Biometrika (1990), 77, 4, *pp.* 747-61
Printed in Great Britain

An approximation to maximum likelihood estimates in reduced models

BY D. R. COX
Nuffield College, Oxford OX1 1NF, U.K.
AND NANNY WERMUTH
Psychologisches Institut, Johannes Gutenberg-Universität Mainz, D-6500 Mainz, Federal Republic of Germany

SUMMARY

An approximation to the maximum likelihood estimates of the parameters in a model can be obtained from the corresponding estimates and information matrices in an extended model, i.e. a model with additional parameters. The approximation is close provided that the data are consistent with the first model. Applications are described to log linear models for discrete data, to models for multivariate normal distributions with special covariance matrices and to mixed discrete-continuous models.

Some key words: Asymptotic theory; Concentration matrix; Conditional independence; Covariance matrix; Covariance selection; Generalized least squares; Graphical chain model; Information matrix; Log linear model.

1. INTRODUCTION

On the whole in maximum likelihood fitting, the more parameters the greater the difficulty in computing estimates. There are exceptions however. For instance, the model of direct interest, which we call the reduced model, may be derivable by assigning fixed values to some of the parameters in another model, which we call the extended model, for which estimates are available in simple closed form.

The object of the present paper is to show how maximum likelihood estimates under the reduced model can be found to a close approximation from those in the extended model. For example, this yields simple noniterative procedures for the fitting of models involving a multivariate normal distribution with a covariance matrix for which some elements of the concentration matrix, i.e. inverse covariance matrix, are required to vanish, the covariance selection models of Dempster (1972). The extended model here involves an arbitrary covariance matrix, for which explicit maximum likelihood estimation is via the observed covariance matrix. The resulting approximate maximum likelihood estimates for reduced models can either be used directly, or can form effective initial values for an iterative calculation.

The formulae reduce in a special case to those for the deletion of explanatory variables in least squares estimation of regression coefficients (Cochran, 1938).

2. GENERAL THEORY

We consider an extended model defined by a parameter $\phi = (\theta, \gamma)$, where θ and γ are in general vector parameters and a reduced model obtained by fixing γ at some given

point, denoted by γ_0. The usual regularity conditions for maximum likelihood theory are assumed.

Under the extended model there is a maximum likelihood estimate $\hat{\phi} = (\hat{\theta}_e, \hat{\gamma}_e)$ with information matrix and asymptotic covariance matrix respectively

$$I(\theta, \gamma) = \begin{bmatrix} I_{\theta\theta} & I_{\theta\gamma} \\ I_{\gamma\theta} & I_{\gamma\gamma} \end{bmatrix}, \quad \text{cov}(\hat{\phi}) = \{I(\theta, \gamma)\}^{-1} = \begin{bmatrix} I^{\theta\theta} & I^{\theta\gamma} \\ I^{\gamma\theta} & I^{\gamma\gamma} \end{bmatrix} = \begin{bmatrix} \Sigma_{\theta\theta} & \Sigma_{\theta\gamma} \\ \Sigma_{\gamma\theta} & \Sigma_{\gamma\gamma} \end{bmatrix}. \tag{1}$$

The usual relations analogous to regression adjustments hold; for instance the covariance matrix of $\hat{\theta}$ given a fixed value $\gamma = \gamma_0$ is

$$\Sigma_{\theta\theta.\gamma} = \Sigma_{\theta\theta} - \Sigma_{\theta\gamma}\Sigma_{\gamma\gamma}^{-1}\Sigma_{\gamma\theta} = (I_{\theta\theta})^{-1}. \tag{2}$$

Linear regression coefficients may as usual be computed in two ways, namely as

$$-I_{\theta\theta}^{-1} I_{\theta\gamma} = \Sigma_{\theta\gamma}\Sigma_{\gamma\gamma}^{-1}. \tag{3}$$

One of these may be much simpler than the other, depending on the sizes and forms of the matrices involved.

It would be possible to develop the following argument directly in terms of the log likelihood but, because it is changes in the log likelihood that are important, it is more convenient to work directly with the gradient evaluated at the true parameter point, denoted by U with components U_θ, U_γ.

Then under extended and reduced models respectively, we have, using (1), that

$$I_{\theta\theta}(\hat{\theta}_e - \theta) + I_{\theta\gamma}(\hat{\gamma}_e - \gamma) \doteq U_\theta, \quad I_{\theta\theta}(\hat{\theta}_r - \theta) \doteq U_\theta. \tag{4}$$

When $\gamma = \gamma_0$ we thus get that

$$\hat{\theta}_r \doteq \hat{\theta}_e + I_{\theta\theta}^{-1} I_{\theta\gamma}(\hat{\gamma}_e - \gamma_0).$$

That is, if we write

$$\tilde{\theta}_r = \hat{\theta}_e + I_{\theta\theta}^{-1} I_{\theta\gamma}(\hat{\gamma}_e - \gamma_0) \tag{5}$$

$$= \hat{\theta}_e - \Sigma_{\theta\gamma}\Sigma_{\gamma\gamma}^{-1}(\hat{\gamma}_e - \gamma_0) \tag{6}$$

then $\tilde{\theta}_r$ is close to $\hat{\theta}_r$, given $\gamma = \gamma_0$. Indeed $\tilde{\theta}_r - \hat{\theta}_r = O_p(n^{-1})$, where n is the sample size.

Note that the approximation holds in probability only when $\gamma = \gamma_0$. Thus for data that are inconsistent with $\gamma = \gamma_0$, $\tilde{\theta}_r$ may differ appreciably from $\hat{\theta}_r$ and indeed in some cases $\tilde{\theta}_r$ may lie outside the parameter space, for example, may correspond to a probability outside $[0, 1]$ or to a negative variance.

It follows from (2) that asymptotically $\tilde{\theta}_r$ is in general more precise than $\hat{\theta}_e$, having the covariance matrix associated with the linear regression of $\hat{\theta}_e$ conditional on $\hat{\gamma}_e$. In general $\hat{\theta}_e = \tilde{\theta}_r$ if $\hat{\gamma}_e = \gamma_0$ or if $I_{\theta\gamma} = 0$, that is when the parameters are orthogonal (Cox & Reid, 1987). More generally the adjustment from $\hat{\theta}_e$ to $\tilde{\theta}_r$ will be smaller, and usually more accurate, if near orthogonality can be achieved.

In applying these results, it will often be best to replace the expected information matrix by asymptotic equivalents, in particular the observed information matrix at $(\hat{\theta}_e, \hat{\gamma}_e)$.

Note that when maximum likelihood estimates are also quasilikelihood estimates retaining their essential asymptotic properties under incompletely specified models, our results still apply.

3. A SIMPLE EXAMPLE

Before addressing in §§ 4–6 the types of problem which motivated this work, it is useful to look at a relatively simple example.

