Statistica Sinica **20** (2010), 823-836

CHANGING PARAMETERS BY PARTIAL MAPPINGS

Michael Wiedenbeck and Nanny Wermuth

GESIS Leibniz Institute for the Social Sciences and Chalmers/University of Gothenburg

Abstract: Changes between different sets of parameters are often needed in multivariate statistical modeling, such as transformations within linear regression or in exponential models. There may, for instance, be specific inference questions based on subject matter interpretations, alternative well-fitting constrained models, compatibility judgements of seemingly distinct constrained models, or different reference priors under alternative parameterizations.

We introduce and discuss a partial mapping, called partial replication, and relate it to a more complex mapping, called partial inversion. Both operations are used to decompose matrix operations, to explain recursion relations among sets of linear parameters, to change between different types of linear models, to approximate maximum-likelihood estimates in exponential family models under independence constraints, and to switch partially between sets of canonical and moment parameters in exponential family distributions or between sets of corresponding maximum-likelihood estimates.

Key words and phrases: Exponential family, independence constraints, matrix operators, partial inversion, partial replication, reduced model estimates, REML-estimates, sandwich estimates.

1. Definitions and Properties of Two Operators

1.1. Partial inversion

We start with a linear function connecting two real-valued column vectors y and x via a square matrix M of dimension d for which all principal submatrices are invertible,

$$My = x. \tag{1.1}$$

The index sets of rows and columns of M coincide and are ordered as $V = (1, \ldots, d)$. For an arbitrary subset a of V, we consider first a split of V ordered as (a, b) with $b = V \setminus a$. Such a split will correspond later to a change in conditioning sets of variables.

A linear operator applied to (1.1), called partial inversion, may be used to invert M in a sequence of steps and forms a starting point for proving many properties of graphical Markov models, see Wermuth and Cox (2004, 2008), Marchetti and Wermuth (2009), and Wermuth (2009). Partial inversion is a minimally modified version of Gram-Schmidt orthogonalisation and of the sweep operator (Dempster (1969)). The changes are such that an operator with attractive features results, for proofs of its properties given below in this subsection, see Wermuth, Wiedenbeck and Cox (2006).

After partial inversion applied to rows and columns a of M, denoted by $\text{inv}_a M$, the argument and image at (1.1), relating to a, are exchanged

$$\text{inv}_a M \begin{pmatrix} x_a \\ y_b \end{pmatrix} = \begin{pmatrix} y_a \\ x_b \end{pmatrix}, \qquad \text{inv}_a M = \begin{pmatrix} M_{aa}^{-1} & -M_{aa}^{-1} M_{ab} \\ M_{ba} M_{aa}^{-1} & M_{bb.a} \end{pmatrix}. \tag{1.2}$$

The matrix $M_{bb.a} = M_{bb} - M_{ba} M_{aa}^{-1} M_{ab}$ is often called the Schur complement of M_{bb}, after Issai Schur (1875-1941), and M_{aa}^{-1} denotes the inverse of the submatrix M_{aa} of M.

Partial inversion with respect to b applied to $\text{inv}_a M$ is denoted in several equivalent ways depending on the context

$$\text{inv}_b(\text{inv}_a M) = \text{inv}_b \circ \text{inv}_a M = \text{inv}_V M = \text{inv}_{ab} M,$$

where $\text{inv}_V M$ yields the inverse of M. Some basic properties of partial inversion are

$$\begin{aligned} &\text{(i) } \text{inv}_a \circ \text{inv}_a M = M, \\ &\text{(ii) } (\text{inv}_a M)^{-1} = \text{inv}_b M, \\ &\text{(iii) } \text{inv}_a M = \text{inv}_b (M^{-1}). \end{aligned} \tag{1.3}$$

The operator is commutative and can be undone. More precisely, for three disjoint subsets α, β, γ of V, with $a = \alpha \cup \beta$

$$\begin{aligned} &\text{(i) } \text{inv}_\alpha \circ \text{inv}_\beta M = \text{inv}_\beta \circ \text{inv}_\alpha M, \\ &\text{(ii) } \text{inv}_{\alpha\beta} \circ \text{inv}_{\beta\gamma} M = \text{inv}_{\alpha\gamma} M. \end{aligned} \tag{1.4}$$

With $c = \beta \cup \gamma$, (1.4) gives also $\text{inv}_a \circ \text{inv}_c M = \text{inv}_{a \triangle c} M$, where $a \triangle c = \alpha \cup \gamma$ denotes the symmetric difference of a and c, that is the union without the intersection.

1.2. Partial replication

We now introduce another operator to be applied to (1.1), called partial replication, which represents a partial mapping and provides a decomposition of partial inversion, see (1.8) below.

After partial replication applied to rows and columns a of M, denoted by $\text{rep}_a M$, the argument relating to a and denoted by y_a, is replicated while the relation for the argument of $b = V \setminus a$ is preserved as at (1.1)

$$\text{rep}_a M \begin{pmatrix} y_a \\ y_b \end{pmatrix} = \begin{pmatrix} y_a \\ x_b \end{pmatrix}, \qquad \text{rep}_a M = \begin{pmatrix} I_{aa} & 0_{ab} \\ M_{ba} & M_{bb} \end{pmatrix}. \tag{1.5}$$

Partial replication is denoted in a way analogous to partial inversion

$$\mathrm{rep}_b(\mathrm{rep}_a M) = \mathrm{rep}_b \circ \mathrm{rep}_a M = \mathrm{rep}_V M = \mathrm{rep}_{ab} M,$$

where $\mathrm{rep}_V M$ yields the identity matrix I.

