Bernoulli **12**(5), 2006, 841–862

Covariance chains

NANNY WERMUTH[1], D.R. COX[2] and GIOVANNI M. MARCHETTI[3]

[1]*Mathematical Statistics, Chalmers/Gothenburg University, Chalmers tvärgata 3, 41296 Göteborg, Sweden. E-mail: nanny.wermuth@uni-mainz.de*
[2]*Nuffield College, Oxford OX1 1NF, UK. E-mail: david.cox@nuff.ox.ac.uk*
[3]*Dipartimento di Statistica, Università degli Studi di Firenze 'Giuseppe Parenti', Viale Morgagni 59, Italy. E-mail: Giovanni.marchetti@unifi.it*

Covariance matrices which can be arranged in tridiagonal form are called covariance chains. They are used to clarify some issues of parameter equivalence and of independence equivalence for linear models in which a set of latent variables influences a set of observed variables. For this purpose, orthogonal decompositions for covariance chains are derived first in explicit form. Covariance chains are also contrasted to concentration chains, for which estimation is explicit and simple. For this purpose, maximum-likelihood equations are derived first for exponential families when some parameters satisfy zero value constraints. From these equations explicit estimates are obtained, which are asymptotically efficient, and they are applied to covariance chains. Simulation results confirm the satisfactory behaviour of the explicit covariance chain estimates also in moderate-size samples.

Keywords: canonical parameters; exponential families; graphical chain models; independence equivalence; latent variables; linear least-squares regressions; moment parameters; orthogonal decompositions; parameter equivalence; reduced models; structural equation models

1. Introduction

1.1. General remarks

Independence can arise in linear multivariate systems in a number of ways. There may be vanishing least-squares regression coefficients in a system defined recursively, vanishing elements in a concentration matrix, i.e. in the inverse of the covariance matrix, each reflecting a zero partial correlation given all remaining variables, or there may be vanishing elements in a covariance matrix, reflecting zero marginal correlations. In particular, we study covariance chains in which a series of component variables is such that only adjacent pairs have non-zero correlation, in contrast to concentration chains in which a sequence of only adjacent pairs have non-zero partial correlation. In the special context of stationary time series the former correspond to moving-average processes and the latter to autoregressive processes.

Covariance chains may be present directly for the component variables or for the residuals in a system of linear equations. Both can result from recursive processes in which the covariances of adjacent pairs are generated by latent, i.e. hidden, variables. Covariance chains for residuals are valuable in clarifying relations between the important concepts of parameter equivalence and independence equivalence.

1350–7265

A Gaussian covariance chain is an exponential family model with zero constraints on the moment parameters and with canonical parameters which are all non-vanishing. Therefore we solve the more general task of finding explicit asymptotically efficient estimates for an exponential family with zero constraints on the moment parameters. The results are studied in more detail for covariance chains.

1.2. Constrained covariance structures

We consider mean-centred random vector variables Y with component variables Y_i for i in $N = (1, \ldots, d)$, having an invertible covariance matrix Σ, with elements σ_{ij}, and concentration matrix Σ^{-1}, with elements σ^{ij}. Distinct types of constraints on the covariance structure of Y are captured by different types of graph which consist of a set of nodes N and of sets of edges connecting node pairs. Node i of N represents the component variable Y_i, and each missing edge coincides in different types of linear model with a parameter constrained to be zero. The types of graph considered in this paper are either subclasses of independence graphs, in which every edge represents a particular conditional relation of a variable pair, or graphs attached to structural equation models, for which this need not hold.

Independence graphs have at most one edge for each pair of nodes, and each missing edge corresponds to one particular independence statement. The statement that Y_i is linearly independent of Y_j given Y_C where C is the subset of the remaining nodes, $C = N\backslash\{ij\}$, means the vanishing of the corresponding partial correlation coefficient, i.e. $\rho_{ij|C} = 0$. In the case of a Gaussian distribution of Y, this is equivalent to the factorization of the joint conditional density of Y_i and Y_j given Y_C. The precise meaning of a conditioning set C is defined with the type of graph. Of the different types of independence graph, we describe in the following the parent graph, covariance graph and concentration graph (see also Cox and Wermuth 1993; Wermuth and Cox 1998, 2004).