Example 1: *Curved exponential family.* Estimation in a curved exponential family provides an illustration of the above results (Efron, 1975). To take a specific example, let $(\bar{Y}_1, \bar{Y}_2)$ be independently normally distributed with mean $(\theta, a\theta^2)$, where a is a known constant and let $\text{var}(\bar{Y}_1) = \text{var}(\bar{Y}_2) = \sigma_0^2/n$, where σ_0^2 is known. The extended model has an arbitrary mean vector, written for the present purpose as $(\theta, a\theta^2 + \gamma)$. Thus $\hat{\theta}_e = \bar{Y}_1$, $\hat{\gamma}_e = \bar{Y}_2 - a\bar{Y}_1^2$. A direct calculation yields at $\gamma = 0$ the information matrix and its inverse

$$I(\theta, \gamma) = \frac{n}{\sigma_0^2}\begin{bmatrix} 1+4a^2\theta^2 & 2a\theta \\ \bullet & 1 \end{bmatrix}, \quad I^{-1}(\theta, \gamma) = \frac{\sigma_0^2}{n}\begin{bmatrix} 1 & -2a\theta \\ \bullet & 1+4a^2\theta^2 \end{bmatrix}.$$

Thus, on replacing θ in I^{-1} by the consistent estimate $\bar{Y}_1 = \hat{\theta}_e$, (5) yields

$$\tilde{\theta}_r = \bar{Y}_1 + \frac{2a\bar{Y}_1}{1+4a^2\bar{Y}_1^2}(-a\bar{Y}_1^2 + \bar{Y}_2). \tag{7}$$

It follows from the above matrices that asymptotically

$$\text{var}(\tilde{\theta}_r) = (\sigma_0^2/n)(1+4a^2\theta^2)^{-1}, \quad \text{var}(\hat{\theta}_e) = \sigma_0^2/n.$$

The structure of $\tilde{\theta}_r$ can be examined directly by writing $\bar{Y}_j = E(\bar{Y}_j) + (\sigma_0^2/n)^{\frac{1}{2}}Z_j$, so that Z_1, Z_2 are independent $N(0, 1)$ random variables. Then

$$\tilde{\theta}_r = \theta + \frac{\sigma_0}{\sqrt{n}}\left(Z_1 + \frac{2a\theta Z_2}{1+4a^2\theta^2}\right) + O_p(n^{-1}). \tag{8}$$

4. APPLICATIONS TO DISCRETE DATA

We now apply the results of § 2 to models for multivariate discrete data in which certain contrasts of log probabilities are zero (Birch, 1963; Goodman, 1970; Cox & Snell, 1989). Many of these models can be interpreted in terms of schemes of conditional independence for some, but by no means all, of which maximum likelihood estimates are available in simple closed form (Haberman, 1974; Darroch, Lauritzen & Speed, 1980).

We take as extended model the saturated model in which all probabilities vary freely. We therefore reparameterize in terms of the parameters of interest in the reduced model and the parameters to be set to zero and evaluate the information matrix for the new parameters.

Let there be p discrete variables with levels $(0, \ldots, I_j - 1)$ $(j = 1, \ldots, p)$ so that there are $L = I_1 \ldots I_p$ levels or cells in the contingency table. We write

$$\pi = (\pi_1, \ldots, \pi_L)^{\mathrm{T}} \quad \nu = (1/\pi_1, \ldots, 1/\pi_L)^{\mathrm{T}}$$

for vectors of probabilities and reciprocals of probabilities. For asymptotic calculations we suppose that none of the π_j is very small. Write, for example, D_π for the diagonal matrix formed from the elements π. For simplicity, we concentrate on the arguments for p binary variables, i.e. take $I_j = 2$ $(j = 1, \ldots, p)$, $L = 2^p$.

We take as extended model an arbitrary multinomial distribution over L cells, so that from n independent observations $\hat{\pi}_e$ is formed from cell proportions and

$$n \operatorname{cov}(\hat{\pi}_e) = D_\pi - \pi\pi^{\mathrm{T}}. \tag{9}$$

The Jacobian of the transformation from π to $\log \pi$ is D_π^{-1}, so that asymptotically

$$n \operatorname{cov}(\log \hat{\pi}_e) = D_\pi^{-1} - ee^{\mathrm{T}}, \tag{10}$$

where e is a vector of ones. Because $D_\pi^{-1}\pi = e$, it follows that

$$(D_\pi - \pi\pi^{\mathrm{T}})(D_\pi^{-1} - ee^{\mathrm{T}}) = I - \pi e^{\mathrm{T}},$$

where I is the identity matrix and, the right-hand side being idempotent, it follows that (9) and (10) are generalized inverses. This is connected with the result that in a full exponential family the canonical parameter and associated moment parameters have mutually inverse information and covariance matrices and with the more general notion of dual parameters (Dempster, 1971).

One important reparameterization is in terms of the factorial contrasts of $\log \pi$ defined by

$$\lambda = (\bar{\lambda}, \lambda_0^A, \lambda_0^B, \lambda_{00}^{AB}, \ldots)^{\mathrm{T}} = T^{-1} \log \pi,$$

where T defines contrasts in a standard factorial system (Yates, 1937; Good, 1958), via

$$T = \begin{bmatrix} 1 & 1 \\ 1 & -1 \end{bmatrix} \otimes \begin{bmatrix} 1 & 1 \\ 1 & -1 \end{bmatrix} \otimes \ldots \otimes \begin{bmatrix} 1 & 1 \\ 1 & -1 \end{bmatrix},$$

with $T^{-1} = 2^{-p}T$.

The asymptotic covariance matrix of $\hat{\lambda}_e$ is

$$T^{-1} \operatorname{cov}(\log \hat{\pi}_e) T^{-1} = L^{-2}\{T^{(i)\mathrm{T}} D_\nu T^{(j)}\} + (T^{-1}e)(T^{-1}e)^{\mathrm{T}},$$

where $T^{-1}e = (1\,0 \ldots 0)^{\mathrm{T}}$, and $T^{(i)}$ is a column of T. Thus

$$n \operatorname{cov}(\hat{\lambda}) = L^{-2}\left(\begin{array}{c|ccc} \sum \pi_j^{-1} - L^2 & \nu^{\mathrm{T}}T^{(2)} & \ldots & \nu^{\mathrm{T}}T^{(p)} \\ \hline & & & \\ & T^{(i)\mathrm{T}} D_\nu T^{(j)} & & \end{array}\right), \tag{11}$$

where $T^{(i)\mathrm{T}} D_\nu T^{(j)} = \Sigma\, \pi_l^{-1}$ for $i = j$ and all remaining products are elements of $\{\nu^{\mathrm{T}}T^{(2)}, \ldots, \nu^{\mathrm{T}}T^{(p)}\}$, since elementwise products of columns of the design matrix just reproduce one of the other columns in the design matrix.

To fit a reduced model in which certain of the λ's are zero, we pick out the rows and columns of (11) so that an upper block refers to the parameters θ of ultimate interest and the lower block to the parameters γ fixed at 0 under the reduced model. Application of the results of § 2 is then immediate, incidentally without matrix inversion.