By direct computation, a basic property of partial replication is

$$\mathrm{rep}_a \circ \mathrm{rep}_a M = \mathrm{rep}_a M,$$

and the following matrix forms relate to the components of $\mathrm{inv}_b M$

$$\begin{aligned}
&\text{(i)}\ \ (\mathrm{rep}_a M)^{-1} = \begin{pmatrix} I_{aa} & 0_{ab} \\ -\,M_{bb}^{-1} M_{ba} & M_{bb}^{-1} \end{pmatrix} = \mathrm{rep}_a(\mathrm{inv}_b M), \\
&\text{(ii)}\ M(\mathrm{rep}_a M)^{-1} = \begin{pmatrix} M_{aa.b} & M_{ab} M_{bb}^{-1} \\ 0_{ba} & I_{bb} \end{pmatrix} = \mathrm{rep}_b(\mathrm{inv}_b M).
\end{aligned} \tag{1.6}$$

The last matrix product has been used in the numerical technique of block Gaussian elimination as one special form in which the Schur complement $M_{aa.b}$ of M_{aa} occurs.

Two derived properties of partial replication are to be listed next in the same order as the corresponding properties of partial inversion in (1.4). As before, we take three disjoint subsets α, β, γ of V and $a = \alpha \cup \beta$, then

$$\begin{aligned}
&\text{(i)}\ \ \mathrm{rep}_\alpha \circ \mathrm{rep}_\beta M = \mathrm{rep}_\beta \circ \mathrm{rep}_\alpha M, \\
&\text{(ii)}\ \mathrm{rep}_{\alpha\beta} \circ \mathrm{rep}_{\beta\gamma} M = \mathrm{rep}_{\alpha\beta\gamma} M.
\end{aligned} \tag{1.7}$$

Thus, partial replication shares with partial inversion the property of commutativity (i), but in contrast to partial inversion, it cannot be undone, but has the expansion property (1.7)(ii). Nevertheless, partial inversion can be expressed in terms of partial replication. By direct computation

$$\mathrm{inv}_a M = (\mathrm{rep}_a M)(\mathrm{rep}_b M)^{-1}. \tag{1.8}$$

1.3. Partial inversion combined with partial replication

By direct computation, basic properties of the two operators combined are

$$\begin{aligned}
&\text{(i)}\ \ \ \mathrm{inv}_a \circ \mathrm{rep}_a M = \mathrm{rep}_a M, \\
&\text{(ii)}\ \ \mathrm{inv}_b \circ \mathrm{rep}_a M = (\mathrm{rep}_a M)^{-1} = \mathrm{rep}_a \circ \mathrm{inv}_b M, \\
&\text{(iii)}\ \mathrm{rep}_b \circ \mathrm{inv}_b M = M(\mathrm{rep}_a M)^{-1}.
\end{aligned} \tag{1.9}$$

Thus, the components of $\mathrm{inv}_b M$ in (1.6) are expressed here with (1.9)(ii) and (iii).

For a partition of V into α, β, γ, some of the derived properties are

$$\begin{aligned} &\text{(i) } \mathrm{inv}_\alpha \circ \mathrm{rep}_\beta M = \mathrm{rep}_\beta \circ \mathrm{inv}_\alpha M, \\ &\text{(ii) } \mathrm{inv}_{\alpha\beta} \circ \mathrm{rep}_{\beta\gamma} M = \mathrm{inv}_\alpha \circ \mathrm{rep}_{\beta\gamma} M. \end{aligned} \tag{1.10}$$

Thus, a contraction property is obtained with (1.10) (ii) instead of the expansion property (1.7) (ii) of partial replication or the symmetric difference in (1.4) (ii) of partial inversion.

Partial replication on $c = \beta \cup \gamma$ applied to a matrix partial inverted on $a = \alpha \cup \beta$ gives, just like (1.8), a matrix product involving two partially replicated matrices, one with respect to $a \triangle c = \alpha \cup \gamma$, the other with respect to $b = \gamma \cup \delta$ as

$$\mathrm{rep}_c \circ \mathrm{inv}_a M = (\mathrm{rep}_{a \triangle c} M)(\mathrm{rep}_b M)^{-1}. \tag{1.11}$$

For the proof of (1.11), we recall from the definitions of partial replication (1.5) and partial inversion (1.2) that

$$\mathrm{rep}_b M y = \begin{pmatrix} x_a \\ y_b \end{pmatrix}, \qquad \mathrm{inv}_a M \begin{pmatrix} x_a \\ y_b \end{pmatrix} = \begin{pmatrix} y_a \\ x_b \end{pmatrix}.$$

Partial replication of $\mathrm{inv}_a M$ with respect to $c = \beta \cup \gamma$ and direct computation give

$$(\mathrm{rep}_c \circ \mathrm{inv}_a M)(\mathrm{rep}_b M) y = (\mathrm{rep}_{\alpha\gamma} M) y,$$

thus completing the proof. A matrix proof is given in the Appendix.

In the following sections, the two matrix operators are applied to quite different statistical themes. In Section 2, they simplify proofs of a number of results known for linear models. In Sections 3 and 4, linear relations are obtained for sets of parameters and for sets of estimates in nonlinear models within the exponential family of distribution.

