of MacDonald and Gardner (2000) were in favor of the standardized residuals. García-Pérez and Núñez-Antón (2003) considered the moment-corrected Pearsonian residuals and concluded that they behave the same as the standardized when the marginal distributions of the table are uniform while standardized residuals behave slightly better for peaked marginal distributions. The residuals presented in Sect. 2.2.4 are the most known and widely used. However, a variety of alternative residuals have been suggested in the literature. For example, Brown (1974) and Simonoff (1988) introduced the deleted residuals, for which each expected cell frequency is estimated by the model of quasi-independence, fitted on the data table with this particular cell replaced by a structural zero.

Residuals are a crucial tool for detecting outliers in a contingency table (Simonoff 1988). On outlier detection for two-way tables see, among others, Fuchs and Kennet (1980), Kotze and Hawkins (1984), and Lee and Yick (1999). For outlier detection and measures of influence, see Hastie and Pregibon (1992) and Lee and Fung (1997). For outlier identification in multi-way contingency tables, see Kuhnt (2004) and references cited therein. Alternatively, outliers are treated via algebraic statistics (see Sect. 10.4) by Rapallo (2012).

### 4.9.3 On Graphical Log-Linear Models

The connection of log-linear to graphical models is due to Darroch et al. (1980), while important early contributions are by Edwards and Kreiner (1983) and Wermuth and Lauritzen (1983). Classical reference sources on graphical models are Whittaker (1990) and Lauritzen (1996). Graphical models with missing data are dealt in Lauritzen (1995). Conditional independence graphs for multi-way log-linear models along with more complex multigraphs and the construction of fundamental conditional independencies for non-decomposable log-linear models are discussed in Khamis (2011). For graphical models with causal motivation, distinguishing between explanatory and response variables, see in Sect. 8.4.2.

### 4.9.4 On Collapsibility

Collapsibility, discussed in Sect. 4.8, is an important concept associated with the dimension reduction of multi-way contingency tables without affecting the underlying association structure information. Issues of collapsibility are tied related to Simpson's paradox. For more on Simpson's paradox we refer to Simpson (1951), Blyth (1972), and Samuels (1993).

There exist various notions of collapsibility, starting with the classical *parametric collapsibility* (Bishop et al. 1975, Chap. 2). Further necessary and sufficient conditions of parametric collapsibility, less restrictive than those by Bishop et al. (1975), are provided by Whittemore (1978), who introduced also the term of *strict collapsibility*. Additional to strict collapsibility, Ducharme and Lepage (1986) considered the *pseudo collapsibility* and tested the various types of collapsibility based on the table's nominal odds ratios. A geometric approach for exploring collapsibility is provided by Shapiro (1982). Vellaisamy and Vijay (2007) stated the results of Whittemore in an alternative form using the technique of Möbious inversion and further established new results on collapsibility and strict collapsibility.

Asmussen and Edwards (1983) approached collapsibility via graphical models and defined the *model-collapsibility*. They linked collapsibility to model's decomposability and to the idea of invariance of models when some variables are unobserved. They also showed that model-collapsibility is often equivalent to estimate-collapsibility. The different types of collapsibility conditions are reviewed in Whittaker (1990, Sect. 12.5). Model-collapsibility is also considered in Khamis (2011). Vellaisamy and Vijay (2010) obtained necessary and sufficient conditions for the strict collapsibility based on the interaction parameters of the conditional log-linear model adopted for the layers of the conditioning variables. They considered also the model-collapsibility for hierarchical log-linear models under the conditioning framework and provided connections between the strict and the model-collapsibility. Model- and estimate-collapsibility and their equivalence for conditional graphical models for multi-way contingency tables are considered by Liu and Guo (2012).

### 4.9.5 Information-Theoretic Approach in Contingency Table Analysis

A pioneering approach in categorical data analysis is the *minimum discrimination information* (MDI) approach, based on information theory. It is based on the discrimination information function of Kullback (1959), which is defined on two probability distributions and is a measure of closeness between them.

Based on the principle of MDI, the MDI estimates are BAN estimates obtained by minimizing the discrimination information function between the observed frequencies and the expected under the assumed model or hypothesis. For the cell probabilities of two-way tables with fixed marginals, Ireland and Kullback (1968a) proposed the MDI estimators, illustrating also how their procedure is extended to multi-way tables. Further applications of the results derived in Ireland and Kullback (1968a) and connections to previously ad hoc considered estimators by Fisher (1934) for the 2 × 2 case are given in Ireland and Kullback (1968b).