For a parent graph, G^N_{par}, the nodes are ordered $N = (1, \ldots, d)$ and each edge is drawn as an (i,j) arrow starting from a parent node j and pointing to an offspring node $i < j$. The constraints defined by a parent graph are expressed in terms of nodes as the offspring node i being independent of node $j > i$ given all its parent nodes. The graph has an attached system of recursive linear equations with uncorrelated residuals which is also called a path analysis model or a linear triangular system.

For a covariance graph, G^N_{cov}, each edge is drawn as an (i,j) dashed (broken) line. Each missing (i,j) line defines marginal independence of Y_i and Y_j. This means in the attached linear covariance graph model that $\sigma_{ij} = \rho_{ij} = 0$, where ρ_{ij} denotes a simple correlation coefficient. In contrast, for a concentration graph, G^N_{con}, each edge is drawn as an (i,j) full (solid) line. Each missing (i,j) line defines independence of Y_i and Y_j given $Y_{N\backslash\{ij\}}$, implying that in the attached linear concentration graph model $\sigma^{ij} = \rho_{ij|N\backslash\{ij\}} = 0$.

For instance, the three chordless graphs in Figure 1 define with the missing edges distinct sets of independence constraints for the same three pairs. The model for the parent graph in Figure 1(a) is arguably best known as a linear Markov model (Markov 1912). The linear covariance graph model for Figure 1(b) is a covariance chain, and the linear concentration

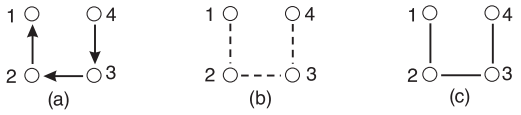

Figure 1. Three chain graphs respresenting distinct sets of independence constraints for the same three pairs: (a) a parent graph with $\rho_{14|23} = \rho_{13|24} = \rho_{24|3} = 0$; (b) a covariance graph with $\rho_{14} = \rho_{13} = \rho_{34} = 0$; (c) a concentration graph with $\rho_{14|23} = \rho_{13|24} = \rho_{24|13} = 0$.

graph model for Figure 1(c) is a concentration chain (1, 2, 3, 4). The graph in Figure 1(a) has attached to it the equations

$$Y_1 = \alpha Y_2 + \varepsilon_1, \qquad Y_2 = \gamma Y_3 + \varepsilon_2, \qquad Y_3 = \delta Y_4 + \varepsilon_3, \qquad Y_4 = \varepsilon_4,$$

where the residuals ε are uncorrelated so that the equation parameters are least-squares regression coefficients (Cramér 1946: 302). These equations also imply the concentration chain of Figure 1(c) as the structure in the concentration matrix; the parent graph is said to generate or induce the other type of graph. The covariance graph induced by the parent graph of Figure 1(a) is a complete graph, i.e. it has no missing edge.

A covariance matrix may be constrained in a more complex way by graphs other than independence graphs. A recursive regression graph, G^N_{rec}, is one of these types of graph. It is a parent graph combined with a residual covariance graph so that it contains two types of edge, arrows and dashed lines, and it may have two edges for a node pair. Recursive regression graphs have an attached system of linear equations with correlated residuals, which have been called recursive equations by Goldberger (1964), and form a subclass of structural equation models. Graphs attached to general structural equations have an independence interpretation (Koster 1999) but due to the correlated residuals there need not be a direct relation between an equation parameter and any linear least-squares regression coefficient.

We concentrate in this paper on constrained linear models for which stepwise data-generating processes can be specified by parent graphs which may include latent variables. This ensures that for each linear model considered here, densities of general type, generated over the same parent graph, satisfy corresponding probabilistic independence constraints.

1.3. Parameter equivalence and independence equivalence

There is parameter equivalence between two models if, except possibly in subset of lower dimension, there is a one-to-one correspondence between the sets of defining parameters. Then, given the first set of parameters, all parameters in the second set can be expressed uniquely in terms of those of the first, and vice versa, so that they imply in a linear model covariance matrices with the same constraints. Parameter equivalence can be exploited for estimation since it implies that maximum-likelihood estimates are in the same type of one-to-one correspondence (Fisher 1922: 327).

An example of parameter equivalence of a covariance chain and a linear triangular system is given in Section 6.1. A small non-trivial example of parameter equivalence

concerns the models for Figures 1(a) and 1(c), where the equivalence may be recognized from the orthogonal decomposition of Σ^{-1} (see Section 2.2).