2. Decompositions of Partial Inversion

2.1. Partially inverted covariance matrices

Partial replication provides with (1.8) a decomposition of partial inversion. Let Σ denote the joint covariance matrix of mean-centered random vector variables Y_a, Y_b, then

$$\begin{aligned} (\mathrm{rep}_b \Sigma)\,(\mathrm{rep}_a \Sigma)^{-1} &= \begin{pmatrix} \Sigma_{aa} & \Sigma_{ab} \\ 0_{ba} & I_{bb} \end{pmatrix} \begin{pmatrix} I_{aa} & 0_{ab} \\ -\,\Sigma_{bb}^{-1}\Sigma_{ba} & \Sigma_{bb}^{-1} \end{pmatrix} = \begin{pmatrix} \Sigma_{aa|b} & \Pi_{a|b} \\ -\,\Pi_{a|b}^{\mathrm{T}} & \Sigma_{bb}^{-1} \end{pmatrix} \\ &= \mathrm{inv}_b \Sigma, \end{aligned}$$

where $\Sigma_{bb}^{-1} = \Sigma^{bb.a}$ is the marginal concentration matrix of Y_b and

$$\Sigma_{aa|b} = \Sigma_{aa} - \Sigma_{ab}\Sigma_{bb}^{-1}\Sigma_{ba}, \qquad \Pi_{a|b} = \Sigma_{ab}\Sigma_{bb}^{-1},$$

are the parameter matrices in linear least-squares regression of Y_a on Y_b, i.e., in the linear multivariate regression model defined by

$$Y_a = \Pi_{a|b} Y_b + \epsilon_a,\ \mathrm{E}\,(\epsilon_a) = 0,\ \mathrm{Cov}\,(\epsilon_a, Y_b) = \mathrm{E}\,(\epsilon_a Y_b^{\mathrm{T}}) = 0,\ \mathrm{Cov}\,(\epsilon_a) = \Sigma_{aa|b}. \quad (2.1)$$

The interpretation of $\Pi_{a|b}$ as a matrix of regression coefficients results by post multiplication with Y_b^{T} and taking expectations, that is with $\mathrm{E}\,(Y_a Y_b^{\mathrm{T}}) - \Pi_{a|b} \mathrm{E}\,(Y_b Y_b^{\mathrm{T}}) = 0$.

By (1.3)(iii), we know that $\mathrm{inv}_b \Sigma = \mathrm{inv}_a \Sigma^{-1}$, therefore the three types of parameter matrices can equivalently be expressed in terms of the components $\Sigma^{aa}, \Sigma^{ab}, \Sigma^{ba}, \Sigma^{bb}$ of the concentration matrix Σ^{-1} as

$$\Sigma_{aa|b} = (\Sigma^{aa})^{-1},\ \Pi_{a|b} = -(\Sigma^{aa})^{-1} \Sigma^{ab},\ \Sigma_{bb}^{-1} = \Sigma^{bb.a} = \Sigma^{bb} - \Sigma^{ba} (\Sigma^{aa})^{-1} \Sigma^{ab}. \quad (2.2)$$

Zero constraints on $\Pi_{a|b}$ have been studied especially in econometrics. The possible existence of multiple solutions of estimating equations, derived by maximizing the Gaussian likelihood function, in a Zellner model, also called seemingly unrelated regressions (Zellner (1962)), has more recently been demonstrated by Drton and Richardson (2004).

Zero constraints on concentrations, such as in Σ_{bb}^{-1}, had been introduced as a tool for parsimonious estimation of covariances, called covariance selection (Dempster (1972)). The Gaussian likelihood function has a unique maximum for all possible sets of zero concentrations though for some models, iterative algorithms are needed to find the solution.

Zero constraints on covariances, such as in $\Pi_{a|b}$, have been studied by Anderson (1969, 1973), see also Wermuth, Cox and Marchetti (2006).

With $\rho_{ij.k}$ denoting a partial correlation, the hypotheses for a single zero for pair (i, j) in $\Pi_{a|b}$, $\Sigma_{aa|b}$, Σ_{bb}^{-1} differ and are equivalent, respectively, to

$$\rho_{ij.b\setminus j} = 0, \quad \rho_{ij.b} = 0, \quad \rho_{ij.b\setminus\{i,j\}} = 0.$$

In the context of chain graph models, see Wermuth and Cox (2004), Drton (2009), the notion of multivariate regression and covariance selection is extended to general types of distributions, in which the correlation coefficient is of no or of little relevance. Then, the corresponding hypotheses are those of conditional independence, say of Y_a and Y_b given Y_c, written compactly as $a \perp\!\!\!\perp b|c$. The above three hypotheses turn then, respectively, into

$$i \perp\!\!\!\perp j|b \setminus j, \qquad i \perp\!\!\!\perp j|b, \qquad i \perp\!\!\!\perp j|b \setminus \{i, j\}$$

and are captured by graphs, each having a different type of edge. Then, transformation of matrix representation of graphs mimic the transformations for linear parameter matrices discussed here; see Marchetti and Wermuth (2009) and Wermuth (2009) for partial inversion and partial closure of paths in graphs.

2.2. Three recursion relations among linear parameter matrices

Let variables be partitioned according to $V = (a, \gamma, \delta)$, again with $b = \gamma \cup \delta$, but $a = \alpha$, then which systematic changes are to be expected among the two sets of parameters obtained with $\text{inv}_b\Sigma = \text{inv}_{\gamma\delta}\Sigma$ and $\text{inv}_\delta\Sigma$? Written explicitly, the two matrices are

$$\text{inv}_b\Sigma = \left(\begin{array}{c:cc} \Sigma_{aa|b} & \Pi_{a|\gamma.\delta} & \Pi_{a|\delta.\gamma} \\ \hdashline \sim & \Sigma^{\gamma\gamma.a} & \Sigma^{\gamma\delta.a} \\ \sim & \cdot & \Sigma^{\delta\delta.a} \end{array}\right), \quad \text{inv}_\delta\Sigma = \left(\begin{array}{cc:c} \Sigma_{aa|\delta} & \Sigma_{a\gamma|\delta} & \Pi_{a|\delta} \\ \cdot & \Sigma_{\gamma\gamma|\delta} & \Pi_{\gamma|\delta} \\ \hdashline \sim & \sim & \Sigma^{\delta\delta.a\gamma} \end{array}\right),$$

where $\cdot$ indicates an entry in a symmetric matrix and $\sim$ an entry in a matrix that is symmetric except for the sign.