This procedure has been tested on a number of empirical examples all confirming that $\tilde{\theta}_r$ and $\hat{\theta}_r$ agree to a close approximation, provided that the data are consistent with the reduced model, $\gamma = \gamma_0$, as judged from the studentized interactions.

Example 2: 2×2 *table with near independence.* It is worth illustrating these results on a simple example where direct calculation is possible. Consider a 2×2 contingency table defined by two binary random variables A and B, with $\pi_{ij} = \operatorname{pr}(A = i, B = j)$, n_{ij} being the corresponding cell frequency, $\Sigma\, n_{ij} = n$. Suppose that the π_{ij} are arbitrary in the extended model and that the reduced model corresponds to independence.

We start from

$$\pi_{00} = \exp(\xi+\mu+\zeta)/\Gamma, \quad \pi_{10} = \exp(-\xi+\mu-\zeta)/\Gamma,$$
$$\pi_{01} = \exp(\xi-\mu-\zeta)/\Gamma, \quad \pi_{11} = \exp(-\xi-\mu+\zeta)/\Gamma,$$

where

$$\Gamma = \{\exp(\xi+\mu+\zeta)+\exp(-\xi+\mu-\zeta)+\exp(\xi-\mu-\zeta)+\exp(-\xi-\mu+\zeta)\}$$

is a normalizing constant playing the role of an overall mean.

In the notation of § 2, $\theta = (\xi, \mu)$, $\gamma = \zeta$, $\phi = (\xi, \mu, \zeta)$. In a notation consistent with non-binary data, we sometimes write, for example, λ_i^A, the main effect of A at level i, etc.

Table 1 is a numerical example, illustrating close agreement between $\tilde{\theta}_r$ and $\hat{\theta}_r$, arising in part because the sample two-factor interaction is small.

Example 3: 2^3 *table with no three-factor interaction.* We now illustrate how the procedure of § 2 provides a noniterative accurate approximation in the fitting of a model for which no explicit maximum likelihood solution is available (Bartlett, 1935). Table 2 gives counts,

Table 1. *Numerical example of* 2×2 *table: estimates and covariance matrices*

(*a*) *Estimates*

	Level of A	Level of B			Type of			
l	i	j	n_{ij}	$\hat{\pi}_{ij,c}$	interaction	$\hat{\theta}_e$	$\tilde{\theta}_r$	$\hat{\theta}_r$
1	0	0	10	0·09	$\bar{\lambda}$	−1·548	−1·540	−1·540
2	1	0	30	0·27	$\lambda_0^A = \xi$	−0·504	−0·490	−0·490
3	0	1	20	0·18	$\lambda_0^B = \mu$	−0·301	−0·279	−0·279
4	1	1	50	0·46	$\lambda_{00}^{AB} = \zeta$	−0·046	0	0

(*b*) *Covariance matrix*

0·0036	0·0060	0·0040	0·0023
.	0·0127	0·0023	0·0040
.	.	0·0127	0·0060
.	.	.	0·0127

Covariance matrix is $\frac{1}{4}T(D_\pi^{-1} - ee^{\mathrm{T}})T$, with π estimated by $\hat{\pi}_c$.
Studentized interaction is, for example, for $\lambda_{00}^{AB} = -0{\cdot}046/\sqrt{0{\cdot}0127} = -0{\cdot}40$.
Chi-squared goodness of fit, 0·17 (1 d.f.) based on $\tilde{\theta}_r$ or $\hat{\theta}_r$.

Table 2. *Symptoms of gestosis*

Level of A	Level of B	Level of C		Type of interaction	Studentized			
i	j	k	n_{ijk}	at levels 0	interaction	$\hat{\theta}_e$	$\tilde{\theta}_r$	$\hat{\theta}_r$
0	0	0	2342	$\bar{\lambda}$	—	−3·8673	−3·8712	−3·8713
1	0	0	609	λ^A	3·04	0·2068	0·2020	0·2021
0	1	0	44	λ^B	17·80	1·2103	1·2136	1·2138
1	1	0	6	λ^{AB}	2·73	0·1857	0·1969	0·1964
0	0	1	45	λ^C	16·52	1·1233	1·1253	1·1255
1	0	1	36	λ^{AC}	3·85	0·2615	0·2741	0·2736
0	1	1	9	λ^{BC}	8·41	0·5718	0·5707	0·5705
1	1	1	14	λ^{ABC}	0·29	0·0195	0·0000	0·0000

Chi-squared goodness of fit (1 df) based on $\tilde{\theta}_r$ or $\hat{\theta}_r$, 0·075.

studentized interactions and fitted parameters. The data (Wermuth & Koller, 1976) concern symptoms of gestosis, an illness with still unknown aetiology, occurring during pregnancy. The symptoms are oedema, present $i=1$; proteinuria, present $j=1$; hypertension, present $k=1$.

Maximum likelihood fitting of the reduced model requires iteration. The procedures of § 2 provide a noniterative accurate approximation.

5. APPLICATION TO SOME MULTIVARIATE NORMAL MODELS

In § 4 we dealt with a class of problems connected with discrete random variables. We now turn to two rather different types of reduced model connected with the multivariate normal distribution.

In the first, the extended model is that for a random sample from an arbitrary multivariate normal distribution, whereas under the reduced model certain elements of the concentration matrix, i.e. the inverse of the covariance matrix, vanish (Dempster, 1972). The zero elements correspond to assumptions of conditional independence.

In the second, the extended model is that for the so-called general linear model in which a vector of response variables has linear regression on a fixed set of explanatory variables with multivariate normal errors and a regression of the same form for each component response but with functionally independent parameters, i.e. regression coefficients. In the reduced model some of these regression coefficients are zero, so that the regressions of the separate components are allowed to have different structures (Haavelmo, 1943).

In both cases the maximum likelihood estimates under the extended model take a simple form. In the first case these are determined by the sample mean and covariance matrix, or its inverse, and in the second they are the separate least squares estimates and the covariance matrix of residuals.

Suppose, first, then that $Y=(Y^{(1)},\ldots,Y^{(q)})$ has the multivariate normal distribution, $\text{MN}_q(\mu,\Sigma)$ with mean μ and covariance matrix Σ and that a random sample Y_t $(t=1,\ldots,n)$ is available.

It is convenient to list the distinct elements of Σ as a vector $\sigma=(\sigma_{11},\sigma_{12},\sigma_{13},\ldots,\sigma_{qq})$. Under the extended model

$$\hat{\mu}=(\bar{Y}^{(1)},\ldots,\bar{Y}^{(q)}),\quad \hat{\sigma}=(S_{11},S_{12},\ldots,S_{qq}),$$

where $\hat{\mu}$ is the vector of sample means and

$$S_{ij}=n^{-1}\sum(Y_t^{(i)}-\bar{Y}^{(i)})(Y_t^{(j)}-\bar{Y}^{(j)}).$$

Now $\hat{\mu}$ and S_{ij} are independently distributed with $\hat{\mu}\sim N(\mu,n^{-1}\Sigma)$ and asymptotically $\hat{\sigma}\sim\text{MN}\{\sigma,n^{-1}\,\text{Iss}\,(\Sigma)\}$, where $\text{Iss}\,(\Sigma)$ is the matrix with elements (Isserlis, 1918)

$$n\,\text{cov}\,(\hat{\sigma}_{ij},\hat{\sigma}_{kl})=\sigma_{ik}\sigma_{jl}+\sigma_{il}\sigma_{jk}. \tag{12}$$

Here, however, we need the information or covariance matrix for the estimate not of σ but of ω, the vector formed from the elements of the concentration matrix $\Omega=\Sigma^{-1}$. This is achieved via the simple asymptotic result that $\hat{\omega}\sim N\{\omega,\text{Iss}\,(\Omega)\}$.

