International Statistical Review (1998), **66**, 2, 181–199, Printed in Mexico

On the Application of Conditional Independence to Ordinal Data

Nanny Wermuth[1] and D.R. Cox[2]

[1]*Center of Survey Research and Methodology, ZUMA, B2,1, 68159 Mannheim, Germany*
[2]*Nuffield College, Oxford, OX1 1NF, UK*

Summary

A special log linear parameterization is described for contingency tables which exploits prior knowledge that an ordinal scale of the variables is involved. It is helpful, in particular, in guiding the possible merging of adjacent levels of variables and may simplify interpretation if higher-order interactions are present. Several sets of data are discussed to illustrate the types of interpretation that can be achieved. The simple structure of the maximum likelihood estimates is derived by use of Lagrange multipliers.

Key words: Contingency tables; Contrast matrices; Design matrices; Graphical Markov Models; Level comparisons; Log linear parameters; Maximum likelihood estimation.

1 Introduction

Many different methods have been proposed for the analysis of ordinal data; for systematic reviews, see the books of Agresti (1984, 1990) and Clogg & Shidadeh (1994) and also McCullagh & Nelder (1989, pp.151–155). However, despite recent intensive interest in using the notion of conditional independence (see Wermuth & Cox, 1998a; 1998b) to simplify multivariate systems (Edwards, 1995; Cox & Wermuth, 1996; Lauritzen, 1996; Wermuth, 1998) such possibilities have not been set out in detail for ordinal variables.

The outline of the present paper is as follows. First we present a motivating 5×5 example and give for general two-way tables a detailed account of the suggested special log linear parameterization. It has the advantage of clear interpretation because the vanishing of each interaction term indicates independence in a specific associated 2×2 subtable. Vanishing of a suitable set of such parameters suggests the possible merging of adjacent levels of variables. After an explanation of the model, analyses of further two-way tables are shown.

We then present a direct extension to higher dimensional tables. It is demonstrated by examples that this is particularly useful whenever the log linear model for the whole table is such that a table of two-way frequencies is a component of the minimal sufficient statistics or the same type of independence for a variable pair holds at all level combinations of the other variables.

Finally we give a new derivation via Lagrange multipliers of the simple structure of the maximum likelihood estimates under such models and relate the result to more general considerations.

2 The $I \times J$ Table

2.1 An Example of Ordinal Variables in a 5 × 5 Table

From two general social surveys of adults in West Germany (Central archive, 1993) the observed counts shown in Table 1 were obtained on type of formal schooling and age group for all respondents in years 1991 and 1992.

Table 1

Observed and fitted counts, West Germany 1991/92; n = 3673

Type of	Age group					
schooling	18 - 29	30 - 44	45 - 59	60 - 74	>74	sum
basic, incomplete	12	13	12	20	7	64
	(7.7)	(17.7)	(17.2)	(16.4)	(4.9)	
basic, complete	215	507	493	460	137	1812
	(219.3)	(502.3)	(487.8)	(464.4)	(138.3)	
medium	277	300	192	126	38	933
	(277)	(300)	(192)	(126.4)	(37.6)	
upper medium	52	91	47	15	6	211
	(52)	(91)	(47)	(16.2)	(4.8)	
intensive	233	225	102	74	19	653
	(233)	(225)	(102)	(71.7)	(21.3)	
sum	789	1136	846	695	207	3673

The displayed fitted counts, which agree well with the observed counts, correspond to both of the following constraints: (1) independence of schooling from age group for the last two categories of age above 60 years and (2) independence of age group from the two lowest categories of type of schooling. These categories mean that at most basic, i.e. compulsory, education has been completed. A good fit of the observed values to this special *reduced model* points to the possibility of combining levels 1 and 2 of the row variable and levels 4 and 5 of the column variable for a simplified interpretation of the association or, to state it differently, of concentrating on the association in the reduced 4 × 4 table.

As will be explained in detail below, under the chosen reduced model all seven 2 × 2 subtables in the first two rows and in the last two columns of the 5 × 5 table have an odds ratio of one, and all other fitted counts match the observed counts.

Table 2

Observed and fitted column percentages for Table 1

Type of	Age group				
schooling	18 - 29	30 - 44	45 - 59	60 - 74	>74
basic, incomplete	2	1	1	3	3
	(1)	(2)	(2)	(2)	(2)
basic, complete	27	45	58	66	66
	(28)	(45)	(58)	(67)	(67)
medium	35	26	23	18	18
	(35)	(26)	(23)	(18)	(18)
upper medium	7	8	6	2	3
	(7)	(8)	(6)	(2)	(2)
intensive	30	20	12	11	9
	(30)	(20)	(12)	(10)	(10)
	100%	100%	100%	100%	100%

The column percentages in Table 2 show that the observed conditional distributions of schooling are very close for age groups 60–74 and over 74 years and that they are identical for the fitted counts.

Table 3

Observed and fitted row percentages for Table 1

Type of schooling	Age group 18 - 29	30 - 44	45 - 59	60 - 74	>74	
basic, incomplete	19	20	19	31	11	100%
	(12)	(28)	(27)	(26)	(8)	
basic, complete	12	28	27	25	8	100%
	(12)	(28)	(27)	(26)	(8)	
medium	30	32	21	14	4	100%
	(30)	(32)	(21)	(14)	(4)	
upper medium	25	43	22	7	3	100%
	(25)	(43)	(22)	(8)	(2)	
intensive	36	34	16	11	3	100%
	(36)	(34)	(16)	(11)	(3)	

Most of the observed row percentages for age groups in the first two rows of Table 3 look similar, but some of the pairwise differences for these two categories 'basic schooling incomplete' versus 'basic schooling complete' appear to be quite large. By use of associated tests for goodness of fit these are nevertheless judged to be mere random fluctuations. The reason is that there are relatively few respondents at the first level of the row variable so that corresponding counts in row 1 of Table 1 are estimated only imprecisely. Again, for the fitted counts the two conditional distributions by construction do not differ.

Table 4

Observed and fitted studentized γ-parameters for Table 1

Levels, subtables row	column	Studentized γ-terms observed	fitted	Levels, subtables row	column	Studentized γ-terms observed	fitted
12	12	1.9	0	12	34	−1.6	0
23	12	−6.7	−6.5	23	34	−2.7	−3.0
34	12	2.5	2.5	34	34	−2.3	−2.3
45	12	−3.0	−3.0	45	34	2.5	2.4
12	23	0.1	0	12	45	−0.3	0
23	23	−3.7	−3.7	23	45	0.1	0
34	23	−1.1	−1.1	34	45	0.6	0
45	23	−0.6	−0.6	45	45	−0.8	0

Table 4 contains *studentized interactions* corresponding to our parameterization for ordinal variables, that is they are log odds ratios for neighbouring levels divided by their estimated standard deviation. Hence for large samples and under the hypothesis of a zero log odds ratio a studentized value is to be treated like a value from a standard Gaussian distribution in which absolute values larger than 3 are very unlikely and therefore point to a dependence in the corresponding 2×2 subtable. For instance, for the row variable (type of schooling) at the first two levels, 1, 2, and for the column variable (age group) at the last two levels, 4,5, the observed studentized value of the log odds ratio is -0.3 and it is equal to zero under the reduced model.

2.2 Interactions Based on Level Comparisons

For a general $I \times J$ contingency table of variables A, B we write for probabilities $\pi_{ij} = \Pr(A = i, B = j)$ and denote observed counts by n_{ij} for levels $i = 1, \ldots, I$ of A and $j = 1, \ldots, J$ of B.