The matrices $\Sigma_{aa|b}$, $\Pi_{a|b}$ are as defined before, but the latter is split into two components corresponding to the two explanatory variables Y_γ, Y_δ, as $\Pi_{a|b} = (\Pi_{a|\gamma.\delta} \quad \Pi_{a|\delta.\gamma})$ Similarly, e.g. $\Sigma^{\gamma\delta.a}$ is the component of $\Sigma^{bb.a}$ corresponding to Y_γ, Y_δ.

By (1.4) (ii), we know $\text{inv}_\delta\Sigma = \text{inv}_\gamma(\text{inv}_b\Sigma)$, so that (1.8) applied to this form of $\text{inv}_\delta\Sigma$, and (1.9) applied to the resulting inverse matrix, gives

$$\text{inv}_\delta\Sigma = (\text{rep}_\gamma \circ \text{inv}_b\Sigma)(\text{rep}_{a\delta} \circ \text{inv}_\delta\Sigma)$$

or, written explicitly,

$$\text{inv}_\delta\Sigma = \begin{pmatrix} \Sigma_{aa|b} & \Pi_{a|\gamma.\delta} & \Pi_{a|\delta.\gamma} \\ 0_{\gamma a} & I_{\gamma\gamma} & 0_{\gamma\delta} \\ -\Pi^{\mathrm{T}}_{a|\delta.\gamma} & \Sigma^{\delta\gamma.a} & \Sigma^{\delta\delta.a} \end{pmatrix} \begin{pmatrix} I_{aa} & 0_{a\gamma} & 0_{a\delta} \\ \Sigma_{\gamma a|\delta} & \Sigma_{\gamma\gamma|\delta} & \Pi_{\gamma|\delta} \\ 0_{\delta a} & 0_{\delta\gamma} & I_{\delta\delta} \end{pmatrix}. \qquad (2.3)$$

The second matrix term in (2.3) results with equations (1.9) (ii) as $(\text{rep}_{a\delta} \circ \text{inv}_b\Sigma)^{-1} = \text{rep}_{a\delta} \circ \text{inv}_\delta\Sigma$. This leads to the following equalities that extend those of (2.2):

$$\Sigma_{\gamma\gamma|\delta} = (\Sigma^{\gamma\gamma.a})^{-1}, \quad \Pi_{\gamma|\delta} = -(\Sigma^{\gamma\gamma.a})^{-1}\Sigma^{\gamma\delta.a} \quad \Pi_{a|\gamma.\delta} = \Sigma_{a\gamma|\delta}\Sigma^{-1}_{\gamma\gamma|\delta} = -(\Sigma^{aa})^{-1}\Sigma^{a\gamma}.$$

Several known recursion relations are obtained directly with the matrix product (2.3), one for covariances (Anderson (1958, Sec. 2.5)) in position (a, a) as

$$\Sigma_{aa|\delta} = \Sigma_{aa|\gamma\delta} + \Pi_{a|\gamma.\delta}\Sigma_{\gamma a|\delta},$$

one for concentrations (Dempster (1969, Chap. 4)) in position (δ, δ) as

$$\Sigma^{\delta\delta.a\gamma} = \Sigma^{\delta\delta.a} + \Sigma^{\gamma\delta.a}\Pi_{\gamma|\delta},$$

and one for linear least-squares regression coefficients (Cochran (1938)) in position (a, δ) as

$$\Pi_{a|\delta} = \Pi_{a|\delta.\gamma} + \Pi_{a|\gamma.\delta}\Pi_{\gamma|\delta}. \tag{2.4}$$

Each of the above three equations relates a marginal to a conditional parameter matrix and quantifies the modifications that occur by changing the conditioning set. For extensions of (2.4) to nonlinear relations, in particular to conditions under which no change or at least no change in the direction of dependence may occur, see Cox and Wermuth (2003), Ma, Xie and Geng (2006), and Cox (2007).

2.3. Changing to different splits of three types of linear parameter matrices

Let $V = (\alpha, \beta, \gamma, \delta)$, again with $a = \alpha \cup \beta$, $b = \gamma \cup \delta$ and another split of V be $c = \beta \cup \gamma$ and $d = \alpha \cup \delta$. Then a change in parameters is defined implicitly by $\text{inv}_b\Sigma$ and $\text{inv}_d\Sigma$. These are, with partitions according to $(\alpha, \beta, \gamma, \delta)$,

$$\text{inv}_b\Sigma = \left(\begin{array}{cc|cc} \Sigma_{\alpha\alpha|b} & \Sigma_{\alpha\beta|b} & \Pi_{\alpha|\gamma.\delta} & \Pi_{\alpha|\delta.\gamma} \\ \cdot & \Sigma_{\beta\beta|b} & \Pi_{\beta|\gamma.\delta} & \Pi_{\beta|\delta.\gamma} \\ \hline \sim & \sim & \Sigma^{\gamma\gamma.a} & \Sigma^{\gamma\delta.a} \\ \sim & \sim & \cdot & \Sigma^{\delta\delta.a} \end{array}\right), \text{inv}_d\Sigma = \left(\begin{array}{c|cc|c} \Sigma^{\alpha\alpha.c} & \Pi^{\mathrm{T}}_{\beta|\alpha.\delta} & \Pi^{\mathrm{T}}_{\gamma|\alpha.\delta} & \Sigma^{\alpha\delta.c} \\ \hline \sim & \Sigma_{\beta\beta|d} & \Sigma_{\beta\gamma|d} & \Pi_{\beta|\delta.\alpha} \\ \sim & \cdot & \Sigma_{\gamma\gamma.d} & \Pi_{\gamma|\delta.\alpha} \\ \hline \cdot & \sim & \sim & \Sigma^{\delta\delta.c} \end{array}\right).$$

If we let $\mathrm{E}\,(Y) = 0$ and

$$Z = \Sigma^{-1}Y, \tag{2.5}$$

then

$$\mathrm{Cov}\,(Z) = \mathrm{E}\,(ZZ^{\mathrm{T}}) = \Sigma^{-1}, \qquad \mathrm{Cov}\,(Y, Z) = \mathrm{E}\,(YZ^{\mathrm{T}}) = I,$$

so that the covariance matrix of Z is the concentration matrix of Y and components Y_i of Y and Z_j of Z are uncorrelated, whenever $i \neq j$. Equation (2.5) specifies covariance selection as a linear model if the assumption of a Gaussian distribution of Z is added.