To see this suppose first that μ is known, equal to zero without loss of generality. Then $Z=Y\Omega\sim\text{MN}\,(O,\Omega)$ so that the vector of average sums of squares and products about the origin formed from Z_t $(t=1,\ldots,n)$ is asymptotically $\text{MN}\,\{\omega,n^{-1}\,\text{Iss}\,(\Omega)\}$. Replacement of Ω by $\hat{\Omega}$ and the mean squares and products about the origin by those

about the sample mean leaves the asymptotic result unchanged. Note finally that the matrix of average sums of squares and products about the origin is $n^{-1}\hat{\Omega}^{\mathrm{T}}Y^{\mathrm{T}}Y\hat{\Omega} = n^{-1}\hat{\Omega}$.

To use these results, we need as before to define the θ's and γ's appropriately and to pick out the rows and columns of Iss (Ω) relevant for θ. Because of orthogonality the estimation of μ can be disregarded. Similar arguments would not hold if the reduced model were to be specified by zero marginal correlations, i.e. by zero elements of Σ.

In the extended model the canonical parameters are $\mu^{\mathrm{T}}\Omega$ and Ω. We have, however, worked with the mixed orthogonal parameterization (μ, Ω).

Example 4: *Smallest known non-decomposable covariance matrix.* Table 3 gives means and sample covariance matrix for four variables measured on 684 female students (Spielberger, 1983; Spielberger et al., 1983). The variables are X, anxiety state; Y, anger state; Z, anxiety trait; U, anger trait. Trait variables are viewed as stable personality characteristics of a person and the state variables as denoting behaviour in specific situations. A model suggested by psychological theory has $X \perp\!\!\!\perp U \mid (Y, Z)$ and $Y \perp\!\!\!\perp Z \mid (X, U)$. The fit of the model is good as shown by the studentized interactions in Table 4, and $\tilde{\theta}_r$ and $\hat{\theta}_r$ agree closely and in this instance differ little from $\hat{\theta}_e$.

To study the second regression problem, let Y be an $n \times q$ matrix of response variables with independent rows, each multivariate normal with covariance Σ and with $E(Y) = x\beta$, where x is an $n \times p$ matrix of known explanatory variables and β is a $p \times q$ matrix of unknown parameters. Then under the extended model, $\hat{\beta} = (x^{\mathrm{T}}x)^{-1}x^{\mathrm{T}}Y$, corresponding to applying ordinary least squares to each component variable in turn. Further the

Table 3. 684 *female students*: *Means and covariance matrix of* 4 *values*

(*a*) *Covariance matrix*

X	37·1926	24·9311	21·6056	15·6907
Y	.	44·8472	17·8072	21·8565
Z	.	.	32·2462	18·3523
U	.	.	.	43·1191

(*b*) *Means*

18·8744	15·2265	21·2019	23·4217
X	Y	Z	U

Table 4. 684 *female students*: *Estimates of studentized concentrations and concentrations and precisions under extended and reduced models*

Type of parameter	Studentized concentration	$\hat{\theta}_e$	$\tilde{\theta}_r$	$\hat{\theta}_r$
ω_{xx}	—	0·0560	0·0567	0·0568
ω_{xy}	10·70	−0·0213	−0·0214	−0·0213
ω_{yy}	—	0·0404	0·0397	0·0397
ω_{xz}	−11·12	−0·0267	−0·0267	−0·0267
ω_{yz}	−0·66	0·0012	0·0000	−0·0001
ω_{zz}	—	0·0576	0·0567	0·0567
ω_{xu}	0·99	0·0017	0·0000	−0·0002
ω_{yu}	−7·95	−0·0019	−0·0116	−0·0115
ω_{zu}	−7·93	−0·0142	−0·0137	−0·0136
ω_{uu}	—	0·0346	0·0348	0·0348

Chi-squared goodness of fit (2 d.f.), 1·95 for $\tilde{\theta}_r$, 2·10 for $\hat{\theta}_r$.

maximum likelihood estimate of Σ is formed from the matrix of residual sums of squares and products and is independent of $\hat{\beta}$, and so can be disregarded in the following calculations. The covariance matrix of the vector formed by 'stacking' the columns of $\hat{\beta}$ is $\hat{\Sigma}\otimes(x^{\mathrm{T}}x)^{-1}$. Again to apply the results of § 2 to a model in which some of the components of β are zero, we pick out the relevant rows and columns.

Example 5: *Simplest linear regression problem requiring iterative estimation.* Suppose that there are two component variables, $q=2$, and that $Y^{(1)}$ is assumed to have regression on the first explanatory variable, $Y^{(2)}$ on the second. For simplicity suppose that both regressions are through the origin, and that we have n independent pairs. The extended model is

$$E(Y)=\begin{bmatrix} x_{11} & x_{12} \\ \vdots & \vdots \\ x_{n1} & x_{n2} \end{bmatrix}\begin{bmatrix} \theta_1 & \gamma_2 \\ \gamma_1 & \theta_2 \end{bmatrix},$$

with unknown covariance matrix for the pairs $(Y^{(1)}, Y^{(2)})$. The reduced model is obtained from the special case $\gamma_1=\gamma_2=0$. The 4×4 information matrix of $(\hat{\theta}_{1e}, \hat{\theta}_{2e}, \hat{\gamma}_{1e}, \hat{\gamma}_{2e})$ is

$$\begin{bmatrix} \omega_{11}\Sigma x_{t1}^2 & \omega_{12}\Sigma x_{t1}x_{t2} & \omega_{11}\Sigma x_{t1}x_{t2} & \omega_{12}\Sigma x_{t1}^2 \\ \bullet & \omega_{22}\Sigma x_{t2}^2 & \omega_{12}\Sigma x_{t2}^2 & \omega_{22}\Sigma x_{t1}x_{t2} \\ \bullet & \bullet & \omega_{11}\Sigma x_{t2}^2 & \omega_{12}\Sigma x_{t1}x_{t2} \\ \bullet & \bullet & \bullet & \omega_{22}\Sigma x_{t1}^2 \end{bmatrix}.$$