Parameter equivalence of two constrained models implies that they have the same number of free parameters and that they are independence equivalent. The latter means that they have coinciding sets of independencies, though derived from different graphs. For linear models with independence graphs of one type of edge, such as in Figure 1, independence equivalence also implies parameter equivalence, provided the two graphs have coinciding sets of constrained variable pairs. We therefore summarize next a number of simple graphical criteria for independence equivalence in terms of V-configurations, i.e. of subgraphs induced by three nodes which have two edges.

An induced concentration graph is independence equivalent to its generating parent graph if and only if the parent graph does not contain the configuration

$$\circ \rightarrow \circ \leftarrow \circ$$

(see Wermuth 1980; Frydenberg 1990). Similarly, an induced covariance graph is independence equivalent to its generating parent graph (see Wermuth and Cox 2004) if and only if the parent graph does not contain any of the configurations

$$\circ \leftarrow \circ \leftarrow \circ, \qquad \circ \leftarrow \circ \rightarrow \circ.$$

These conditions also explain two related results on independence equivalence. A concentration graph G^N_{con} can be independence equivalent to a parent graph in node set N if and only if G^N_{con} is decomposable, i.e. it does not contain a chordless q-cycle of length $q \geqslant 4$ (Frydenberg 1990). The proof is that such a chordless cycle cannot be fully oriented, i.e. have all undirected edges changed into arrows, without creating either a cycle or a configuration $\circ \rightarrow \circ \leftarrow \circ$. A covariance graph G^N_{cov} can be independence equivalent to a parent graph in node set N if and only if G^N_{cov} does not contain a chordless chain of length $q \geqslant 4$ (Pearl and Wermuth 1994). The proof is that such a chordless chain cannot be fully oriented without creating at least one configuration of the type $\circ \leftarrow \circ \leftarrow \circ$, $\circ \leftarrow \circ \rightarrow \circ$.

This means, in particular, that there is no parent graph in node set N which is independence equivalent to a concentration graph with a chordless cycle of size larger than 3 or to a covariance graph with a chordless chain larger than 3. Thus, covariance matrices satisfying corresponding independencies cannot be directly generated by triangular systems in the observed variables, but can possibly be generated by a triangular system including in addition some unobserved variables (Cox and Wermuth 2000; Pearl and Wermuth 1994).

Hitherto, however, it has been a matter of conjecture that the two necessary conditions for parameter equivalence, namely independence equivalence and having the same number of parameters, taken together are generally not sufficient for parameter equivalence of two linear models. In Section 4 we settle this conjecture using models in which a missing edge in the associated graph need not represent a conditional independence statement. We illustrate further how triangular decompositions of covariance chains can be used to study the interpretation and equivalence of recursive regression graph models.

1.4. Estimation in concentration versus covariance graph models

Estimation in general Gaussian concentration graph models was introduced as covariance selection by Dempster (1972), studied further by Speed and Kiiveri (1986), and may require iterative calculations. The unique maximum-likelihood estimate of the constrained covariance matrix satisfies

$$\hat{\sigma}_{ij} = s_{ij} \text{ for } i = j \text{ or } i\text{—}j \text{ in } G^N_{\text{con}}, \qquad \text{and } \hat{\sigma}^{ij} = 0 \text{ otherwise,}$$

where s_{ij} is the observed covariance for Y_i and Y_j, an element of $\mathcal{S}$, the maximum-likelihood estimate of the covariance matrix of Y under the saturated model, i.e. when there are no constraints.

Gaussian covariance graph models have been studied as linear in covariance structures by Anderson (1969, 1973; Anderson and Olkin 1986). His maximum-likelihood equations can be written, with $\hat{\Sigma}^{-1}$ denoting the maximum-likelihood estimate of the concentration matrix, as

$$\hat{\sigma}^{ij} = [\hat{\Sigma}^{-1}\mathcal{S}\hat{\Sigma}^{-1}]_{i,j} \text{ for } i = j \text{ or } i \text{ --- } j \text{ in } G^N_{\text{cov}}, \qquad \text{and } \hat{\sigma}_{ij} = 0 \text{ otherwise.}$$

To solve these equations typically requires iterative calculations. A cyclic fitting procedure to compute maximum-likelihood estimates has been shown to converge to a local maximum (Drton & Richardson 2003). We derive a different form of the likelihood equations from more general results for exponential families in Section 6, to obtain explicit approximations to the maximum-likelihood estimates of Σ which are asymptotically efficient.