The implicit change defined by $\text{inv}_b\Sigma$ and $\text{inv}_d\Sigma$ applied to (2.5) is then, by (1.2) between

$$\text{inv}_a\Sigma^{-1}\begin{pmatrix} Z_\alpha \\ Z_\beta \\ Y_\gamma \\ Y_\delta \end{pmatrix} = \begin{pmatrix} Y_\alpha \\ Y_\beta \\ Z_\gamma \\ Z_\delta. \end{pmatrix}, \qquad \text{inv}_c\Sigma^{-1}\begin{pmatrix} Y_\alpha \\ Z_\beta \\ Z_\gamma \\ Y_\delta \end{pmatrix} = \begin{pmatrix} Z_\alpha \\ Y_\beta \\ Y_\gamma \\ Z_\delta \end{pmatrix}. \tag{2.6}$$

The first set of linear models in (2.6) can be written compactly by using the notation of Section 2.1 as

$$Y_a = \Pi_{a|b}Y_b + \Sigma_{aa|b}Z_a, \qquad \Sigma^{bb.a}Y_b = Z_b + \Pi^{\mathrm{T}}_{a|b}Z_a,$$

and there are the same type of expressions for the second set. Thus, equations (2.6), typically with added sets of zero constraints, specify linear multivariate regression models (2.1) for Y_a regressed on Y_b and for Y_c regressed on Y_d, and marginal concentration matrix models for Y_b and for Y_d. The two sets of parameter matrices are therefore

$$(\Sigma_{aa|b}, \quad \Pi_{a|b}, \quad \Sigma^{bb.a}), \qquad (\Sigma_{cc|d}, \quad \Pi_{c|d}, \quad \Sigma^{dd.c}).$$

The change in the two sets is obtained by the following reversible transformation, since for $a = \alpha \cup \beta$ and $c = \beta \cup \gamma$ and starting from $\mathrm{inv}_a\Sigma^{-1}$, one needs to remove α and add γ, i.e., to partially invert on the symmetric difference $a \bigtriangleup c = \alpha \cup \gamma$, see (1.4) (i), (ii), to get

$$\mathrm{inv}_c\Sigma^{-1} = \mathrm{inv}_{a\bigtriangleup c} \circ \mathrm{inv}_a\Sigma^{-1} = (\mathrm{rep}_{a\bigtriangleup c} \circ \mathrm{inv}_a\Sigma^{-1})(\mathrm{rep}_{a\bigtriangleup d} \circ \mathrm{inv}_a\Sigma^{-1})^{-1}. \quad (2.7)$$

The second equality results from (1.8) after noting that the complement of $a \bigtriangleup c$ is $a \bigtriangleup d$.

If instead the main interest is in the change of basis from $(Z_a^{\mathrm{T}}, Y_b^{\mathrm{T}})^{\mathrm{T}}$ to $(Y_\alpha^{\mathrm{T}}, Z_c^{\mathrm{T}}, Y_\delta^{\mathrm{T}})^{\mathrm{T}}$, then this is achieved by (1.5) and (1.11) to give

$$(\mathrm{rep}_b\Sigma^{-1})\, Y = \begin{pmatrix} Z_a \\ Y_b \end{pmatrix}, \quad (\mathrm{rep}_c \circ \mathrm{inv}_a\Sigma^{-1})(\mathrm{rep}_b\Sigma^{-1})\, Y = (\mathrm{rep}_{a\bigtriangleup c}\Sigma^{-1})\, Y = \begin{pmatrix} Y_\alpha \\ Z_c \\ Y_\delta \end{pmatrix}.$$

3. Estimation in Reduced Exponential Families

For a full exponential family, we take the log likelihood, after disregarding terms that do not depend on the unknown parameter, in the form

$$s^T\phi - K(\phi),$$

where ϕ is the canonical parameter and s the sufficient statistic, a realization of the random variable S. The cumulant generating function of S under the full exponential family, see for example Cox (2006, Chap. 6),

$$K(\phi + t) - K(\phi),$$

gives the mean or moment parameter, $\eta = \nabla K(\phi)$, as the gradient of $K(\phi)$ with respect to ϕ, i.e., as a vector of first derivatives. The maximum-likelihood estimate of η is $\hat{\eta} = s$. The gradient of η with respect to ϕ gives the covariance matrix of S but also the concentration matrix of the maximum-likelihood estimate $\hat{\phi}$ of the canonical parameter, that is

$$\mathrm{Cov}\,(S) = \nabla\nabla^{\mathrm{T}}K(\phi) = \mathrm{con}(\hat{\phi}),$$

where $\hat{\phi}$ denotes for simplicity both a maximum-likelihood estimate of ϕ and the corresponding random variable. R. A. Fisher had interpreted $-\nabla\nabla^{\mathrm{T}}K(\phi)$ as the information about the canonical parameter contained in a single observation. Our notation $\mathsf{I} = \nabla\nabla^{\mathrm{T}}K(\phi)$ reminds of this.

Given $\hat{\eta}$ and $-\bar{\mathsf{I}}$, the observed information matrix, i.e., minus I evaluated at s, studentized statistics for testing that an individual component η_i of η is zero are obtained for a large sample size n as $\hat{\eta}_i/\,\bar{\mathsf{I}}_{ii}$. Similarly, since the random variable $\mathsf{I}^{-1}(\hat{\eta} - \eta)$ has mean zero and covariance matrix I^{-1} and hence the same mean and variance as $(\hat{\phi} - \phi)$, a studentized statistic for testing that an individual component ϕ_i of ϕ is zero, may for large n be computed as $\hat{\phi}_i/\,(\bar{\mathsf{I}}^{-1})_{ii}$.