Thus by (5)

$$\tilde{\theta}_r=\hat{\theta}_e+\begin{bmatrix} \hat{\omega}_{11}\Sigma x_{t1}^2 & \hat{\omega}_{12}\Sigma x_{t1}x_{t2} \\ \bullet & \hat{\omega}_{22}\Sigma x_{t2}^2 \end{bmatrix}^{-1}\begin{bmatrix} \hat{\omega}_{11}\Sigma x_{t1}x_{t2} & \hat{\omega}_{12}\Sigma x_{t1}^2 \\ \hat{\omega}_{12}\Sigma x_{t2}^2 & \hat{\omega}_{22}\Sigma x_{t1}x_{t2} \end{bmatrix}\begin{bmatrix} \hat{\gamma}_{1e} \\ \hat{\gamma}_{2e} \end{bmatrix}.$$

Kenny (1979, p. 236) reported for 724 U.S. cities correlations between X, no. of police per cap, 1969; Y, no. of reported burglaries per cap, 1969; Z, no. of police per cap, 1968; U, no. of reported burglaries per cap, 1968. See Table 5. It is sensible to consider the hypothesis that each 1969 value is predicted from the corresponding 1968 value but that the information on the other variable in 1968 does not improve prediction, that is $X \perp\!\!\!\perp U \mid Z$ and $Y \perp\!\!\!\perp Z \mid U$.

Table 5. 724 *U.S. cities. Correlations between four variables*

X	1	0·39	0·86	0·43
Y	.	1	0·35	0·89
Z	.	.	1	0·47
U	.	.	.	1

Table 6 shows studentized regression coefficients under the extended and reduced models. Although there is evidence of lack of fit, $\tilde{\theta}_r$ and $\hat{\theta}_r$ agree closely.

One distinction between the two problems analysed in this section is that in the second conditioning takes place on a pair of variables, whereas in the first problem all variables are treated as random, in fact as multivariate normal. Provided that interest is focused on the appropriate regression coefficients and not on other aspects of the system under study, the arguments for conditioning are powerful and include the substantial extra generality achieved by evading assumptions on the distributional form of the explanatory variables.

Table 6. 724 *U.S. cities. Estimates of studentized regression coefficients under the extended and reduced models*

Type of parameter	Studentized regression coefficients	$\hat{\theta}_e$	$\tilde{\theta}_r$	$\hat{\theta}_r$
$\theta_1 = \beta_{xz.w}$	39·37	0·844	0·881	0·882
$\gamma_1 = \beta_{yw.z}$	1·54	0·033	0	0
$\gamma_2 = \beta_{yz.w}$	−4·63	−0·088	0	0
$\theta_2 = \beta_{yw.z}$	49·22	0·931	0·886	0·886
$\theta_3 = \omega_{11}$		0·260	0·261	0·261
$\theta_4 = \omega_{12}$		0·065	0·068	0·068
$\theta_5 = \omega_{22}$		0·202	0·208	0·208

Chi-squared goodness of fit (2 d.f.), 22·013 for $\tilde{\theta}_r$, 22·015 for $\hat{\theta}_r$.

Within the framework of maximum likelihood estimates a conditional formulation can be converted into an unconditional one, indeed in many ways. For if the conditional log likelihood is $l(\theta; x)$, where x is a vector of 'fixed' explanatory variables and $g(x; \delta)$ is the density of x as a random variable, where the parameter spaces of θ and δ are separate, the overall log likelihood is $l(\theta; x) + \log g(x; \delta)$, and entirely separate maximum likelihood estimation of δ and θ follows. Note that it is crucial that all information relevant to θ is contained in the first term. Given first a joint distribution of all variables, including x, this concentration of information in the conditional distribution will hold, if at all, only for a particular choice of parameters.

Given the conditional log likelihood $l(\theta; x)$, the simplest choice of $g(x; \delta)$ giving an easily handled joint distribution is likely to be the exponential family distribution generated by the pure functions of x entering $l(\theta; x)$ and having moment parameter δ. Thus for normal theory simple linear regression the conditional log likelihood involves, as well as the sufficient statistics $\Sigma y_t, \Sigma y_t^2, \Sigma x_t y_t$, the functions $\Sigma x_t, \Sigma x_t^2$. It is therefore sufficient to give the x's the exponential family distribution generated by $\Sigma x_t, \Sigma x_t^2$, equivalent to a normal distribution for the x_t. The maximum likelihood estimates of Σx_t, Σx_t^2 are those of the marginal parameters and maximum likelihood analysis of the bivariate normal model with all parameters unknown is, for the regression coefficient, the same as maximum likelihood analysis of the conditional model.

By an extension of this argument, the second example above could have been treated via the maximum likelihood analysis of a random sample from a 4-variate normal distribution in which X_1, X_2 are assigned unknown means and marginal covariance matrix. Conceptually the conditional approach is usually preferable, although there may sometimes be computational advantages in the unconditional analysis. The connection between logistic regression and log linear models is another illustration of these points, where the assigned distribution for x is multinomial on the observed points, a useful notion only when the number of such points is small.

6. Application to mixed discrete-continuous models

Finally we consider applications in which a mixture of discrete and continuous variables is involved, in particular via graphical association models (Lauritzen & Wermuth, 1989). Let there be p discrete variables with $I_1, \ldots, I_p$ levels, so that $L = I_1 \ldots I_p$ is the total number of level combinations or cells, the corresponding vector of probabilities being

$\pi^{\mathrm{T}} = (\pi_1, \ldots, \pi_L)$. Suppose also that there are q continuous variables $Y^{(1)}, \ldots, Y^{(q)}$ and in the lth cell let $\mu_i(l) = E(Y^{(i)})$, $\sigma_{ij}(l) = \mathrm{cov}(Y^{(i)}, Y^{(j)})$, $\Sigma(l)$ and $\Omega(l) = \{\Sigma(l)\}^{-1}$ being the corresponding covariance and concentration matrices. For a CG-distribution the joint distribution within a cell is multivariate normal and in the important homogeneous case $\Sigma(l) = \Sigma$. We write, as in § 5, $\mu^{\mathrm{T}}(l)$ and $\sigma^{\mathrm{T}}(l)$ for column vectors of parameters.

Now the parameterization in terms of expected counts $n\pi^{\mathrm{T}}$, means $\mu^{\mathrm{T}} = (\mu^{\mathrm{T}}(1), \ldots, \mu^{\mathrm{T}}(L))$ and covariances $\sigma^{\mathrm{T}} = (\sigma^{\mathrm{T}}(1), \ldots, \sigma^{\mathrm{T}}(L))$, is orthogonal with covariance matrix formed from three blocks of the respective components. Replacement of σ^{T} by ω^{T}, the corresponding concentration matrix preserves the block diagonal structure. In the homogeneous case the separate covariance or concentration matrices for each cell are replaced by a single matrix.

If models involving conditional independencies among the discrete components considered marginally are of interest we follow the procedures of § 4. If, with the CG-family, conditional independencies involving both discrete and continuous components are considered (Wermuth & Lauritzen, 1990), we must transform to the canonical parameters $\Omega = \Sigma^{-1}$ and $\zeta^{\mathrm{T}} = \mu^{\mathrm{T}}\Omega$.