1.5. Outline of the paper

Section 2 gives previous results needed to establish the explicit properties of covariance chains given in Section 3. In Section 4 these results are applied to address the specific issues of interpretation outlined in the previous paragraphs. Section 5 discusses an approximation to maximum-likelihood estimates for constrained exponential families, and Section 6 applies these estimates to covariance chains. The paper concludes with a small simulation study to confirm the satisfactory behaviour of the estimates in moderate-size samples.

2. Notation and previous results

2.1. Interpretation of linear covariance graph models

For linear covariance graph models, independence statements additional to the defining ones may be obtained using the recursion relation for covariances (Anderson 1958: Section 2.5),

$$\sigma_{ij|C} = \sigma_{ij} - \sigma_{iC}\Sigma^{-1}_{CC}\sigma^{\mathrm{T}}_{jC}, \tag{1}$$

where σ_{iC}, σ_{jC} and Σ_{CC} are disjoint submatrices of the covariance matrix Σ. The covariance

in (1) is also called the partial covariance of Y_i and Y_j given Y_C, defined as the covariance of $Y_{i|C}$ and $Y_{j|C}$, where

$$Y_{i|C} = Y_i - \Pi_{i|C} Y_C$$

denotes variable Y_i after linear least-squares regression on Y_C. The parameters in this regression are the regression coefficient vector $\Pi_{i|C}$ and partial variances $\sigma_{ii|C}$, defined by

$$\Pi_{i|C} = \sigma_{iC}\Sigma_{CC}^{-1}, \qquad \sigma_{ii|C} = \sigma_{ii} - \Pi_{i|C}\sigma_{iC}^{\mathrm{T}}. \tag{2}$$

From equation (1) and because σ_{ij} is proportional to the simple correlation coefficient ρ_{ij}, it follows that two zero marginal covariances involving i imply a conditional linear independence statement

$$\sigma_{ij} = 0 \text{ and } \sigma_{iC} = 0 \text{ imply } \sigma_{ij|C} = 0. \tag{3}$$

2.2. Linear least-squares regressions and orthogonal decompositions

Linear least-squares regressions of Y_i on $Y_{r(i)}$ for $i = d-1, \ldots, 1$ with $r(i) = (i+1, \ldots, d)$ define a process of successive orthogonalization (Gram 1883; Schmidt 1907; Dempster 1969: Chapter 4), since a set of new variables is defined successively from Y_d, Y_{d-1}, ..., Y_1 such that each is orthogonal to the new variables obtained at previous steps. The new variables are the residuals ε_i in linear least-squares regressions of Y_i on $Y_{r(i)}$ and have diagonal covariance matrix $\Delta = \mathrm{cov}(\varepsilon)$.

This gives the following interpretation of the elements of an orthogonal decomposition (A, Δ^{-1}) of the concentration matrix, i.e. of the representation $A^{\mathrm{T}}\Delta^{-1}A = \Sigma^{-1}$. Here, the matrix A is upper-triangular and contains 1s along the diagonal. Off-diagonal elements in the vector $a_{i,r(i)}$ of A are minus least-squares regression coefficients, the diagonal elements of Δ the corresponding variances:

$$-a_{i,r(i)} = \Pi_{i|r(i)}, \qquad \delta_{ii} = \sigma_{ii|r(i)}.$$

There is a dual decomposition of Σ given by (B, Δ) with $B = A^{-1}$ and the representation $B\Delta B^{\mathrm{T}} = \Sigma$. Elements of B are linear least-squares coefficients obtained by marginalizing for each pair (i,j) over variables corresponding to nodes between i and j, called the intermediate nodes $i+1, \ldots, j-1$. To obtain this result, Cochran's (1938) recursion relation for regression coefficients is useful, where $\beta_{i|k.C}$ denotes the coefficient of Y_k in linear least-squares regression of Y_i on all variables with nodes listed after the conditioning sign, $|$,

$$\beta_{i|k.C} = \beta_{i|k.jC} + \beta_{i|j.kC}\beta_{j|k.C}. \tag{4}$$

Direct computation using $BA = I$ and (4) gives the elements in row i of B as

$$b_{ij} = \beta_{i|j.r(j)}. \tag{5}$$

Then both the covariance and the concentration matrix have an orthogonal basis in which

$$\Delta = A\Sigma A^{\mathrm{T}}, \qquad \Delta^{-1} = B^{\mathrm{T}}\Sigma^{-1}B.$$