We define an IJ by 1 vector of log linear parameters γ in terms of a suitable $IJ \times IJ$ matrix, C_{IJ}, and a column vector of probabilities π^{AB} as

$$\gamma^{AB} = C_{IJ} \log \pi^{AB}.$$

Because C_{IJ} contains weights defining contrasts of log probabilities we call C the *contrast matrix*. For instance, in a 3×2 table $\gamma^{AB} = C_{32} \log \pi^{AB}$ is for our special representation

$$\begin{pmatrix} \gamma_{-}^{AB} \\ \gamma_{12}^{A} \\ \gamma_{23}^{A} \\ \gamma_{12}^{B} \\ \gamma_{12.12}^{A.B} \\ \gamma_{23.12}^{A.B} \end{pmatrix} = \begin{pmatrix} 1 & 1 & 1 & 1 & 1 & 1 \\ -1 & 1 & 0 & -1 & 1 & 0 \\ 0 & -1 & 1 & 0 & -1 & 1 \\ -1 & -1 & -1 & 1 & 1 & 1 \\ 1 & -1 & 0 & -1 & 1 & 0 \\ 0 & 1 & -1 & 0 & -1 & 1 \end{pmatrix} \begin{pmatrix} \log \pi_{11} \\ \log \pi_{21} \\ \log \pi_{31} \\ \log \pi_{12} \\ \log \pi_{22} \\ \log \pi_{32} \end{pmatrix}.$$

For example the interaction $\gamma_{23.12}^{A.B}$ is then defined by

$$\gamma_{23.12}^{A.B} = \log \pi_{21} - \log \pi_{31} - \log \pi_{22} + \log \pi_{32} = \log \frac{\pi_{21}\pi_{32}}{\pi_{31}\pi_{22}},$$

the log odds ratio in subtable 2,3 of A and 1,2 of B.

For variables with more categories and probabilities listed such that the levels of A change fastest, i.e.

$$\pi^{AB} = (\pi_{11}, \pi_{21}, \ldots, \pi_{I1}, \pi_{12}, \pi_{22}, \ldots, \pi_{I2}, \pi_{13}, \ldots, \pi_{IJ})^T,$$

the log linear parameters are generated in the following type of lexicographical order

$$\gamma_{-}^{AB}, \gamma_{12}^{A}, \gamma_{23}^{A}, \ldots, \gamma_{I-1,I}^{A}, \gamma_{12}^{B}, \gamma_{12.12}^{A.B}, \gamma_{23.12}^{A.B}, \ldots, \gamma_{I-1,I.12}^{A.B}, \gamma_{13}^{B}, \gamma_{12.13}^{A.B}, \ldots, \gamma_{I-1,I.J-1,J}^{A.B}.$$

In general we denote by $\gamma_{ii'.jj'}^{A.B}$ the log odds in the 2×2 subtable of levels i, i' of A and levels j, j' of B, the variable names being omitted if the context is clear. These are local odds ratios (Goodman, 1979a, b) to be contrasted with global odds ratios (Williams & Grizzle, 1972; Dale, 1984; Molenberghs & Lesaffre, 1994). If merging of adjacent levels is of potential interest, local ratios are likely to be more helpful, as in our examples below, whereas if the levels represent a well-judged spacing of an underlying continuous latent variable, global ratios may prove a better base for study of the underlying continuous distribution.

The matrix of contrasts C_{IJ} for the log probabilities may be computed via the Kronecker product of I by I and J by J matrices C_I and C_J of level comparisons for the row and the column variables, respectively, where e.g.

$$C_4 = \begin{pmatrix} 1 & 1 & 1 & 1 \\ -1 & 1 & 0 & 0 \\ 0 & -1 & 1 & 0 \\ 0 & 0 & -1 & 1 \end{pmatrix}, \quad C_3 = \begin{pmatrix} 1 & 1 & 1 \\ -1 & 1 & 0 \\ 0 & -1 & 1 \end{pmatrix}, \quad C_2 = \begin{pmatrix} 1 & 1 \\ -1 & 1 \end{pmatrix}.$$

We use the left Kronecker product to get

$$C_{IJ} = C_J \otimes C_I$$

in which the second matrix, C_I, is multiplied in turn by the elements of the first matrix, C_J, to give for instance for $I = 3$ and $J = 2$

$$C_{32} = C_2 \otimes C_3 = \begin{pmatrix} 1 & 1 & 1 & 1 & 1 & 1 \\ -1 & 1 & 0 & -1 & 1 & 0 \\ 0 & -1 & 1 & 0 & -1 & 1 \\ -1 & -1 & -1 & 1 & 1 & 1 \\ 1 & -1 & 0 & -1 & 1 & 0 \\ 0 & 1 & -1 & 0 & -1 & 1 \end{pmatrix}.$$

To illustrate further the meaning of interactions based on level comparisons we choose a 3×3 table and show the four special types of independence that result by having zero log odds ratios in neighbouring 2×2 subtables.

(1) There is independence in the 2×2 subtable with levels 2,3 of A and levels 1,2 of B ($A_i \perp\!\!\!\perp B_j$ given $i = 2, 3;\ j = 1, 2$) or
$\gamma_{23.12} = 0$.

(2) There is independence in the 3×2 subtable with all levels of A and levels 1,2 of B ($A \perp\!\!\!\perp B_j$ given $j = 1, 2$) or
$\gamma_{12.12} = \gamma_{23.12} = 0$.

(3) There is independence as in 2. and, in addition, in the 2×3 subtable with all levels of B and levels 1, 2 of A ($A \perp\!\!\!\perp B_j$ given $j = 1, 2$ and $A_i \perp\!\!\!\perp B$ given $i = 1, 2$) or
$\gamma_{12.12} = \gamma_{23.12} = \gamma_{12.23} = 0$.

(4) Variable A is independent of variable B ($A \perp\!\!\!\perp B$) or
$\gamma_{12.12} = \gamma_{23.12} = \gamma_{12.23} = \gamma_{23.23} = 0$.

For a simplified description of the association between the variables it is possible to merge levels 1 and 2 of B in case (2) and of both variables in case (3) .

The key point is however that the four degrees of freedom defining the dependence between A and B in the 3×3 table are captured by four log odds ratios with individual specific implications especially appropriate for ordinal variables.

2.3 Log Linear Parameters Fitted by Maximum Likelihood

Maximum-likelihood estimates of the log linear parameters are

$$\hat{\gamma}^{AB} = C \log \hat{\pi}^{AB}$$

with fitted probabilities expressed with fitted counts, $\hat{\pi}_{ij} = \hat{m}_{ij}/n$, where

$$(i)\quad \hat{m}_{ij} = n_{ij} \text{ if } ij \in S$$

$$(ii)\quad \hat{m}_{i.} = n_{i.},\quad \hat{m}_{.j} = n_{.j} \text{ for all } i, j$$

$$(iii)\quad \frac{\hat{m}_{ij}\hat{m}_{i'j'}}{\hat{m}_{i'j}\hat{m}_{ij'}} = 1 \quad \text{if } \gamma^{AB}_{ii'.jj'} = 0.$$

Here S denotes the subset of cells of the contingency table unaffected by zero constraints (iii), i.e. not corresponding to any ii', jj' with $\gamma^{A.B}_{ii'.jj'} = 0$ (see also Section 4).

Note that under an unconstrained model with S being of size $(I-1)(J-1)$, which is often called the *saturated model*, each fitted count coincides with the observed count. On the other hand if all $(I-1)(J-1)$ interaction terms $\gamma^{AB}_{ii'.jj'}$ are zero, so that S is empty and $A \perp\!\!\!\perp B$ holds, only the marginal counts are to match the corresponding fitted marginal counts.

The parameterization suggested here permits the formulation of more subtle forms of indepen-

dency than would be possible by effect coding, i.e. symmetric constraints on log odds ratios, most appropriate for nominally scaled variables (Bishop *et al.*, 1975), or as would be possible by indicator coding, i.e. with base-line constraints, most appropriate if one of the levels is indeed a natural base-line category. See Wermuth & Cox (1992) for a more detailed discussion and for relations between different codings and with design (or coding) matrices, the inverses of contrast matrices.

The estimated covariance matrix of the above estimates $\hat{\gamma}$ may be expressed as a conditional covariance matrix given c, with c being the indices of the constraint cells in the vector π (Cox & Wermuth, 1990):

$$\Sigma_{ff.c} = \Sigma_{ff} - \Sigma_{fc}\Sigma_{cc}^{-1}\Sigma_{cf}$$

where Σ is expressed in terms of the contrast matrix, C, and the estimated covariance matrix of fitted probabilities $\hat{\text{cov}}(\log \hat{\pi})$ as

$$\Sigma = C\,\hat{\text{cov}}(\log \hat{\pi})\,C^T$$

$$\hat{\text{cov}}(\log \hat{\pi}) = n^{-1}(\mathcal{D} - ee^T)$$

with f denoting the remaining indices in the vector, e being a column vector of ones, n the sample size, and $\mathcal{D}$ a diagonal matrix formed from reciprocals of the estimated probabilities. To state it differently, $\Sigma_{ff.c}$ is the submatrix for the unconstrained parameters in positions (f, f) after sweeping Σ on c.