The MDI approach offers a complete treatment for categorical data inference. The corresponding statistic is asymptotically $X^2$ distributed under the assumed model and is used for testing model fit. Furthermore, the procedure can be applied for testing hypotheses about parameters of the model or linear combinations of them and provides indication of outlier cells and the analysis of information table, in analogy to the analysis of variance table. It is a platform of unified treatment for contingency tables of any order and dimension as well as for categorical data not in a contingency table form. For applications of this approach on contingency tables see Ku and Kullback (1974), references cited therein, and the book by Gokhale and Kullback (1978a). A clarifying short review is given by Gokhale and Kullback (1978b). The MDI approach is identical to the ML approach for *internal constrained problems* (ICP) while for *external constrained problems* (ECP) the two approaches are equivalent in probability under the null hypothesis or the assumed model. For more on ICP and ECP, we refer to Gokhale and Kullback (1978b) and Read and Cressie (1988, Sect. 3.5).

The MDI approach is itself a special case of the *minimum power divergence* approach. The *power divergence* family is introduced by Cressie and Read (1984) and unifies all major approaches considered for discrete multivariate data analysis. Its dynamism lies on the fact that the individual special cases are obtained through a single parameter $\lambda$. The power divergence goodness-of-fit statistic for comparing the frequency vector $\mathbf{Y} = (Y_1, \dots Y_{n_y})'$ to the estimated of the expected under the assumed null hypothesis (or model) $\hat{\propto} = (\hat{\propto}_1, \dots \hat{\propto}_{n_y})'$ is defined as

$$2I^{\lambda}(\mathbf{Y} : \hat{\propto}) = \frac{2}{\lambda(\lambda+1)} \sum_{i=1}^{n_y} Y_i \left[ \left( \frac{Y_i}{\hat{\propto}_i} \right)^{\lambda} - 1 \right], \ -\infty < \lambda < \infty, \ \lambda \neq -1, 0 . \quad (4.36)$$

The cases $\lambda = -1$ and $\lambda = 0$ are defined by the continuous limits of (4.36) for $\lambda \to -1$ and $\lambda \to 0$. It forms a parametric family of goodness-of-fit statistics, controlled by the parameter $\lambda$. Pearson's $X^2$ is (4.36) with $\lambda = 1$, while (4.36) converges to the LR statistic $G^2$ for $\lambda \to 0$. Further, for $\lambda \to -1$, it converges to the MDI statistic mentioned above. The Neyman-modified $X^2$ statistic (Neyman 1949) is obtained for $\lambda = -2$ and the Freeman–Tukey statistic (Freeman and Tukey 1950) for $\lambda = -1/2$. Under the null hypothesis tested and under certain regularity conditions, (4.36) is asymptotically $X^2$ distributed and all members of this goodness-of-fit statistics family are asymptotically equivalent. In terms of test power and of small sample approximation, Cressie and Read (1984) suggested the value $\lambda = 2/3$. Statistical inference for multivariate discrete data based on the power divergence is studied extensively in Read and Cressie (1988), also under sparseness assumptions.

Associated with statistic (4.36) is the *power divergence measure*, which measures the divergence of two probability distributions. If $\pi = (\pi_1, \dots \pi_K)'$ and $\mathbf{q} = (q_1, \dots q_K)'$ are two probability vectors, then the power divergence specifies their divergence as

$$2I^{\lambda}(\pi : \hat{\mathbf{q}}) = \frac{2}{\lambda(\lambda+1)} \sum_{i=1}^{K} \pi_i \left[ \left( \frac{\pi_i}{q_i} \right)^{\lambda} - 1 \right], \ -\infty < \lambda < \infty, \ \lambda \neq -1, 0 , \quad (4.37)$$

with the cases $\lambda = -1$ and $\lambda = 0$ being defined as above.

The power divergence belongs to the even broader family of $\phi$-divergences. For $\pi$ and $\mathbf{q}$ as above, the $\phi$-divergence between $\pi$ and $\mathbf{q}$ (or Csiszar's measure of information in $\mathbf{q}$ about $\pi$) is defined by

$$\mathrm{I}^{\mathrm{C}}(\pi,\mathbf{q}) = \sum_{i=1}^{K} q_i \phi(\pi_i/q_i), \tag{4.38}$$

where $\phi$ is a real-valued strictly convex function on $[0,\infty)$ with $\phi(1) = \phi'(1) = 0$, $0\phi(0/0) = 0$, $0\phi(y/0) = \lim_{x\to\infty}\phi(x)/x$ (see Pardo 2006). Setting $\phi(x) = x\log x$, (4.38) is reduced to the Kullback–Leibler divergence measure that corresponds to the LR statistic $G^2$. For $\phi(x) = (1-x)^2$, Pearson's divergence is derived, related to Pearson's $X^2$ statistic. If $\phi(x) = \frac{x^{\lambda+1}-x}{\lambda(\lambda+1)}$, (4.38) becomes the power divergence measure (4.37).

For $\phi$-divergence-based inference and for special applications to log-linear models and categorical data analysis, we refer to Pardo (2006), references therein, and to Martìn and Pardo (2008). Minimum power divergence and minimum $\phi$-divergence estimators generalize the MLEs, retaining their properties and meanwhile exhibiting robustness properties (see Basu et al. 1998 and Pardo 2006). In Sect. 7.4 we discuss generalized association models, connected to $\phi$-divergence.