For instance, for $d = 4$, the elements of the two triangular matrices A, B are

$$A = \begin{pmatrix} 1 & -\beta_{1|2.34} & -\beta_{1|3.24} & -\beta_{1|4.23} \\ 0 & 1 & -\beta_{2|3.4} & -\beta_{2|4.3} \\ 0 & 0 & 1 & -\beta_{3|4} \\ 0 & 0 & 0 & 1 \end{pmatrix}, \; B = \begin{pmatrix} 1 & \beta_{1|2.34} & \beta_{1|3.4} & \beta_{1|4} \\ 0 & 1 & \beta_{2|3.4} & \beta_{2|4} \\ 0 & 0 & 1 & \beta_{3|4} \\ 0 & 0 & 0 & 1 \end{pmatrix}. \tag{6}$$

Since a coefficient $\beta_{i|j.C}$ is a multiple of the partial correlation coefficient $\rho_{ij|C}$, the representation in equation (5), in essence given by Heywood (1931), and equations (4), (6) explain the interpretation and relations of zero constraints on elements of A and B.

2.3. Linear least-squares regression coefficients and concentrations

Linear least-squares regression coefficients have an interpretation in terms of concentrations (Dempster 1969: Section 6.4). We denote the elements in the concentration matrix of Y by σ^{ij} and a submatrix corresponding to component Y_s by $\Sigma^{ss} = [\Sigma^{-1}]_{s,s}$. After partitioning Y with $s \dot{\cup} v = \{1, \ldots, d\}$ we accordingly write

$$\begin{pmatrix} \Sigma^{ss} & \Sigma^{sv} \\ \Sigma^{vs} & \Sigma^{vv} \end{pmatrix} = \begin{pmatrix} \Sigma_{ss} & \Sigma_{sv} \\ \Sigma_{vs} & \Sigma_{vv} \end{pmatrix}^{-1}.$$

With $\Sigma^{ss}\Sigma_{sv} + \Sigma^{sv}\Sigma_{vv} = 0$ and using the matrix analogue of equation (2), this gives

$$\Pi_{s|v} = -(\Sigma^{ss})^{-1}\Sigma^{sv} = \Sigma_{sv}\Sigma_{vv}^{-1}. \tag{7}$$

With $\Sigma^{ss}\Sigma_{ss} + \Sigma^{sv}\Sigma_{vs} = I$ and using (7), we obtain

$$\Sigma_{ss|v} = (\Sigma^{ss})^{-1}. \tag{8}$$

Application of equation (8) to the partial covariance matrix of $Y_{i|C}$ and $Y_{j|C}$, with $s = \{i, j\}$ and $v = C$ denoting all variables except the pair (i, j), gives

$$\Sigma^{ss} = \begin{pmatrix} \sigma^{ii} & \sigma^{ij} \\ \cdot & \sigma^{jj} \end{pmatrix} = \begin{pmatrix} \sigma_{ii|C} & \sigma_{ij|C} \\ \cdot & \sigma_{jj|C} \end{pmatrix}^{-1},$$

where the . notation indicates here that in a symmetric matrix the (j,i)th element coincides with the (i,j)th element.

Therefore the elements of the overall concentration matrix can be written as

$$\frac{1}{\sigma^{ii}} = \sigma_{ii|jC}, \qquad -\frac{\sigma^{ij}}{\sigma^{ii}} = \beta_{i|j.C}. \tag{9}$$

This means, in particular, that a diagonal element σ^{ii} is a measure of precision for the linear least-squares regression of Y_i on all other variables. Also, the off-diagonal element σ^{ij} is zero if and only if $Y_{i|C}$ and $Y_{j|C}$ are linearly independent, i.e. when Y_i is linearly independent of Y_j given all remaining variables Y_C.