If the probabilities are estimated under the saturated model this gives an approximation to the maximum likelihood estimate of the covariance matrix under the reduced model, while the precise maximum likelihood estimate under the reduced model results if the probabilities are estimated under this reduced model.

2.4 A Note on Computation

The actual fitting may be performed as for other log linear contingency table models with iteratively reweighted least squares or iterative proportional fitting algorithms. Also, a cyclic fitting algorithm may be used. For covariance selection models such an algorithm had been proposed by Dempster (1972), a programmed algorithm was given in Wermuth & Scheidt (1977) and its convergence properties studied by Speed & Kiiveri (1986). While for covariance selection the cyclic fitting algorithm involves repeated inversion of covariance matrices it is much simpler here: if the odds ratio in any of the 2×2 tables with levels ii' of A and jj' of B is replaced by the one expected under independence then the corresponding log linear interaction parameter $\hat{\gamma}_{ii'.jj'}$ is zero.

The cyclic fitting algorithm to compute the fitted counts can be described as follows. At step 1 the starting values $\tilde{m}^{1,1}$ are the observed counts. At step s, $(s = 1, \ldots, d_r)$ of cycle T we have notional counts $\tilde{m}^{T,s}$ and modify them so that $\tilde{m}^{T,s+1}$ has independence in the relevant 2×2 subtable corresponding to the $(s + 1)$'th constraint. After the last of d_r steps of an iteration cycle we obtain $\tilde{m}^{T+1,1}$. If the change compared to $\tilde{m}^{T,1}$ is small enough the iteration stops.

To avoid complications related to observed zero cells we have found it useful to add 0.01 to all cells in such a case. This increases the formal sample size by 1% of the total number of cells of the table.

Sometimes closed form maximum likelihood estimation is possible, which can also be achieved by simple computational steps. For instance, two steps are needed if, as in the initial example of Section 2, there is independence of the row and of the column variable for a subset of levels of the other variable. Then the independency of the type $A \perp\!\!\!\perp B_j$ can be directly fitted and to the resulting fitted counts the independency of type $A_i \perp\!\!\!\perp B$ (or vice versa). Corresponding tests are used within the software DIGRAM when looking for mergeable levels (Kreiner, 1990).

Wilks's (1938) likelihood ratio test of goodness of fit of a reduced model against the saturated model gives a statistic which has approximately a chi-squared distribution on as many degrees of freedom as log linear interaction terms are set to zero. In chi-squared distributions values smaller

than the degrees of freedom occur with probability larger than one half, hence indicate a good or excellent goodness of fit.

The test statistic compares the fitted counts to the observed counts via

$$\chi^2 = 2(\log n_{ij} - \log \hat{m}_{ij}).$$

Note that for interactions based on level comparisons nonzero contributions can arise only from cells affected by zero constraints (see also Section 4).

Provided there is a well fitting reduced model, say $\mathcal{M}_*$, with fitted counts $\hat{m}^*_{ij}$, the goodness of fit of a further model with additional constraints and fitted values $\hat{m}_{ij}$ can be tested as above with n_{ij} replaced by $\hat{m}^*_{ij}$.

2.5 Further Examples of I × J Tables

Before turning to a fuller discussion of the initial example in the context of a higher dimensional table we now discuss briefly more analyses and interpretation of three two-way contingency tables in terms of level comparisons.

The 2 × 5 contingency table for 32574 newborns shown in Table 5 was used by Graubard & Korn (1987) to illustrate poor properties of midrank scores. The variables are malformations of the newborn's sex organs A, ($i = 1$: yes, $i = 2$: no), and average number of alcoholic drinks the mother had daily during pregnancy, B. The authors note that the distinction between less than one and no drink per day is likely to be unreliable. In addition, with just 165 women reporting to have had more than two alcoholic drinks per day there is in this group essentially no reliable information on the rare response. For three of the contrasts the studentized values, that is the fitted values divided by their standard deviations, are small. Also the likelihood ratio chi-squared statistic of 1.4 on 3 degrees of freedom, shows very good agreement between the observed counts and the reduced model.

Table 5

Counts, percentages as observed and as fitted under the reduced model defined by $A \perp B_j$ given $j = 1, 2$ and $A \perp B_j$ given $j = 3, 4, 5$, together with fitted studentized parameters for level comparisons; for data on malformations in newborns; $n = 32574$

	Levels j of B				
Levels i of A	Mother's number of alcoholic drinks per day				
Newborn's malformation of sex organs	0	< 1	1-2	3-5	≥ 6
Observed counts for i=1 (yes): n_{1j}	48	38	5	1	1
	0.28%	0.26%	0.63%	0.79%	2.37%
Fitted counts for i=1	46.6	39.4	5.8	.9	.3
Observed counts: $n_{1j} + n_{2j}$	17114	14502	793	127	38
	Fitted studentized γ-parameters				
		$\gamma_{12.12}$	$\gamma_{12.23}$	$\gamma_{12.34}$	$\gamma_{12.45}$
Saturated model (with observed counts)		0.314	−1.85	−0.204	−.859
Reduced model (with fitted counts)		0	−2.52	0	0

The studentized value −2.5 of $\hat{\gamma}_{12.23}$ in the reduced model is also the studentized log odds ratio in the 2 × 2 table obtained after combining levels 1 and 2 as well as levels 3, 4 and 5 of variable B. This conclusion captures quantitatively what had been reported by the investigators, namely that at most a binary variable B carries information about the association with A. The increase in the studentized value in the reduced model is analogous to that achieved when eliminating unimportant explanatory variables in multiple regression.

Table 6 is a 4 × 5 contingency table for two ordinal variables obtained for 417 adults in an epidemiological cohort study in Denmark, the so-called Glostrup study. The variables are self-reported health status, A (i =1, very good; i =2, fair; i =3, bad; i =4, very bad) and habits concerning

cigarette smoking five years earlier, B (j =1, never smoked; j =2, did not smoke then; j = 3, 4, and 5 smoked fewer than 10, between 10 and 20 and more than 20 cigarettes per day, respectively). Table 7 shows the dependence after combining levels into a 3 × 2 table. The choice of levels to be merged is made in the light of the studentized estimates of the γ's under the saturated model as shown in Table 6.

Table 6

Counts and studentized level comparison parameters $\hat{\gamma}_{ii'.jj'}$ fitted under three models for a 4 × 5 table of ordinal variables; variable A, self-reported health status, with levels $i = 1, \ldots, 4$: very good, fair, bad, very bad; variable B, smoking habits five years ago, with levels $j = 1, \ldots, 5$: never smoked, quit smoking, smoked less than 10, between 10 and 20, more than 20 cigarettes per day; for data of the Glostrup-study; n = 417

Levels		Fitted counts			Levels		Studentized γ−terms		
i	j	$\mathcal{M}_1$	$\mathcal{M}_2$	$\mathcal{M}_3$	ii'	jj'	$\mathcal{M}_1$	$\mathcal{M}_2$	$\mathcal{M}_3$
1	1	16	15.53	15.53	-	-	-	-	-
2	1	73	73.06	73.06	12	-	9.35	11.49	11.49
3	1	6	6.71	6.42	23	-	−11.39	−12.18	−12.18
4	1	1	0.71	0.99	34	-	−3.66	−4.01	−4.01
1	2	15	15.53	15.53	-	12	−0.64	0.00	0.00
2	2	75	73.06	73.06	-	23	0.51	−1.37	−1.37
3	2	6	6.71	6.42	-	34	1.93	3.27	2.55
4	2	0	0.71	0.98	-	45	−3.78	−6.04	−6.04
1	3	13	12.94	12.94	12	12	0.23	0.00	0.00
2	3	59	60.88	60.88	23	12	−0.05	0.00	0.00
3	3	7	5.58	5.35	34	12	−0.64	0.00	0.00
4	3	1	0.59	0.82	12	23	−0.24	0.00	0.00
1	4	10	8.42	8.42	23	23	0.69	0.00	0.00
2	4	81	84.21	84.21	34	23	0.62	0.00	0.00
3	4	17	15.31	15.92	12	34	1.27	2.11	2.11
4	4	3	3.06	2.45	23	34	1.23	2.00	2.38
1	5	1	2.58	2.58	34	34	0.00	0.70	0.00
2	5	29	25.79	25.79	12	45	1.03	0.00	0.00
3	5	3	4.69	4.88	23	45	−0.95	0.00	0.00
4	5	1	0.94	0.75	34	45	0.70	0.00	0.00

$\mathcal{M}_1$: the saturated model (fitted counts equal the observed counts)

$\mathcal{M}_2$: $A \perp B_j$ given $j = 1, 2, 3$ and $A \perp B_j$ given $j = 4, 5$

$\mathcal{M}_3$: the intersection of model $\mathcal{M}_2$ and $A_i \perp B$ given $i = 3, 4$

Note that the studentized values of the remaining γ's in the 4 × 5 table coincide with those in the reduced model of Table 7.