We transform first to the canonical parameters $\Omega = \Sigma^{-1}$, $\zeta^{\mathrm{T}} = \mu^{\mathrm{T}}\Omega$ and δ with

$$\delta_l = \log \pi_l - \tfrac{1}{2}[\mu^{\mathrm{T}}(l)\mu(l) + \log\{2\pi|\Sigma(l)|\}]$$

and in a further step to canonical interaction parameters. For instance, for two discrete random variables, A, B and two continuous variables X, Y, the quadratic ψ, the linear η, and discrete canonical interactions, λ, in a nonhomogeneous CG-distribution are

$$\psi^X = M[\omega_{xx}(l)], \quad \psi^{XY} = M[\omega_{xy}(l)], \quad \psi^Y = M[\omega_{yy}(l)],$$

$$\eta^X = M[\zeta_x(l)], \quad \eta^Y = M[\zeta_y(l)], \quad \lambda = M[\delta],$$

where, for example, $[\omega_{xx}(l)]$ denotes the $L \times 1$ vector of precisions for X and M is the Kronecker product of inverses of design matrices for the discrete variables.

Thus, if $I_1 = 2$, $I_2 = 3$ we may take

$$M = \begin{bmatrix} 1 & 1 \\ 1 & -1 \end{bmatrix}^{-1} \otimes \begin{bmatrix} 1 & 1 & 1 \\ 1 & 0 & -2 \\ 1 & -1 & 1 \end{bmatrix}^{-1}.$$

The conditional independence hypothesis $A \perp\!\!\!\perp B \,|\, (X, Y)$ is in this case reflected in zero two-factor interactions in all six interaction vectors, i.e. in the last two terms of $\psi^X, \psi^{XY}, \ldots, \lambda$ being zero.

The covariance matrices of estimates of the canonical parameters can be computed via the block-diagonal structure of the orthogonal estimates $\hat{\pi}, \hat{\mu}, \hat{\sigma}$, using the appropriate transformation matrices. No explicit matrix inversion is needed.

The concentration matrix of the moment estimates is

$$\{n\,\mathrm{cov}(\hat{\pi}, \hat{\mu}, \hat{\sigma})\}^{-1} = \mathrm{diag}\{D_\nu - ee^{\mathrm{T}}, D_{[\pi_l \Omega(l)]}, D_{[\pi_l D_l \,\mathrm{lss}(\omega) D_l]}\},$$

where the second and third components are themselves block-diagonal square matrices of sizes qL and $q'L$ respectively, where $q' = \frac{1}{2}q(q+1)$, and D_f is a diagonal matrix with elements $\frac{1}{2}$ for precisions, ω_{ii} and 1 for concentrations, ω_{ij} $(i \neq j)$.

A transformation from $(\pi^{\mathrm{T}}, [\mu(l)]^{\mathrm{T}}, [\sigma(l)]^{\mathrm{T}})$ to $(\pi^{\mathrm{T}}, [\pi_l\mu(l)]^{\mathrm{T}}, [\pi_l\sigma(l)]^{\mathrm{T}})$ and back involves the transformation matrices

$$T_1 = \begin{bmatrix} I_L & 0 & 0 \\ D_{[\mu(l)]} & D_\pi \otimes I_q & 0 \\ D_{[\sigma(l)]} & 0 & D_\pi \otimes I_{q'} \end{bmatrix}, \quad T_1^{-1} = \begin{bmatrix} I_L & 0 & 0 \\ -D_{[\nu_l\mu(l)]} & D_\nu \otimes I_q & 0 \\ -D_{[\nu_l\sigma(l)]} & 0 & D_\nu \otimes I_{q'} \end{bmatrix}.$$

Further a transformation from $(\pi^{\mathrm{T}}, [\pi_l\mu(l)]^{\mathrm{T}}, [\pi_l\sigma(l)]^{\mathrm{T}})$ to

$$(\pi^{\mathrm{T}}, [\pi_l\mu(l)]^{\mathrm{T}}, [\pi_l\{\sigma(l) + a(l)\}]^{\mathrm{T}})$$

with $a_{ij}(l) = \mu_i(l)\mu_j(l)$ and back involves T_2, T_2^{-1}, where

$$T_2 = \begin{bmatrix} I_L & 0 & 0 \\ 0 & I_{qL} & 0 \\ -D_{[a(l)]} & D_{[B(l)]} & I_{q'L} \end{bmatrix}, \quad T_2^{-1} = \begin{bmatrix} I_L & 0 & \\ 0 & I_{qL} & 0 \\ D_{[a(l)]} & -D_{[B(l)]} & I_{q'L} \end{bmatrix},$$

where $B(l)$ is the matrix of derivatives of $a(l)$ with respect to μ_l. For instance for $q = 3$ variables X, Y, Z we have, suppressing subscripts, that

$$B^{\mathrm{T}} = \begin{bmatrix} 2\mu_x & \mu_y & \mu_z & 0 & 0 & 0 \\ 0 & u_x & 0 & 2\mu_y & \mu_z & 0 \\ 0 & 0 & \mu_x & 0 & \mu_y & 2\mu_z \end{bmatrix}.$$

The covariance matrix of counts, sums, sums of squares and products is

$$n^2 T_2 T_1 \operatorname{cov}(\hat{\pi}, \hat{\mu}, \hat{\sigma}) T_1^{\mathrm{T}} T_2^{\mathrm{T}},$$

the inverse being the covariance matrix of $(\hat{\delta}, \hat{\zeta}, -D_f\hat{\omega})$, a fact exploited by Dempster (1973) in deriving the covariance matrix for estimates of canonical parameters in a homogeneous CG-distribution. His results are reproduced by replacing $D_{[\sigma(l)]}$ in the last line of T_1 by a matrix of zeros, and replacing $D_\pi \otimes I_q$ by I_q.

The asymptotic covariance matrix of the canonical parameter is, with minus denoting a generalized inverse,

$$\operatorname{cov}(\hat{\delta}, \hat{\zeta}, \hat{\omega}) = n^{-1} T_3^{-\mathrm{T}} T_2^{-\mathrm{T}} T_1^{-\mathrm{T}} \operatorname{cov}(\hat{\pi}, \hat{\mu}, \hat{\sigma})^{-} T_1^{-1} T_2^{-1} T_3^{-1},$$

where $T_3 = \operatorname{diag}(I_L, I_{qL}, -D_f \otimes I_{q'L})$, and with explicit expressions for the other inverses being given above.

Finally, the variances of canonical interactions are needed to compute studentized interactions as a basis for judging the adequacy of fit of the reduced model.