Zero constraints on elements of the regression coefficient matrix $\Pi_{u|v}$ have yet another independence interpretation, since its elements are $\beta_{i|j.v\setminus j}$. For instance, for $u = 1, 2$ and $v = 3, 4$, we have

$$\Pi_{u|v} = \begin{pmatrix} \beta_{1|3.4} & \beta_{1|4.3} \\ \beta_{2|3.4} & \beta_{2|4.3} \end{pmatrix}.$$

2.4. Partial variances and determinants of covariance matrices

Let $T = \Delta^{-1}A$ and $R(i) = (i, r(i))$. Then T is upper triangular and, analogously to equation (9), its elements in row i are such that

$$\frac{1}{t_{ii}} = \sigma_{ii|r(i)}, \qquad -\frac{t_{i,r(i)}}{t_{ii}} = \Pi_{i|r(i)}. \tag{10}$$

Thus, it contains in row i the precision and concentrations relating to Y_i when considering $Y_{R(i)}$ alone, i.e. row i of the concentration matrix $(\Sigma_{R(i),R(i)})^{-1}$. Elements in the first row of T are the overall precision σ^{11} and the overall concentrations σ^{1j}. Let D_s denote the determinant $|\Sigma_{ss}|$ of Σ_{ss} and $N = \{1, \ldots, d\}$. Then

$$|\Sigma| = D_N = \prod \sigma_{ii|r(i)}, \qquad |\Sigma_{ss|C}| = \frac{D_N}{D_C}, \qquad \sigma_{ii|r(i)} = \frac{D_{R(i)}}{D_{r(i)}}. \tag{11}$$

3. Parameters and generating processes for covariance chains

3.1. Constraints on partial regression coefficients and variances

For a covariance chain $(1, \ldots, d)$ the decomposition defined in Section 2.2 as $B\Delta B^{\mathrm{T}} = \Sigma$ implies that $b_{ik} = 0$ for $k > i + 1$ and from equation (5) that

$$b_{i,i+1} = \beta_{i|i+1.r(i+1)}. \tag{12}$$

Equation (1) implies that conditioning on other than neighbouring nodes leaves a covariance and a variance in the chain unchanged. In particular,

$$\sigma_{i,i+1|r(i+1)} = \sigma_{i,i+1}, \qquad \sigma_{ii|r(i+1)} = \sigma_{ii}. \tag{13}$$

For instance, the covariance matrix of Y_1, Y_2 given $Y_{r(2)}$ is, with (2),

$$\beta_{1|2.r(2)} = \frac{\sigma_{12}}{\sigma_{22|r(2)}}, \qquad \sigma_{11|r(1)} = \sigma_{11} - \frac{\sigma_{12}^2}{\sigma_{22|r(2)}}.$$

To obtain the other regression coefficients and variances, note that all elements in $\sigma_{i,r(i)}$ are zero except for the first. Then use the definition of parameters in linear least-squares regressions (2) as well as those of row $i + 1$ in the concentration matrix of $Y_{R(i+1)}$, (see (10)), where $R(i + 1) = (i + 1, r(i + 1)) = r(i)$, to give

$$\Pi_{i|r(i)} = \sigma_{i,i+1}\, t_{i+1,R(i+1)}, \quad \sigma_{ii|r(i)} = \sigma_{ii} - \frac{\sigma_{i,i+1}^2}{\sigma_{i+1,i+1|r(i+1)}}. \tag{14}$$

3.2. The orthogonal decomposition of the covariance matrix

The orthogonal decomposition of Σ for a covariance chain $(1, \ldots, d)$ inherits the simple chain structure in the following way. The elements of Δ are, by (14),

$$\delta_{dd} = \sigma_{dd},$$

$$\delta_{ii} = \sigma_{ii} - \frac{\sigma^2_{i,i+1}}{\sigma_{i+1,i+1|r(i+1)}} \qquad \text{for } i = 1, \ldots, d-1. \tag{15}$$

The elements of B are, by (14) and (15),

$$b_{i,i+1} = \frac{\sigma_{i,i+1}}{\delta_{i+1,i+1}} \qquad \text{for } i = 1, \ldots, d-1,$$

$$b_{ik} = 0, \qquad \text{for } k > i+1, \tag{16}$$

Example 1. For a covariance chain (1, 2, 3, 4) the orthogonal decomposition of Σ has the same simple chain structure with

$$B = \begin{pmatrix} 1 & \sigma_{12}/\delta_{22} & 0 & 0 \\ 0 & 1 & \sigma_{23}/\delta_{33} & 0 \\ 0 & 0 & 1 & \sigma_{34}/\delta_{44} \\ 0 & 0 & 0 & 1 \end{pmatrix},$$

and $\delta_{44} = \sigma_{44}$, $\delta_{33} = \sigma_{33} - \sigma^2_{34}/\sigma_{44}$, $\delta_{22} = \sigma_{22} - \sigma^2_{23}/\delta_{33}$ and $\delta_{11} = \sigma_{11} - \sigma^2_{12}/\delta_{22}$.