The likelihood ratio chi-squared statistic for the fit of $\mathcal{M}_2$ against the saturated model $\mathcal{M}_1$,i.e. against the observed counts, has value 3.4 on 9 degrees of freedom, written shorter as $\chi^2_9 = 3.4$. The additional fit of $\mathcal{M}_3$ against $\mathcal{M}_2$ is still excellent with $\chi^2_1 = 1.5$.

The association in the 3 × 2 table remaining after combining levels accordingly is in the expected direction; the risk of reporting poor health is 16.6%, when 10 or more cigarettes had been smoked daily five years ago, about twice as high as 7.7%, when the cigarette consumption was lower. Almost exactly reversed are the chances of reporting good health, with 16% versus 7.6%.

By contrast, a reduced model but no level combination is recommended in the 5 × 5 contingency table shown in Table 8. It captures responses of 268 chronic pain patients to two items, that are questions of a questionnaire in which a score for 'Avoidance of social contacts' is constructed as the sum score of answers to six similar questions; items are on a five point ordinal scale. All bivariate distributions look very similar to the one in Tables 8 and 9.

An interpretation of the fitted model is that persons choosing ordinal level i as response to the

Table 7

Counts and percentages as observed after combining levels of both variables of Table 6, together with studentized level comparison parameters; n = 417

Health status	Number of cigarettes smoked 5 years ago: < 10 per day Counts	Percent	≥ 10 per day Counts	Percent	Studentized γ-terms
good	44	16.2	11	7.6	$\gamma_{12.12}$: 2.11
fair	207	76.1	110	75.9	$\gamma_{23.12}$: 2.38
poor	21	7.7	24	16.6	
Sum	272	100.0	145	100.1	

Table 8

Counts and studentized level comparison parameters as observed on two very similar five-point items of a questionnaire; data for chronic pain patients; n = 268

Observed counts n_{ij}						Studentized γ-terms				
	j						jj'			
i	1	2	3	4	5	ii'	12	23	34	45
1	33	16	4	1	0	12	2.63	0.63	−0.50	−0.23
2	23	33	13	2	0	23	0.72	2.11	0.44	0.08
3	12	24	24	6	1	34	0.69	0.88	2.54	−0.32
4	2	8	13	15	2	45	−0.96	0.07	0.63	2.06
5	3	4	7	12	10					

first question tend to choose the same level $j = i$ or possibly a neighbouring level $j = i - 1$ or $j = i + 1$ as response to the second question. This describes plausible behaviour since there is no reversal in the pooling of the six items of this questionnaire and since the content of the questions is very similar.

The fit of the model is good, since the likelihood ratio chi-squared statistic has value 6.12 on 12 degrees of freedom. It is interesting that choosing linear scores for the items is judged to be inappropriate by fitting models with orthogonal polynomial comparisons: the quadratic by quadratic interaction is needed in addition to the linear by linear one. It might be worthwhile to explore whether score construction can be improved by exploiting these features, possibly by modifying the codes attached to the extreme responses of the items (Cox & Wermuth, 1994a).

3 Contingency Tables of More Than Two Dimensions

3.1 Interactions Based on Level Comparisons

There is a direct formal extension to more than two variables. For a saturated model in which the probabilities are unconstrained other by having to sum to unity, we can write for variables $A, B, C, \ldots$

$$\gamma^{ABC\ldots} = C_{IJK\ldots}\pi^{ABC\ldots}$$

where $C_{IJK\ldots}$ is the contrast matrix defined via level comparisons for the single variables.

For instance, for a 3 × 4 × 2 table C_{IJK} can be computed as

$$C_{342} = C_2 \otimes C_4 \otimes C_3 \ (= C_2 \otimes C_{34} = C_{42} \otimes C_3),$$

and reduced models result by having some of γ terms equal to zero. In general only hierarchical models are of interest, i.e models with no lower order term being set to zero unless all higher order terms involving the same levels of the variables are also zero.

Table 9

Counts and studentized level comparison parameters for the same data as in Table 8 as fitted under a model with $\gamma_{ii',jj'} = 0$ *for all* $\{ii'\} \neq \{jj'\}$; $n = 268$

Fitted counts $\hat{m}_{ij}$						Studentized γ-terms				
	j						jj'			
i	1	2	3	4	5	ii'	12	23	34	45
1	33.00	15.11	4.80	0.97	0.12	12	3.9	0.0	0.0	0.0
2	20.71	36.18	11.50	2.33	0.28	23	0.0	4.4	0.0	0.0
3	11.99	20.94	27.77	5.63	0.67	34	0.0	0.0	4.9	0.0
4	4.34	7.59	10.06	16.09	1.93	45	0.0	0.0	0.0	2.9
5	2.96	5.18	6.87	10.99	10.00					

The log linear γ-parameters will be given again in the lexicographic order mentioned above in Section 2.2 if in the column vector of joint probabilities the levels of the first variable A change fastest and of the last variable slowest.

In general, the simple independence interpretation of zero individual interaction parameters is lost. There are however at least two exceptions.

If the bivariate distribution of two variables, say of A, B is a minimal sufficient component of the model, then all higher order interactions terms involving A, B vanish and $\gamma^{AB}_{ii'.jj'} = 0$ if and only if in the two dimensional AB table the odds ratios are equal to one in the subtables with levels ii' of A and levels jj' of B.

For instance if the joint probabilities factorize as $\pi^{ABC}_{ijk} = \pi^{AB}_{ij}\pi^{BC}_{jk}/\pi^{B}_{j}$, so that $A \perp\!\!\!\perp C \mid B$, then any further pairwise conditional association may also be studied in the marginal bivariate distributions of AB and BC.

This prerequisite is in particular satisfied if the variables follow a so-called quadratic exponential distribution (Cox & Wermuth, 1994b), that is if no higher order than two-factor interactions are present for a given contingency table. This is readily tested with the help of standard log linear models (Bishop *et al.*, 1975; Edwards, 1995) and seems to occur quite frequently in observational studies, see also Section 3.3.

On the other hand there is conditional independence for certain fixed levels of two variables, say ii' and jj' of A, B at all levels of the remaining variables $C, \ldots$, if and only if $\gamma^{AB}_{ii'.jj'} = 0$ and all higher order terms involving levels ii' and jj' are also zero. Thus, the same procedure as in Section 2 can be applied to each of the $K \times L \times \ldots$ two-dimensional AB subtables of the larger $ABCD, \ldots$-table, separately.

In this case it is not the vanishing of individual two factor γ terms, say of A, B, which has a simple interpretation but only their vanishing jointly with all higher order γ terms involving the same levels of A and B. This is particularly useful if a higher order interaction is concentrated only on a small subset of levels whereas conditional independence holds for other subsets of levels for a given pair as is the case in the BCE table of Section 3.4.

3.2 Relations Between Conditional and Joint Distribution Models

For several discrete variables a reduced model may be formulated in at least two ways, as a graphical log linear model (Darroch *et al.*, 1980) for all variables considered jointly or via a system of univariate recursive conditional distributions (Goodman, 1973; Cox & Wermuth, 1996). The two formulations coincide if and only if the log linear model is a so-called decomposable model (Wermuth & Lauritzen, 1983).