At a maximum of the likelihood function under a full exponential model, also called often the saturated or the largest covering model, it holds that

$$\bar{\mathsf{I}}^{-1}\hat{\eta} = z, \qquad z = \bar{\mathsf{I}}^{-1}s, \tag{3.1}$$

which is in the form of (2.5) so that the results of the previous section apply to it, especially as used below for (3.4).

We now consider special reduced exponential models, those given by constraints $\eta_c = 0$ for some subset of elements of η, writing $\eta = (\eta_u, \eta_c)$. By differentiating the Lagrangian

$$s^T\phi - K(\phi) - \lambda^T\eta_c$$

with respect to ϕ, the maximum likelihood estimating equations are obtained as

$$\hat{\eta}_u = s_u - \hat{\mathsf{I}}_{uc}\hat{\mathsf{I}}_{cc}^{-1}s_c, \quad \hat{\eta}_c = 0, \tag{3.2}$$

where $\hat{\mathsf{I}}$ is the maximum likelihood estimate of I in the reduced model.

Since the solution of maximum-likelihood equations requires in general iterative algorithms and may not be unique, an efficient closed form approximation is useful, see Cox and Wermuth (1990), Wermuth, Cox and Marchetti (2006), that has been called the reduced model estimate of η_u. Equation (3.2) is thereby modified into

$$\tilde{\eta}_u = s_u - \tilde{\mathsf{I}}_{uc}\tilde{\mathsf{I}}_{cc}^{-1}s_c, \quad \tilde{\eta}_c = 0, \tag{3.3}$$

where the (u, c) and (c, c) components of $\hat{\mathsf{I}}$ in (3.2) have been replaced by the corresponding components of the asymptotic covariance matrix $\tilde{\mathsf{I}}$ of S and do not involve unknown parameters.

Equations (3.2) and (3.3) result also, by using (2.1), and $\mathrm{inv}_u\mathsf{I}^{-1} = \mathrm{inv}_c\mathsf{I}$ in (3.1) when the (u, c) and (c, c) components of the matrix I are replaced by $\hat{\mathsf{I}}$ and by $\tilde{\mathsf{I}}$, respectively, and the (u, u) component is evaluated at $\eta_c = 0$. Since the use of reduced model estimates is recommended only in situations in which the constraints agree well with the observed data, choosing the full observed matrix, $\bar{\mathsf{I}}$ of $\hat{\phi}$, under the saturated model and then partially inverting it on c, should not give estimates which differ much from those in (3.2) and (3.3). By (2.1) and (3.1), we then have

$$\bar{\eta}_u = s_u - \bar{\mathsf{I}}_{uc}\bar{\mathsf{I}}_{cc}^{-1} s_c, \quad \bar{\mathsf{I}}^{cc.u}\,\bar{\eta}_c = s_c - (\bar{\mathsf{I}}_{uc}\bar{\mathsf{I}}_{cc}^{-1})^{\mathrm{T}}\, s_u, \tag{3.4}$$

with $\bar{\eta}_u = \tilde{\eta}_u$, Cov $(\bar{\eta}_u) = \bar{\mathsf{I}}_{uu|c}$ and Cov $(\bar{\eta}_c) = \bar{\mathsf{I}}_{cc}$. In the case of a poor fit to the hypothesis $\eta_c = 0$, i.e., with some components of $\bar{\eta}_c$ deviating much from zero, (2.7) permits a direct change to the fit under an alternative hypothesis.

Since $\bar{\mathsf{I}}_{uc}\bar{\mathsf{I}}_{cc}^{-1}$ can be viewed as the observed coefficient of S_c in linear least-squares regression of S_u on S_c, standard results apply for testing that $0 = \mathsf{I}_{uc}\mathsf{I}_{cc}^{-1}$ given an estimate of the appropriate covariance matrix. Expansions into relevant interaction parameters lead to studentized statistics of interaction terms and provide insight into where a possibly poor fit is located. See Cox and Wermuth (1990) for examples, and Lauritzen and Wermuth (1989) for a discussion of interaction parameters in the case of Conditional-Gaussian (CG) distributions and of CG-regressions. The latter contain for instance logistic regression as a special case for which, in general, iterative fitting algorithms are needed to give $\hat{\eta}$ even under the saturated model, see Edwards and Lauritzen (2001).

The closed form estimates of η_u in (3.4) provide a new justification for the reduced model estimates: they result by partially inverting (3.1) with respect to the subset given by the set of unconstrained parameters. These estimates can also be viewed as the generalized least squares estimates of Aitken (1935), which turn for instance for the multinomial distribution to the estimates of Grizzle, Starmer, and Koch (1969), see Cox and Snell (1981, Appendix 1) Cox and Wermuth (1990, Sec. 7).

Under some general regularity conditions, relations as in (3.1) and hence the estimates of η_u in (3.4) arise in more general settings than for the exponential family from asymptotic theory, see for example Cox (2006, Chap. 6). They are then called sandwich estimates, introduced in a special context by Huber (1964), or in the context of generalized linear models they are called approximate residual maximum-likelihood (REML) estimates, derived by Patterson and Thompson (1971).

By similar arguments as above, the maximum-likelihood equations with zero constraints on canonical parameters are obtained in a form comparable to (3.2). From a theoretical viewpoint, these may be more attractive than zero constraints on moment parameters. First, if the constrained canonical parameters exist, then the maximum-likelihood equations have a unique solution. Second, the sets of minimal sufficient statistics are of reduced size. Third, estimates are available in closed form for Gaussian and for multinomial distributions provided the model is decomposable, that is, the associated independence graph can be arranged in a sequence of possibly overlapping but complete prime graphs, see Cox and Wermuth (1999). For non-decomposable models, it is in addition often simple to find a not much larger, decomposable covering model.