The decompositions can look more complex when the chain is written in a different sequence from $(1, \ldots, d)$. For instance, for the covariance chain (3, 1, 2, 4), the matrix Σ and the matrix B of its orthogonal decomposition (5) can be written with (11) as

$$\Sigma = \begin{pmatrix} D_1 & \sigma_{12} & \sigma_{13} & 0 \\ . & D_2 & 0 & \sigma_{24} \\ . & . & D_3 & 0 \\ . & . & . & D_4 \end{pmatrix}, \qquad B = \begin{pmatrix} 1 & \sigma_{12}D_4/D_{24} & \sigma_{13}/D_3 & 0 \\ 0 & 1 & 0 & \sigma_{24}/D_4 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix},$$

while for the covariance chain (2, 3, 4, 1) we obtain

$$\Sigma = \begin{pmatrix} D_1 & 0 & 0 & \sigma_{14} \\ . & D_2 & \sigma_{23} & 0 \\ . & . & D_3 & \sigma_{34} \\ . & . & . & D_4 \end{pmatrix}, \qquad B = \begin{pmatrix} 1 & \sigma_{14}\sigma_{23}\sigma_{34}/D_{234} & -\sigma_{14}\sigma_{34}/D_{34} & \sigma_{14}/D_4 \\ 0 & 1 & \sigma_{23}D_4/D_{34} & 0 \\ 0 & 0 & 1 & \sigma_{34}/D_4 \\ 0 & 0 & 0 & 1 \end{pmatrix}.$$

3.3. The concentration matrix

To obtain the concentration matrix induced by a covariance chain in nodes $R(i)$ we denote the nodes to the left of i by $l(i) = 1, \ldots, i-1$, the nodes to the right of i by $r(i)$, and use

the convention that $D_\varnothing = 1$. Then the elements ω_{ik} of the concentration matrix $\Omega(i)$ of $Y_{R(i)}$ can be written as

$$\begin{aligned} \omega_{ii} D_N &= D_{l(i)} D_{r(i)} && \text{for } i = 1, \ldots, d, \\ \omega_{i,i+1} D_N &= -D_{l(i)} \sigma_{i,i+1} D_{r(i+1)} && \text{for } i = 1, \ldots, d-1, \\ \omega_{ik} &= \frac{\omega_{ij}\omega_{jk}}{\omega_{jj}} && \text{for } k > j > i = 1, \ldots, d-2, \end{aligned} \tag{17}$$

where the last equation defines $\omega_{i,i+2}, \ldots$ recursively, using $\Sigma_{R(i),R(i)}\Omega(i) = I$.

Equation (17) leads to the following compact expression for the concentrations corresponding to zero covariances, denoting by $q = k - i$ the number of edges in a chain segment from node i to node k:

$$\omega_{ik} = (-1)^q D_{l(i)} \sigma_{i,i+1}\, \sigma_{i+1,i+2} \cdots \sigma_{k-1,k} \frac{D_{r(k)}}{D_{R(i)}}, \tag{18}$$

where, for example, with $R(i) = (i, r(i))$ the determinant $D_{R(i)}$ is computed as

$$D_{i,r(i)} = D_i D_{r(i)} - \sigma^2_{i,i+1} D_{r(i+1)}, \tag{19}$$

which is a recursion relation for determinants of tridiagonal symmetric matrices.

Example 2. For a covariance chain (1, 2, 3, 4) the covariance and the concentration matrix are

$$\Sigma = \begin{pmatrix} D_1 & \sigma_{12} & 0 & 0 \\ . & D_2 & \sigma_{23} & 0 \\ . & . & D_3 & \sigma_{34} \\ . & . & . & D_4 \end{pmatrix},$$

$$\Sigma^{-1} = D^{-1}_{1234} \begin{pmatrix} D_{234} & -\sigma_{12} D_{34} & \sigma_{12}\sigma_{23} D_4 & -\sigma_{12}\sigma_{23}\sigma_{34} \\ . & D_1 D_{34} & -D_1\sigma_{23} D_4 & D_1\sigma_{23}\sigma_{34} \\ . & . & D_{12} D_4 & -D_{12}\sigma_{34} \\ . & . & . & D_{123} \end{pmatrix}.$$

The determinant of Σ for this covariance chain is

$$D_{1234} = D_1 D_2 D_3 D_4 (1 - \rho^2_{12} - \rho^2_{23} - \rho^2_{34} + \rho^2_{12}\rho^2_{34}). \tag{20}$$

This can be used to measure the distance from an unrestricted covariance matrix, since it may be viewed as a residual generalized variance (Wilks 1932).