A slight extension to log linear interaction models (Edwards, 1995) shows that a conditional distribution of a response variable gives always the same fitted values as the log linear model (for the response and its explanatory variables) if the marginal table of all explanatory variables is one

of the minimal sufficient components. If the response is binary then the result implies that answers obtained by a logistic regression (Cox & Snell, 1989) coincide with those obtained by fitting the corresponding suitable log linear model. An illustration is given here with the examples in Sections 3.3 and 3.4.

3.3 A 2 × 5 × 2 Table

We use some of the data from a large scale study of the US National Institutes of Health on women and their pregnancies (1972) to investigate the dependence of perinatal mortality, A, on the survival status of the last born child, B, and on the skin colour of the mother, C. The variables are:
Perinatal (fetal or neonatal) death, A ($i = 1$: yes, $i = 2$: no); Survival status of last born child, B ($j = 1$: alive, $j = 2$: child death, $j = 3$: fetal death, $j = 4$: neonatal death, $j = 5$: unknown); Skin colour of mother, C ($k = 1$: light, $k = 2$: dark).

A tentative interpretation of the last category, $j = 5$, where the survival status of previous children is not reported, is that it contains a high proportion of women who have given their last child for adoption or who were to ill or to isolated to produce records about their last born child.

Table 10

Counts, studentized level comparison parameters ($\hat{\gamma}_{ii'.jj'.kk'}$), and percentages of a binary reponse, A, ($100 \times \hat{\pi}_{i|jk}$), fitted under two models in a 2 × 5 × 2 table; response variable is perinatal mortality, A; explanatory variables are survival status of last born child, B, and skin colour of mother, C; n = 22574

Levels of $A\,B\,C$	Fitted counts		Fitted percent for level $i = 1$ of response A		Levels	Fitted studentized γ-terms	
$i\ j\ k$	$\mathcal{M}_1$	$\mathcal{M}_2$	$\mathcal{M}_1$	$\mathcal{M}_2$	$ii'\ jj'\ kk'$	$\mathcal{M}_1$	$\mathcal{M}_2$
1 1 1	270	297.12	2.87	3.16	– – –	–	–
2 1 1	9148	9111.33			12 – –	32.04	65.53
1 2 1	3	3.81	2.70	3.16	– 12 –	−23.51	−69.81
2 2 1	108	116.75			– 23 –	16.75	43.46
1 3 1	134	132.82	7.40	7.33	– 34 –	−22.99	−23.91
2 3 1	1678	1679.18			– 45 –	5.63	5.64
1 4 1	17	19.29	8.95	10.15	– – 12	2.78	5.68
2 4 1	173	170.71			12 12 –	0.11	0.00
1 5 1	56	59.33	12.58	13.33	12 23 –	−2.49	−12.13
2 5 1	389	385.67			12 34 –	−1.94	−2.28
1 1 2	371	343.67	3.41	3.16	12 45 –	−1.88	−1.73
2 1 2	10502	10538.88			12 – 12	−0.89	0.00
1 2 2	5	4.40	3.36	3.16	– 12 12	0.46	0.00
2 2 2	144	135.04			– 23 12	−0.67	0.29
1 3 2	154	155.18	7.22	7.33	– 34 12	3.16	4.50
2 3 2	1963	1961.82			– 45 12	−5.04	−7.88
1 4 2	37	34.71	10.82	10.15	12 12 12	−0.06	0.00
2 4 2	305	307.28			12 23 12	−0.32	0.00
1 5 2	46	42.67	14.38	13.33	12 34 12	−0.69	0.00
2 5 2	274	277.33			12 45 12	0.15	0.00

$\mathcal{M}_1$: the saturated model (fitted counts equal the observed counts)

$\mathcal{M}_2$: $A \perp C \mid B$ in intersection with:

$A \perp B_j$ given $j = 1, 2$ and $B_j \perp C$ given $j = 1, 2$

We expected that information on the survival status will change the prediction of perinatal mortality, but that additional information on the skin colour will not lead to a modification of these predictions. Expressed formally we expect a near zero log odds ratio for the binary variables A and C for each

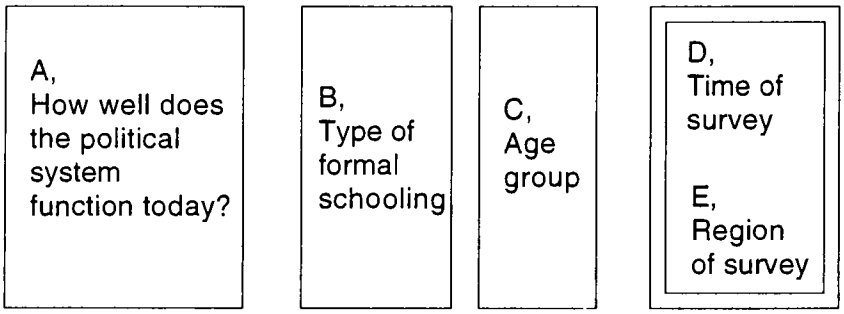

Figure 1. *A first ordering of the variables derived from substance matter knowledge.*

of the five subgroups of women described by the categories of B, i.e. $A \perp\!\!\!\perp C \mid B$.

Provided this holds we examine further, after inspection of the data, consistency with the following two hypotheses. They concern the distinction between the categories 'last born child alive' and 'last born child dead, but not due to birth related causes'. The hypotheses are that this distinction is (i) irrelevant for perinatal death of the just born child and (ii) it is unrelated to skin colour. That is if hypotheses (i) and (ii) hold, there will be a near zero log odds ratio for A and B_j, $j = 1, 2$ at both levels of C, as well as a near zero log odds ratio for B_j $j = 1, 2$ and C at both levels of A.

The risks of perinatal mortality are shown in Table 10 as observed and as fitted for the reduced model which combines the above specified hypotheses. Indeed, risks estimated under the reduced model assumptions differ only little from the observed risks. The chi-squared statistic for the fit of this model against the saturated model confirms the visual impression of a good fit; it has value 7.5 on 7 degrees of freedom. The simplifications achieved supply both a condensed summary of the data via the reduced model specifying $A \perp\!\!\!\perp C \mid B$ and a redefinition of categories by combining levels 1 and 2 of variable B. Both are well justifiable on subject matter grounds.

An interpretation of the evidence in the remaining $2 \times 4 \times 2$ table is as follows. Irrespective of the skin colour of the mother the risk of perinatal mortality increases from a risk 3.2 % for children with a most recent older sibling, who did not die due to birth related causes, to 7.3% and 10.2% if the last born child had died as a foetus or within a week after birth, respectively. The risk increases to 13.3% for $j = 5$, characterizing this clearly as the least favourable category.

The predicted risks for perinatal mortality coincide with probabilities estimated in a logistic regression of the binary variable A on B, C (having 4 and 2 levels) in which only the main effect of B is important. As mentioned before, the reason is that the marginal table of the explanatory variables (BC) is one of minimal sufficient components, i.e. of tables AB and BC, in the log-linear model with $A \perp\!\!\!\perp C \mid B$.

3.4 A $4 \times 5 \times 5 \times 2 \times 2$ *Table*

We return now to the initial example of Section 2 which is taken from a larger context studying the question 'What influences political attitude?'. We analyze here responses from two surveys taken in 1991 and 1992 in Germany. The counts are reproduced in a five dimensional contingency table given in the Appendix with the variable names, shown in Figure 1, abbreviated by A to E. Figure 1 is used to guide analysis by describing some of our knowledge regarding the variables. It is an example of a first ordering of variables for an analysis with graphical Markov models (see Cox & Wermuth, 1996; Wermuth, 1998).

Political attitude (A) is the response variable of primary interest (listed in the first box), for which all other variables are possibly explanatory. There are two variables fixed by design (shown in the doubly lined last box): time (D) and region (E), i.e. West and East Germany, of the survey. For the remaining intermediate variables there is a time order, with type of formal schooling (B) possibly depending on the age group (C) of a respondent, but not vice versa. Each intermediate variable plays the role of both response and explanatory variable, possibly explanatory to some of the variables shown to the left and possibly a response to some of the variables shown to the right.

The levels of the five variables are defined as follows.