4. Switching Partially between Canonical and Moment Parameters

So far, we have considered only linear mappings even though these were relevant both for nonlinear model formulations and for correlated data. We now illustrate how changes between moment and canonical parameters in exponential families may be obtained in terms of partial replication.

A general formulation has been given by Cox (2006, Sec 6.4), including a short proof for orthogonality, i.e., uncorrelatedness of canonical and moment parameters and of the asymptotic independence of the corresponding maximum-likelihood estimates, see also Barndorff-Nielsen (1978, Sec. 9.8). These general results have often been proven for specific members of exponential families, involving the then necessary lengthy, detailed arguments.

In the notation of the present paper, for the moment parameter $\eta = \nabla K(\phi)$ and the canonical parameter ϕ with a maximum-likelihood estimate denoted by $\hat{\phi}$, let $s = (s_a, s_b)$ represent a split of the sufficient statistic into two column vectors and let ϕ_b be replaced by η_b, the corresponding component of η, to give the mixed parameter vector $\psi = (\phi_a, \eta_b)$. Then with

$$\mathsf{I} = \frac{\partial \eta^{\mathrm{T}}}{\partial \phi}, \qquad \mathrm{rep}_a \mathsf{I} = \frac{\partial \psi^{\mathrm{T}}}{\partial \phi} = \begin{pmatrix} I_{aa} & 0_{ab} \\ \mathsf{I}_{ba} & \mathsf{I}_{bb} \end{pmatrix},$$

one obtains, given $\hat{\phi}$, a maximum-likelihood estimate of ψ under the saturated model as

$$\hat{\psi} = \mathrm{rep}_a \bar{\mathsf{I}}\, \hat{\phi}, \qquad \hat{\mathrm{Cov}}\,(\hat{\psi}) = (\mathrm{rep}_a \bar{\mathsf{I}})\, \bar{\mathsf{I}}^{-1} (\mathrm{rep}_a \bar{\mathsf{I}})^{\mathrm{T}} = \begin{pmatrix} \bar{\mathsf{I}}^{-1}_{aa|b} & 0_{ab} \\ 0_{ba} & \bar{\mathsf{I}}_{bb} \end{pmatrix}. \tag{4.1}$$

The two random variables corresponding to $\hat{\phi}_a$ and $\hat{\eta}_b$ are uncorrelated and, given their asymptotic joint Gaussian distribution, they are also asymptotically independent.

One simple example is a joint Gaussian distribution with $\hat{\eta}_b$ the observed mean of Y_b and $\hat{\phi}_a$ the observed overall concentration matrix of Y_a. Another is for two dichotomous variables with $\hat{\eta}_b$ the difference in observed frequencies for Y_b and $\hat{\phi}_a$ the observed log-odds ratios of Y_a.

Two complementary mappings may, respectively, be represented by

$$\begin{pmatrix} \phi_a \\ \eta_b \end{pmatrix} = (\mathrm{rep}_a \mathsf{I})\, \phi, \qquad \begin{pmatrix} \eta_a \\ \phi_b \end{pmatrix} = (\mathrm{rep}_b \mathsf{I})\, \phi,$$

so that for such mappings, we have with the specific choice $M = \mathsf{I}$, as in (1.2) and (1.8),

$$\begin{pmatrix} \eta_a \\ \phi_b \end{pmatrix} = (\text{inv}_a \mathsf{I}) \begin{pmatrix} \phi_a \\ \eta_b \end{pmatrix}, \qquad \text{inv}_a \mathsf{I} = (\text{rep}_a \mathsf{I})(\text{rep}_b \mathsf{I})^{-1}. \tag{4.2}$$

One important consequence of (4.2) is that, given (4.1), the change from a split (a, b) to another split (c, d) with canonical parameters for c and moment parameters for d is possible by using (2.7) after just replacing Σ^{-1} by I. Another application, not treated here, is to constrained chain graph models of different types when the conditional distributions are members of the exponential family.

For the computation of the observed information matrix in the case of both discrete and continuous variables, see Dempster (1973) and Cox and Wermuth (1990). Whenever the population matrix I is replaced by its observed counterpart $\bar{\mathsf{I}}$, the linear relations between sets of parameters in possibly nonlinear models turn into a linear relation between the corresponding maximum-likelihood estimates.

Acknowledgement

We thank the Swedish Research Society and the Swedish Strategic Fund for supporting our cooperation via the Gothenburg Stochastic Center and via the GMM Center, as well as D. R. Cox and G. M. Marchetti and the referees for their most helpful comments.

Appendix

For a direct matrix proof of (1.11), let $N = \text{inv}_a M$. Further, denote the matrix obtained by partial replication with respect to $a \triangle c$ by Q and with respect to b by R,

$$\text{rep}_{a\triangle c} M = \begin{pmatrix} I_{\alpha\alpha} & 0_{\alpha\beta} & 0_{\alpha\gamma} & 0_{\alpha\delta} \\ M_{\beta\alpha} & M_{\beta\beta} & M_{\beta\gamma} & M_{\beta\delta} \\ 0_{\gamma\alpha} & 0_{\gamma\beta} & I_{\gamma\gamma} & 0_{\gamma\delta} \\ M_{\delta\alpha} & M_{\delta\beta} & M_{\delta\gamma} & M_{\delta\delta} \end{pmatrix}, \quad \text{rep}_b M = \begin{pmatrix} M_{\alpha\alpha} & M_{\alpha\beta} & M_{\alpha\gamma} & M_{\alpha\delta} \\ M_{\beta\alpha} & M_{\beta\beta} & M_{\beta\gamma} & M_{\beta\delta} \\ 0_{\gamma\alpha} & 0_{\gamma\beta} & I_{\gamma\gamma} & 0_{\gamma\delta} \\ 0_{\delta\alpha} & 0_{\delta\beta} & 0_{\delta\gamma} & I_{\delta\delta} \end{pmatrix}.$$