The computations are considerably simpler if conditional independencies among the discrete variables are not of interest. From the block-diagonal covariance matrix of $(\hat{\pi}, \hat{\mu}, \hat{\omega})$ that of $(\hat{\pi}, \hat{\zeta}, \hat{\omega})$ can be obtained via the transformation matrix

$$\begin{bmatrix} I_L & 0 & 0 \\ 0 & D_{[\Omega(l)]} & D_{[B(l)^{\mathrm{T}} D f]} \\ 0 & 0 & I_{q'L} \end{bmatrix}.$$

Again suppressing subscripts, it can be shown that in a cell with count n^*

$$n^* \operatorname{cov}(\hat{\zeta}, \hat{\omega}) = \begin{bmatrix} \Omega(I + CD_f B^{\mathrm{T}}) & \Omega C \\ \bullet & \operatorname{Iss}(\omega) \end{bmatrix},$$

where C has the same dimension and form as B with means replaced by corresponding linear canonical parameters; the parameters ζ, ω are orthogonal to π.

Example 6: *Numerical example of conditional independence model for two discrete and one continuous variable.* Table 7 gives counts, means and variance for two binary and one continuous variable. It is the simplest mixed association model requiring iterative fitting for maximum likelihood estimation (Edwards, 1990; Frydenberg & Edwards, 1989). The reduced model has $A \perp\!\!\!\perp B \mid Y$ and in this two two-factor interactions involving A and B are zero (Lauritzen & Wermuth, 1989), namely $\lambda^{AB} = 0$, where λ^{AB} is the standard two-factor interaction calculated from $\log \pi_{ij} - \frac{1}{2}\mu_{ij}^2/\sigma$, and $\eta^{AB} = 0$, the latter being in this special case equivalent to the factorial requirements of no two-factor interaction in means (Cox, 1984). Table 8 gives studentized canonical interactions and approximate and exact maximum likelihood estimates.

Table 7. *Numerical example for* 2 *discrete and one continuous variable*

i	j	n_{ij}	$\bar{y}_{ij}$	$\hat{\sigma}$
0	0	180	32	60
1	0	120	5	60
0	1	20	26	60
1	1	180	−2	60

Table 8. *Numerical example for* 3 *variables. Studentized interactions and parameter estimates*

Type of interaction at levels 0	Studentized interaction	$\hat{\theta}_e$	$\tilde{\theta}_r$	$\hat{\theta}_r$
$\bar{\lambda}$	—	−8·2405	−8·1362	−8·1444
λ^A	−12·8026	−3·9292	−3·8355	−3·8240
λ^B	−1·4631	−0·3208	−0·2914	−0·2919
λ^{AB}	−0·1402	−0·0306	0·0000	0·0000
$\bar{\eta}$	—	0·2542	0·2458	0·2461
η^A	13·6368	0·2292	0·2219	0·2223
η^B	5·9075	0·0542	0·0502	0·0504
η^{AB}	−0·4897	0·0042	0·0000	0·0000
σ^{-1}	—	0·0167	0·0162	0·0162

Chi-squared goodness of fit (2 d.f.): based on $\tilde{\theta}_r$, 1·29; $\hat{\theta}_r$, 1·28.

Example 7: *Effects of child rearing styles*; 2 *discrete and* 2 *continuous variables.* Table 9 summarizes observations on 117 children of two binary variables A, vigilance of child; B, supportive behaviour of father and of, X, anxiety of child; Y, inconsistent behaviour

Table 9. 117 *children. Counts, means and covariance matrices for* 2 *discrete and* 2 *continuous variables*

i	j	n_{ij}	$\bar{x}_{ij}$	Means, $\bar{y}_{ij}$	Covariance matrices		
0	0	22	28·5455	23·3636	X	66·3388	45·0744
					Y	.	48·1405
1	0	39	33·6154	25·4103	X	41·1598	23·0039
					Y	.	43·4727
0	1	29	26·0134	21·2069	X	18·5065	9·0820
					Y	.	19·6124
1	1	27	30·8148	23·4815	X	27·4102	7·4966
					Y	.	33·7311

of mother scored on a continuous scale (Kohlmann, Schuhmacher & Streit, 1988). A reasonable hypothesis is that supportive or unsupportive behaviour of the father will have no direct effect on the child's vigilance, i.e. on whether the child is intensely collecting information on potentially threatening events. This corresponds to the hypothesis $A \perp\!\!\!\perp B|(X, Y)$. Again this model needs iterative fitting and leads to a chi-squared goodness of fit (6 d.f.) of 4·46 on 6 degrees of freedom based on $\hat{\theta}_r$ and of 4·55 based on $\tilde{\theta}_r$. Table 10 summarizes some relevant results of fitting.

Table 10. 117 *children. Estimated interactions involving both A and B*

Type of interaction at levels 0	Studentized interaction	$\hat{\theta}_e$
λ^{AB}	0·74	1·6304
η^{ABX}	−1·00	−0·1410
η^{ABY}	−0·14	−0·0172
ψ^{ABX}	−0·94	−0·0060
ψ^{ABXY}	0·15	0·0008
ψ^{ABY}	−0·37	−0·0025

Note that in these last examples the structure of the CG-models is such that absence of interaction involves a special linking of the parameters defining the continuous and discrete components of the model and hence is not assessed from fairly standard analysis of continuous and discrete aspects separately.

While it is necessary to use canonical parameters in the last examples, it is desirable so far as feasible to keep to orthogonal parameters, in particular to avoid the large differences between $\hat{\theta}_e$ and $\hat{\theta}_r$ consequent on major nonorthogonality. For instance, if the canonical parameters are used in fitting a random sample from a single normal distribution the correlation between the estimates is $-(1+\frac{1}{2}\text{CV}^2)^{-1}$, where CV is the coefficient of variation, standard deviation divided by mean, and hence is often close to 1 with consequent unreliability in the estimation formula (2). A similar conclusion is likely in comparable more complicated cases.

7. Discussion

We are very grateful to a referee for suggesting a connection with the generalized least squares method of Grizzle, Starmer & Koch (1969) for the analysis of log linear models for multinomial data. In fact a rather general relation of this kind holds for certain curved exponential family models embedded in full exponential family models, in particular for generalized linear models (McCullagh & Nelder, 1989).

Suppose that, possibly after reduction by sufficiency, we have a vector Y, with $E(Y) = \eta$, such that for some function $h(.)$, the reduced model is $h(\eta) = x_\theta\theta$ and that the extended model is a full exponential family with moments or canonical statistic Y; for the extended model we write $h(\eta) = z\phi$, where $z = (x_\theta, x_\gamma)$, $\phi^{\mathrm{T}} = (\theta_e^{\mathrm{T}}, \gamma_e^{\mathrm{T}})$. Note that $\hat{\eta}_e = Y$.

Examples include a multinomial model, with Y the cell proportions, π the cell probabilities, arbitrary under the extended model subject to $\Sigma\pi_l = 1$, and with $h(\pi_l) = \log \pi_l$, for instance, and the two regressions model, Example 5. Here the vector Y consists of the right-hand sides of the two least squares equations and the matrix of residual sums of squares and products.