Levels of A	Levels of B	Levels of C	Levels of D	Levels of E
$i = 1$, very poorly	$j = 1$, basic incomplete	$k = 1, 19-29$	$l = 1$, 1991	$r = 1$, West
$i = 2$, poorly	$j = 2$, basic	$k = 2, 30-44$	$l = 2$, 1992	$r = 2$, East
$i = 3$, well	$j = 3$, medium	$k = 3, 45-59$		Germany
$i = 4$, very well	$j = 4$, upper medium	$k = 4, 60-74$		
	$j = 5$, intensive	$k = 5, \geq 75$		

There is some further knowledge about the variables involved, which we want to exploit for analysis.

(1) With two separate states having been formed in 1949 different school systems were established in East and in West Germany. Therefore a strong three-factor BCE interaction is expected in the joint distribution of these three variables, at least for persons aged under 60 at the time of the surveys.

(2) The surveys were planned to be representative for the whole population. Hence if this plan had been successful we expect the same type of association between schooling and age for both years within each of the two regions. To state it differently, we expect $BC \perp\!\!\!\perp D$ given each level of E, and, if this independence holds in each region, it implies the additional independencies $B \perp\!\!\!\perp D$ and $C \perp\!\!\!\perp D$ given each level of E.

(3) Within each region we expect only additive but no interactive effects of B, C, D on the response A.

These different expectations are in the following way well supported by the data. The likelihood ratio χ^2-statistic on 16 degrees of freedom for no three-factor interaction in the BCE table is highly significant with a value larger than 200 ($\chi^2_{16} = 266.73$). We therefore proceed to report results for each level of E, separately, which leads to a so-called split model (Høsgard, 1996). The analysis reported in Section 2 for the BC table in West Germany could be replicated for East Germany as shown with observed counts and those estimated under the model with $A_i \perp\!\!\!\perp B \mid i = 1, 2$ and $A \perp\!\!\!\perp B_j \mid j = 4, 5$ in Table 11. We get a good fit to this model in the West ($\chi^2_7 = 7.7$) and in the East ($\chi^2_7 = 6.5$).

Thus, even though there is an extremely strong BCE interaction, simplification can be achieved; the chi-squared statistic $\chi^2_{14} = 14.2 = 7.7 + 6.5$ corresponds to setting all three-factor and two-factor γ interaction terms to zero which involve on the one hand B at levels 1,2 and C at all levels and on the other hand B at all levels and C at levels 4,5. Hence the information on the three-factor interaction is concentrated in the $4 \times 4 \times 2$ BCE table remaining after combining levels 1 and 2 of B and levels 4 and 5 of C as is summarized in Table 12. As a consequence the chi-squared statistic for three-factor interaction the BCE table after concatenating levels is almost unchanged in value in spite of fewer degrees of freedom: $\chi^2_9 = 259.8$.

The main distinguishing features are that in the West the number of persons with intensive schooling increased from 10% for those aged 60 or more to 30% in the age group 18–29, while it only doubled in the East. However, in the East there were more possibilities to continue formal education after having been successful in work so that the observed large difference is slightly misleading. On the other hand for those aged 18–29 the percentage of persons with only basic schooling or less had

Table 11

Observed and fitted counts, East Germany 1991/92; n = 2366

Type of schooling, B	Age group, C 18 - 29	30 - 44	45 - 59	60 - 74	≥ 75	sum
basic, incomplete	5	10	37	18	5	75
	(3.2)	(10.9)	(33.4)	(22.4)	(5.1)	
basic	35	128	384	259	65	871
	(36.8)	(127.1)	(387.6)	(260.6)	(58.9)	
medium	*301*	*503*	*126*	*50*	*7*	*987*
	(301)	(503)	(126)	(46.5)	(10.5)	
upper medium	10	25	39	11	3	88
	(10)	(25)	(39)	(11.4)	(2.6)	
intensive	76	139	92	34	4	345
	(76)	(139)	(92)	(31)	(7)	
Sum	427	805	678	372	84	2366

Table 12

Column percentages of type of formal schooling, B, for the BC counts given in Tables 1 and 11, after merging categories, 1991/92; n = 6039

New levels	West Germany ($r = 1$) New levels k of C, age group				East Germany ($r = 1$) New levels k of C, age group			
j of B	18-29	30-44	45-59	≥ 60	18-29	30-44	45-59	≥ 60
basic or less	29	46	60	63	10	17	62	76
medium	35	26	23	18	70	62	19	12
upper med.	7	8	6	2	2	3	6	3
intensive	30	20	12	10	18	17	14	8
Count	789	1136	846	902	427	805	678	456

been decreased to 10% in the East but only to about 30% in the West.

A reasonable fit to the hypothesis $BC \perp\!\!\!\perp D$ in the $4 \times 4 \times 2$ table is observed for West Germany ($\chi^2_{15} = 21.5$). For East Germany, where the survey was carried out for the first time in 1991, the fit is less good ($\chi^2_{15} = 38.3$). The statistics to test implications of this hypothesis for two-way tables have fewer degrees of freedom and could detect specific deviations between observed and fitted values, hidden in the statistic with many degrees of freedom. However, they point to a good fit in West Germany where values are $\chi^2_3 = 3.8$ for $B \perp\!\!\!\perp D$ and $\chi^2_3 = 3.3$ for $C \perp\!\!\!\perp D$, while the less good fit in the East concerns both schooling ($\chi^2_3 = 13.8$) and age ($\chi^2_3 = 8.5$). This is likely just to reflect the fact that the surveys in the East were not yet completely representative when conducted the first few times.

Within each of the two regions the dependence of A on BCD has only main effects if the marginals tables AB, AC, AD and BCD are sufficient to describe the relations between all four variables. This hypothesis fits well in the West ($\chi^2_{72} = 73.5$) and in the East ($\chi^2_{72} = 83.9$). As a consequence the conditional distribution of A given BCD can be studied in the marginal two-way tables AB, AC, AD. From these we conclude, after proceeding as in Section 2, that there is no difference for our measure of political attitude A for persons with upper medium and with intensive schooling. This means for reporting the type of dependence that the corresponding levels may be combined. Similarly, there is no gain in distinguishing the categories 'very poorly' and 'poorly' or 'very well' and 'well' for the response A.

The three two-way tables show further that disappointment with the political system increased in the years after reunification of Germany from 1991 to 1992 in both regions. In the West it is higher the shorter the formal schooling and the younger the person is, while there is no systematic dependence of this judgement on schooling and age in the East. For West Germany the joint influence of the three

explanatory variables can be summarized with estimated probabilities as follows.

The political system functions poorly or very poorly (in percent)

	B, type of formal schooling					
	medium plus		medium		basic or less	
C, age group	1991	1992	1991	1992	1991	1992
≥ 60	7	17	12	25	16	33
45 – 59	8	20	13	29	18	37
30 – 44	9	22	15	32	20	20
18 – 29	11	24	17	35	22	43

The effect of time (D) is strongest since the risk for a negative judgement increases at all level combinations of age group and schooling by a factor of two or more. There is an increase in risk by not quite a factor of two from 'more than medium level' schooling (B) to 'basic or less'. The weakest but still significant effect is due to age group (C).

The fitted probabilities shown are identical to those estimated with logistic main effect regressions for both regions (E) with response A having just two levels and B, C, D having 3,4 and 2 levels, respectively. In East Germany, there were relatively more negative judgements regarding the political system than in the West, increasing from 35% in 1991 to 53% in 1992, and the remaining two variables B, C had only nonsignificant additional contributions to predicting A.

Some of the similarities and differences in the dependencies are reflected in the recursive regression graphs shown separately for each of the two regions in Figure 2. Further analyses of the 1994 data on the same variables will show whether the trend to more negative attitudes could by then be stopped or even reversed.