For (1.11) to hold, one needs to show that $QR^{-1} = \text{rep}_c N$, where

$$R^{-1} = \begin{pmatrix} M_{aa}^{-1} & -M_{aa}^{-1} M_{ab} \\ 0_{ba} & I_{bb} \end{pmatrix} = \begin{pmatrix} N_{aa} & N_{ab} \\ 0_{ba} & I_{bb} \end{pmatrix}.$$

The rows of components α and γ in QR^{-1} coincide directly with those of $\text{rep}_c N$. The rows of component β in $\text{rep}_c N$ result since for Q and R, the rows of β coincide so that the product $Q_{\beta V} R^{-1}$ gives zeros whenever the row index within

β for $Q_{\beta V}$ differs from the column index in R^{-1}, and is one otherwise. Finally, for the rows of components δ we have for $Q_{\delta V}R^{-1}$, by the defining equation for partial inversion (1.2),

$$(M_{\delta a}M_{aa}^{-1} \quad M_{\delta\gamma} - M_{\delta a}M_{aa}^{-1}M_{a\gamma} \quad M_{\delta\delta} - M_{\delta a}M_{aa}^{-1}M_{a\delta}) = (N_{\delta a} \quad N_{\delta b}),$$

which completes the proof.

References

Aitken, A. C. (1935). On least squares and linear combination of observations. *Proc. Roy. Soc. Edin.* **55**, 42-48.

Anderson, T. W. (1958). *An Introduction to Multivariate Statistical Analysis.* Wiley, New York.

Anderson, T. W. (1969). Statistical inference for covariance matrices with linear structure. In: *Multivariate Analysis* II (Edited by P.R. Krishnaiah), 55-66. Academic Press, New York.

Anderson, T. W. (1973). Asymptotically efficient estimation of covariance matrices with linear structure. *Ann. Statist.* **1**, 135-141.

Barndorff-Nielsen, O. E. (1978). *Information and Exponential Families in Statistical Theory.* Wiley, Chichester.

Cochran, W. G. (1938). The omission or addition of an independent variate in multiple linear regression. *Suppl. J. Roy. Statist. Soc.* **5**, 171-176.

Cox, D. R. (2006). *Principles of Statistical Inference.* Cambridge University Press, Cambridge.

Cox, D. R. (2007). On a generalization of a result of W. G. Cochran. *Biometrika* **94**, 755-759.

Cox, D. R. and Snell, E. J. (1981). *Applied Statistics.* Chapman and Hall, London.

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

Cox, D. R. and Wermuth, N. (1999). Likelihood factorizations for mixed discrete and continuous variables. *Scand. J. Statist.* **26**, 209-220.

Cox, D. R. and Wermuth, N. (2003). A general condition for avoiding effect reversal after marginalization. *J. Roy. Statist. Soc. Ser. B* **65**, 937-941.

Dempster, A. P. (1969). *Elements of Continuous Multivariate Analysis.* Addison-Wesley, Reading.

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

Dempster, A. P. (1973). Aspects of the multinomial logit model. In: Proc. 3rd Symp. Mult. Anal. (Edited by P.R. Krishnaiah), 129-142. Academic Press, New York.

Drton, M. (2009). Discrete chain graph models. *Bernoulli* **15**, 736-753.

Drton, M. and Richardson, T. S. (2004). Multimodality of the likelihood in the bivariate seemingly unrelated regression model. *Biometrika* **91**, 383-392.

Edwards, D. and Lauritzen, S. L. (2001). The TM algorithm for maximising a conditional likelihood function. *Biometrika* **88**, 961-972.

Grizzle, J. E., Starmer, C. F. and Koch, G. C., (1969). Analysis of categorical data by linear models. *Biometrics* **28**, 137-156.

Huber, P. J. (1964). Robust estimation of a location parameter. *Ann. Math. Stat.* **35**, 7-101.

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

Ma, Z., Xie, X. and Geng, Z. (2006). Collapsibility of distribution dependence. *J. Roy. Statist. Soc. Ser. B* **68**, 127-33.

Marchetti, G. M. and Wermuth, N. (2009). Matrix representations and independencies in directed acyclic graphs. *Ann. Statist.* **37**, 961-978.

Patterson, H. D. and Thompson, R. (1971). Recovery of interblock information when block sizes are unequal. *Biometrika*, **58**, 545-554.

Wermuth, N. (2009). Probability distributions with summary graph structure. Submitted.

Wermuth, N. and Cox, D. R. (2004). Joint response graphs and separation induced by triangular systems. *J. Roy. Statist. Soc. Ser. B* **66**, 687-717.

Wermuth, N. and Cox, D. R. (2008). Distortion of effects caused by indirect confounding. *Biometrika*, **95**, 17-33.

Wermuth, N., Cox, D. R. and Marchetti, G. M. (2006). Covariance chains. *Bernoulli* **12**, 841-862.

Wermuth, N., Wiedenbeck, M., and Cox, D. R. (2006). Partial inversion for linear systems and partial closure of independence graphs. *BIT, Num. Math.* **46**, 883-901.

Zellner, A. (1962). An efficient method of estimating seemingly unrelated regressions and tests for aggregation bias. *J. Amer. Statist. Assoc.* **57**, 348-368.

GESIS, Postfach 12 21 55, 68072 Mannheim, Germany.

E-mail: wiedenbeck@gesis.org

Chalmers/University of Gothenburg, Mathematical Statistics, 41296 Gothenburg, Sweden, Chalmers Tvärgata 3.

E-mail: wermuth@chalmers.se

(Received September 2007; accepted January 2009)