Under the extended model with canonical statistic Y, maximum likelihood estimates are obtained by equating the canonical statistics to their expectations, or equivalently by writing $h(\hat{\eta}_e) = h(Y) = z\hat{\phi}_e$. We write V_e for the asymptotic covariance matrix of $h(Y)$ obtained from the observed information matrix under the extended model; note that V_e is determined by the data and does not involve unknown parameters. The asymptotic covariance matrix of $\hat{\phi}_e$ is $(z^{\mathrm{T}} V_e^- z)^{-1}$, where generalized inverses could be avoided, but are convenient to cover the multinomial case where $\Sigma\pi_l = 1$ and Y is the vector of all cell proportions, also summing to one. The key equation (5) now gives

$$\tilde{\theta}_e = \hat{\theta}_e + (x_\theta^{\mathrm{T}} V_e^- x_\theta)^{-1}(x_\theta V_e^- x_\gamma)\hat{\gamma}_e.$$

An apparently different approach is to apply the method of generalized least squares (Aitken, 1935) to the reduced model using as 'weighting matrix', however, V_e, the covariance matrix under the extended model, i.e.

$$\tilde{\theta}_r^* = (X_\theta^{\mathrm{T}} V_e^- X_\theta)^{-1}\{X_\theta^{\mathrm{T}} V_e^- h(Y)\}.$$

We can write the equation defining $\hat{\phi}_e$ as $(z^{\mathrm{T}} V_e^- z)\hat{\phi}_e = z^{\mathrm{T}} V_e^- h(Y)$. After expressing this in partitioned form corresponding to (θ, γ) it follows that $\tilde{\theta}_r = \tilde{\theta}_r^*$.

Special cases are the procedure of Grizzle et al. (1969), in which V_e is determined via the covariance matrix of cell proportions in a multinomial distribution, and Example 5, with the covariance matrix estimated from the residuals in the extended model fit of separate least squares equations. The exact maximum likelihood estimates satisfy an equation of the same form as that for $\tilde{\theta}_r^*$ but with V_e, and in general X_θ, calculated at $\hat{\theta}$; see, for example, Cox & Snell (1989, p. 176). We do not address the difficult issues of convergence to and uniqueness of solution of the formal maximum likelihood estimating equations. Note, however, that when $h(.)$ is itself linear, then the generalized least squares estimates $\tilde{\theta}_r^*$, and hence $\tilde{\theta}_r$, will be close to the maximum likelihood estimates $\hat{\theta}_r$ whenever V_e calculated from the data is close to V_e calculated at $\hat{\theta}$.

REFERENCES

AITKEN, A. C. (1935). On least squares and linear combination of observations. *Proc. R. Soc. Edin.* A **55**, 42–8.

BARTLETT, M. S. (1935). Contingency table interactions. *J. R. Statist. Soc., Suppl.* **5**, 248–52.

BIRCH, M. W. (1963). Maximum-likelihood in three-way contingency tables. *J. R. Statist. Soc.* B **25**, 220–3.

COCHRAN, W. G. (1938). The omission or addition of an independent variate in multiple linear regression. *J. R. Statist. Soc.*, Suppl **5**, 171–6.

COX, D. R. (1984). Interaction. *Int. Statist. Rev.* **52**, 1–31.

COX, D. R. & REID, N. (1987). Parameter orthogonality and approximate conditional inference (with discussion). *J. R. Statist. Soc.* B **49**, 1–39.

COX, D. R. & SNELL, E. J. (1989). *Analysis of Binary Data*, 2nd ed. London: Chapman and Hall.

DARROCH, J. N., LAURITZEN, S. L. & SPEED, T. P. (1980). Markov fields and log-linear models for contingency tables. *Ann. Statist.* **8**, 522–39.

DEMPSTER, A. P. (1971). An overview of multivariate data analysis. *J. Mult. Anal.* **1**, 316–46.

DEMPSTER, A. P. (1972). Covariance selection. *Biometrics* **28**, 157–75.

DEMPSTER, A. P. (1973). Aspects of the multinomial logit model. In Proc. 3rd Symp. Mult.-An., Ed. P. R. Krishnaiah, pp. 129–42. New York: Academic Press.

EDWARDS, D. (1990). Hierarchical mixed interaction models (with discussion). *J. R. Statist. Soc.* B **52**, 3–20.

EFRON, B. (1975). Defining the curvature of a statistical problem (with applications to second-order efficiency) (with discussion). *Ann. Statist.* **3**, 1189–242.

FRYDENBERG, M. & EDWARDS, D. (1989). A modified iterative proportional scaling algorithm for estimation in regular exponential families. *Comp. Statist. Data Anal.* **8**, 143–53.

GOOD, I. J. (1958). The interaction algorithm and practical Fourier analysis. *J. R. Statist. Soc.* B **20**, 361–72.

GOODMAN, L. A. (1970). The multivariate analysis of qualitative data. *J. Am. Statist. Assoc.* **65**, 226–56.

GRIZZLE, J. E., STARMER, C. F. & KOCH, G. C. (1969). Analysis of categorical data by linear models. *Biometrics* **28**, 137–56.

HABERMAN, S. J. (1974). *The Analysis of Frequency Data.* University of Chicago Press.

HAAVELMO, T. (1943). The statistical implications of a system of simultaneous equations. *Econometrica* **11**, 1-12.

ISSERLIS, L. (1918). On a formula for the product-moment correlation of any order of a normal frequency distribution in any number of variables. *Biometrika* **12**, 134-9.

KENNY, D. A. (1979). *Correlation and Causation.* New York: Wiley.

KOHLMANN, C. W., SCHUHMACHER, A. & STREIT, R. (1988). Parental child rearing behaviour and the development of trait anxiety in children: support as a moderator variable? *Anxiety Res.* **1**, 53-64.

LAURITZEN, S. L. & WERMUTH, N. (1989). Graphical models for associations between variables, some of which are qualitative and some quantitative. *Ann. Statist.* **17**, 31-54.

MCCULLAGH, P. & NELDER, J. A. (1989). *Generalized Linear Models*, 2nd ed. London: Chapman and Hall.

SPIELBERGER, C. D. (1983). *Manual for the State-Trait Anxiety Inventory.* Palo Alto: Consulting Psychologists Press.

SPIELBERGER, C. D., JACOBS, G. A., RUSSELL, S. F. & CRANE, R. S. (1983). Assessment of anger: the state-trait anger scale. In *Advances in Personality Assessment*, **2**, Ed. J. N. Butcher and C. D. Spielberger, pp. 161-89.

WERMUTH, N. & KOLLER, S. (1976). Systematik multivariater Korrelationsmuster angewandt auf die Symptomkorrelation von Krankheiten. In *Klinisch-Statist Forschung*, Ed. S. Koller and J. Berger, pp. 111-20. Stuttgart: Schattauer.

WERMUTH, N. & LAURITZEN, S. L. (1990). On substantive research hypotheses, conditional independence graphs and graphical chain models (with discussion). *J. R. Statist. Soc.* B **52**, 21-72.

YATES, F. (1937). *The Design and Analysis of Factorial Experiments.* Harpenden: Imperial Bureau of Soil Science.

[*Received November* 1989. *Revised March* 1990)