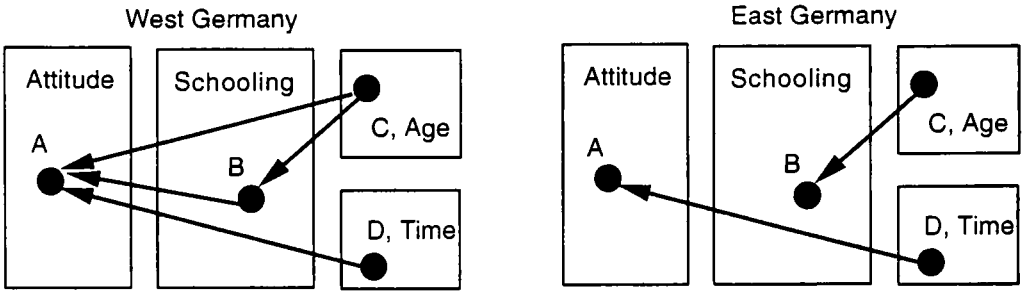

Figure 2. *The two univariate recursive regression graphs resulting from analyses and prior knowledge. They show A depending on B, C, D in West but only on D in East Germany, type of formal schooling depending on age group for both, and BC being independent of D.*

4 Maximum Likelihood Estimates: Derivation and Properties

In this more theoretical section we outline the theory of maximum likelihood fitting for the models used in this paper.

The fitting of a saturated model to multinomial data is by matching fitted counts to observed

counts. We assume for simplicity that no cell frequencies are zero. For example, to fit the model

$$\gamma = C\theta, \quad \theta = D\gamma,$$

where $\theta = \pi^{AB}$ is the vector of log cell probabilities and γ the vector of contrasts, we replace θ by $\hat{\theta}$, the vector of log cell proportions, and compute $\hat{\gamma} = C\hat{\theta}$. The resulting vector of estimated contrasts, supplemented by comparison with their standard deviations, may then be inspected for large and for near zero elements especially starting with the highest order interaction components.

We wish to consider the structure of maximum likelihood estimates when a reduced model is fitted, i.e. one with some of the components of γ set to zero.

Let n^T denote the row vector of observed cell counts, constrained to sum to n, regarded as fixed. We start again from the saturated model in the above form where we arrange that the first column of D has elements all the same. In the log likelihood

$$n^T\theta = n^T D\gamma,$$

for example from the multinomial model of mutually independent trials, the first element of γ, say γ_-, has a coefficient proportional to n which is fixed. The normalizing condition can then be used to replace γ_- by a nonlinear function of the other parameters having fixed coefficient, thus leaving for the saturated model an unconstrained parameter space in the remaining elements of γ.

Now consider a reduced model in which some of the γ_t, other than γ_- are zero. The log likelihood is

$$n^T D_r \gamma_r - k(\gamma_r),$$

where D_r is obtained from D by deleting the first column and columns corresponding to the γ_t that have been set to zero. Thus, D_r is a $T \times (T - 1 - d_r)$ matrix, where d_r is the number of γ_t constrained in the reduced model and $k(\gamma_r)$ is the (fixed) function obtained from the elimination of γ_-. This gives the log likelihood of a $(T - 1 - d_r, T - 1 - d_r)$ full exponential family. Therefore, provided there are enough observations and a technical condition is satisfied, the maximum likelihood estimate of γ exists and is unique. The technical condition (Barndorff-Nielsen, 1978) is that the canonical statistic $n^T D_r$ does not lie on the boundary of its possible values. When, as often is the case, the components of $n^T D_r$ are based on subtotals, this means that the subtotals must not contain zeros; if this happens, one or more components of $\hat{\gamma}$ are formally infinite.

To solve for the maximum likelihood estimates, a possible route is to obtain D, hence D_r, hence the canonical statistics and then to equate them to their expectations.

Some further light on the form of the estimates can be derived by using Lagrange multipliers, w_s. Returning to the representation with the full number T components of γ, each zero γ_s $(s = 1, \ldots, d_r)$ represents a constraint of the form

$$\sum c_{st}\theta_t = 0, \quad \text{or} \quad C_{(s)}\theta = 0,$$

where $C_{(s)}$ is the row of the contrast matrix which corresponds to the position of the zero γ_s in the parameter vector γ. Take therefore the Lagrangian

$$\sum_t n_t \log \pi_t + \sum_s w_s \sum_t c_{st} \log \pi_t - w_0\left(\sum_t \pi_t - 1\right).$$

Formal simplifications are possible whenever each log linear parameter for $t = 2, \ldots, T$ is defined by a contrast in $\log \pi_t$, that is if $\sum_t c_{st} = 0$, $s = 1, \ldots, d_r$. Then each row of the contrast matrix (except for the first) sums to zero. This holds for instance for all contrast matrices C obtained as Kronecker product of C-matrices defined for single variables with level comparisons, with orthogonal polynomials, or with symmetric constraints (which result from an effect coding for the design matrices, $D = C^{-1}$).

After finding a stationary value of the Lagrangian with respect to $\{\pi_t\}$, we have that $w_0 = n$ and

$$\frac{n_t + \sum_s w_s c_{st}}{\hat{\pi}_t} = n.$$

It follows that

$$n\hat{\pi}_t = n_t + \sum_s w_s c_{st}, \quad \text{or} \quad \hat{m}^T = n^T + \sum_s w_s C_{(s)}, \tag{1}$$

where $\hat{m}^T = n\,\hat{\pi}^T$ denotes the fitted vector of counts. Therefore the row vector of differences between fitted and observed counts, $n^T - \hat{m}^T$ is a weighted sum of rows of the comparison matrix, where the weights are Lagrange multipliers.

This implies, in particular, that the maximum likelihood estimate of the cell count matches the observed count, for any cell unaffected by the constraints. These may be identified for level comparisons directly from the subtables defining the constrained log linear parameters. For instance, in a 4×3 table using level contrasts with $\gamma^{AB}_{12.12} = \gamma^{AB}_{12.23} = 0$, the relevant weighted sum of contrasts $\sum_s w_s C_{(s)}$, written as a $I \times J$ table corresponding to the levels of A and B, is

$$w_1 \begin{pmatrix} 1 & -1 & 0 \\ -1 & 1 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{pmatrix} + w_2 \begin{pmatrix} 0 & 1 & -1 \\ 0 & -1 & 1 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{pmatrix} = \begin{pmatrix} w_1 & (-w_1 + w_2) & -w_2 \\ -w_1 & (w_1 - w_2) & w_2 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{pmatrix},$$

so that for this particular reduced model the estimated cell counts in the last two rows match the observed counts and in the first two rows they are the counts expected under independence in this 2×3 table. In general, for interaction contrasts of values 0, 1, or -1, with weights w_s summing to zero over rows i and columns j in the $I \times J$ table of A and B, the interpretation of (1) for a given reduced model is

(i) the estimated cell count matches the observed count, that is $\hat{m}_{t^*} = n\pi_{t^*} = n_{t^*}$ for any cell t^* unaffected by the constraints, that is for which $c_{st^*} = 0$ $(s = 1, \ldots, d_r)$;
(ii) the estimated cell counts match the observed one-dimensional margins of each variable;
(iii) the estimated cell counts satisfy all the constraints.

From (1) results a set of d_r polynomial equations. A sufficient condition for an explicit solution is that these are linear equations. In general an iterative procedure is necessary to find the maximum-likelihood estimates.

There is a connection of these results with the forms of maximum likelihood estimation in other types of exponential model induced from a saturated model by setting some canonical parameters to zero. For example in the multivariate Gaussian covariance selection models of Dempster (1972) the saturated model with arbitrary covariance matrix is constrained by setting some concentrations, that is elements of the inverse covariance matrix, to zero. In the resulting maximum likelihood estimates the elements of the estimated covariance matrix in positions unaffected by the constraints match their observed values, the other estimated elements being such as to satisfy the constraints. From the viewpoint of general theory an interplay is involved between the canonical and moment parameters of the exponential family.

Acknowledgement

We are grateful to the Humboldt Society and the Max Planck Society for supporting our joint work and D.R. Cox acknowledges support by a Leverhulme Emeritus Fellowship. We thank Jochen Hardt, University of Mainz, for the questionnaire data, Svend Kreiner, University of Copenhagen, for the Danish data and for very helpful comments and the Center for Sociological Survey Research and

Methodology (ZUMA Mannheim) for the data on political attitude. The computations were carried out with the help of MATLAB and BMDP. Finally we wish to thank the referees for their extremely constructive criticism.

References

Agresti, A. (1984). *Analysis of ordinal categorical data.* New York: Wiley.

Agresti, A. (1990). *Categorical data analysis.* New York: Wiley.

Barndorff-Nielsen, O.E. (1978). *Information and exponential families in statistical theory.* Chichester: Wiley.

Bishop, Y.M.M., Fienberg, S.E. & Holland, P.W. (1975). *Discrete multivariate analysis: theory and practice.* Cambridge, Mass: MIT Press.

Central Archive for Empirical Social Science Research, (ZA), University of Cologne & Center of Social Survey Research and Methodology, (ZUMA), Mannheim (1993). *ALLBUS 1980–92*, Codebook ZA-number: 1795, Cologne: ZA.

Clogg, C.C. & Shihadeh, E.I. (1994). *Statistical models for ordinal data.* London: Sage.

Cox, D.R. & Snell, E.J. (1989). *Analysis of binary data,* 2nd ed. London: Chapman and Hall.

Cox, D.R. & Wermuth, N. (1990). An approximation to maximum-likelihood estimates in reduced models. *Biometrika*, **77**, 747–761.

Cox, D.R. & Wermuth, N. (1994a). Tests of linearity, multivariate normality and adequacy of linear scores. *Applied Statistics*, **43**, 347–355.

Cox, D.R. & Wermuth, N. (1994b). A note on the binary quadratic exponential. *Biometrika*, **81**, 403–408.

Cox, D.R. & Wermuth, N. (1996). *Multivariate dependencies – models, anylysis and interpretation.* London: Chapman and Hall.

Dale, J.R. (1984). Local versus global association for bivariate ordered response. *Biometrika*, **71**, 507–514.

Darroch, J.N., Lauritzen, S.L. & Speed, T.P. (1980). Markov fields and log-linear models for contingency tables. *Ann. Statist.* **8**, 522–539.

Dempster, A.P. (1972). Covariance selection. *Biometrics*, **28**, 157–175.

Edwards, D. (1995). *Introduction to graphical modelling.* New York: Springer.

Goodman, L. A. (1973). The analysis of multidimensional contingency tables when some variables are posterior to others: a modified path analysis approach. *Biometrika* **60**, 179–192.

Goodman, L.A. (1979a). Simple models for the analysis of association in cross classifications having ordered categories. *J. Amer. Statist. Assoc.*, **74**, 537–552.

Goodman, L.A. (1979b). The analysis of classified data having ordered and/or unordered categories: association models, correlation models and asymmetry models for contingency tables with or without missing entries. *Ann. Statist.*, **13**, 10–69.

Graubard B.I. & Korn, E.L. (1987). Choice of column scores for testing independence in ordered $2 \times K$ tables. *Biometrics*, **43**, 471–476.

Høsgaard, Søren (1996). Split models for contingency tables. Technical Report 2-1996, Biometry Research Unit, Danish Institute of Agricultural Sciences.

Kreiner, S. (1990). Graphical modeling using DIGRAM. In: *Symposium on Applied Statistics*. J. Godt (ed.), Copenhagen: UNI-C.

Lauritzen, S.L. (1996). *Graphical models.* Oxford University Press.

McCullagh, P. & Nelder J.A. (1989). *Generalized linear models.* 2nd ed. London: Chapman and Hall.

Molenberghs, G. & Lesaffre, E. (1994). Marginal modeling of correlated ordinal data using a multivariate Plackett distribution. *J. Amer. Statist. Assoc.*, **89**, 633–644.

National Institute of Health (1972). *The Women and their Pregnancies*, (DHEW Publication No. (NIH) 73–379). Washington: U.S. Government Printing Office.

Speed, T.P. & Kiiveri, H.T. (1986). Gaussian Markov distributions over finite graphs. *Ann. Statist.* **14**, 138–150.

Wermuth, N. (1998). *Graphical Markov models.* In S. Kotz, C. Read & D. Banks (eds) *Encyclopedia of Statistical Sciences*, update volume 2, 284-300. New York: Wiley.

Wermuth, N. & Cox, D.R. (1992). On the relation between interactions obtained with alternative codings of discrete variables. *Methodika*, **VI**, 76–85.

Wermuth, N. & Cox, D.R. (1998a). Statistical dependence and independence. In P. Armitage & T. Colton (eds) *Encyclopedia of Biostatistics*, New York: Wiley. To appear.

Wermuth, N. & Cox, D.R. (1998b). On association models defined over independence graphs. *Bernoulli*, **4**. To appear.

Wermuth, N. & Lauritzen, S. L. (1983). Graphical and recursive models for contingency tables. *Biometrika* **70**, 537–552.

Wermuth, N. & Scheidt, E. (1977). Fitting a covariance selection model to a matrix, Algorithm 105. *Applied Statistics*, **26**, 88–92.

Wilks, S. S. (1938). The large sample distribution of the likelihood ratio for testing composite hypotheses. *Ann. Math. Statist.* **9**, 60–62.

Williams, O.D. & Grizzle J.E. (1972). Analysis of contingency tables having ordered response categories. *J. Amer. Statist. Assoc.*, **67**, 55–63.

Résumé

On décrit une paramétrisation adaptée particulièrement aux des variables ordinales. En particulier celle-ci permettra de guider l'operation comment combiner les niveaux adjacents pour simplifier l'interprétation. Une illustration des possibilités d'interprétation de la méthode à plusieurs ensembles de données est presentée. La méthode de Lagrange expose la structure des estimateurs de maximum vraisemblage.

APPENDIX

Table A1

The 4 × 5 × ×5 × 2 × 2 *table of counts of data on political attitude,* $n = 6039$

Levels of	Level of $E, r = 1$					Level of $E, r = 2$				
A, C, D :	Levels j of B					Levels j of B				
i, k, l	1	2	3	4	5	1	2	3	4	5
1, 1, 1	1	5	10	3	8	0	0	2	0	2
2, 1, 1	3	63	88	22	78	2	13	103	6	29
3, 1, 1	1	25	18	5	9	1	5	53	1	12
4, 1, 1	2	2	0	0	1	0	0	4	0	3
1, 2, 1	0	24	17	3	11	0	3	7	0	1
2, 2, 1	1	135	89	26	68	3	39	198	7	52
3, 2, 1	1	34	14	1	10	4	27	86	7	17
4, 2, 1	0	2	1	0	1	1	7	9	0	1
1, 3, 1	0	26	4	2	5	0	4	0	0	2
2, 3, 1	2	120	62	17	29	14	134	50	5	32
3, 3, 1	1	27	10	2	3	13	61	18	3	18
4, 3, 1	0	6	2	0	0	1	7	3	0	2
1, 4, 1	2	41	12	1	7	0	4	1	0	0
2, 4, 1	6	107	32	4	26	7	81	15	1	12
3, 4, 1	1	18	3	0	2	6	56	8	1	7
4, 4, 1	1	3	0	0	0	1	5	0	0	0
1, 5, 1	1	8	3	1	2	0	3	0	0	0
2, 5, 1	1	28	8	3	6	5	16	1	1	3
3, 5, 1	0	9	2	0	0	0	10	1	0	1
4, 5, 1	1	0	0	0	0	0	1	0	0	0
1, 1, 2	0	6	4	0	7	0	0	0	0	0
2, 1, 2	2	68	101	17	100	1	8	58	1	13
3, 1, 2	3	40	48	3	29	1	8	68	2	16
4, 1, 2	0	6	8	2	1	0	1	13	0	1
1, 2, 2	0	10	7	4	8	0	0	0	0	1
2, 2, 2	4	186	100	47	99	0	28	104	5	26
3, 2, 2	6	102	67	10	25	1	22	86	3	38
4, 2, 2	1	14	5	0	3	1	2	13	3	3
1, 3, 2	1	19	11	2	9	1	2	0	1	0
2, 3, 2	2	182	76	17	42	2	89	25	14	30
3, 3, 2	6	102	24	6	13	3	74	27	14	7
4, 3, 2	0	11	3	1	1	3	13	3	2	1
1, 4, 2	1	11	7	0	4	0	0	0	1	0
2, 4, 2	4	177	57	9	26	1	62	16	3	7
3, 4, 2	5	82	10	1	6	2	46	9	5	6
4, 4, 2	0	21	5	0	3	1	5	1	0	2
1, 5, 2	0	12	2	0	1	0	3	0	0	0
2, 5, 2	3	51	16	1	6	0	18	4	1	0
3, 5, 2	1	22	6	1	4	0	14	1	1	0
4, 5, 2	0	7	1	0	0	0	0	0	0	0

[*Received October 1996, accepted January 1998*]