3.4. The orthogonal decomposition of the concentration matrix

The orthogonal decomposition of Σ^{-1} for a covariance chain $(1, \ldots, d)$ results as follows. The elements of A are

$$-a_{i,i+1} = b_{i,i+1} \qquad \text{for } i = 1, \ldots, d-1,$$

$$a_{ik} = a_{ij}a_{jk} \qquad \text{for } k > j > i, \tag{21}$$

where the last equation defines $a_{i,i+2}, \ldots,$ recursively. Each element a_{ik} of A for $k > i+1$ is, from (15)–(21) also a multiple of the covariances along the chain segment from node i to k having q edges

$$-a_{ik} = (-1)^q \frac{\sigma_{i,i+1}}{\delta_{i+1,i+1}} \frac{\sigma_{i+1,i+2}}{\delta_{i+2,i+2}} \cdots \frac{\sigma_{k-1,k}}{\delta_{kk}}. \tag{22}$$

Example 3. For a covariance chain (1, 2, 3, 4) the matrix elements of the orthogonal decomposition of the concentration matrix Σ^{-1} are, with (22) and (16),

$$A = \begin{pmatrix} 1 & -\beta_{1|2.34} & \beta_{1|2.34}\beta_{2|3.4} & -\beta_{1|2.34}\beta_{2|3.4}\beta_{3|4} \\ 0 & 1 & -\beta_{2|3.4} & \beta_{2|3.4}\beta_{3|4} \\ 0 & 0 & 1 & -\beta_{3|4} \\ 0 & 0 & 0 & 1 \end{pmatrix}.$$

3.5. Generating processes for covariance chains

A chain structure in a covariance matrix of residuals of Y may be generated by a linear triangular system for (Y, X) in which the latent variables X are standardized and orthogonal and where

$$HY + \Gamma X = \varepsilon_y, \qquad X = \varepsilon_x.$$

Here, the coefficient matrix H of Y is upper triangular and the elements of Γ are non-zero, $\gamma_{ij} \neq 0$, only for $i = j$ and $i = j+1$. Then, the linear equations in the observed variables Y are

$$HY = \eta,$$

with $\eta = \varepsilon_y - \Gamma X$, $\kappa = \mathrm{cov}(\eta) = \Delta_{yy} + \Gamma\Gamma^{\mathrm{T}}$, $\Sigma_{yy}^{-1} = H^{\mathrm{T}}\kappa^{-1}H$, and the elements of H have interpretations as least-squares regression coefficients given both observed and latent variables. These may or may not coincide with least-squares regression coefficients given only the observed variables.

The triangular decomposition (G, Δ^{-1}) of κ^{-1} relates to the triangular decomposition (A, Δ^{-1}) of Σ_{yy}^{-1} with

$$A = GH,$$

since $\Sigma_{yy}^{-1} = (H^{\mathrm{T}}G^{\mathrm{T}})\Delta^{-1}(GH) = H^{\mathrm{T}}\kappa^{-1}H$ and orthogonal decompositions are unique for a fixed ordering. This leads to an interpretation of $H = G^{-1}A$, in terms of least-squares regression coefficients of the observed variables contained in A (cf. (6)), and regression coefficients of residuals obtained with the triangular decomposition of the residual covariance matrix κ and contained in G^{-1}. This is exploited for the examples in Section 4.

When $H = I$, a covariance chain results for the observed variables, i.e. directly in Σ_{yy}. For instance, the covariance chain (1, 2, 3, 4) has as a generating parent graph

$$G^{N,L}_{\text{par}}\colon 1 \leftarrow \not\emptyset \rightarrow 2 \leftarrow \not\emptyset \rightarrow 3 \leftarrow \not\emptyset \rightarrow 4,$$

where $\not\emptyset$ denotes hidden variables.

4. Parameter and independence equivalence

4.1. Independence equivalence but no parameter equivalence

We can now consider the recursive regression graph

$$G^{N}_{\text{rec}}\colon 1 \underset{---}{\leftarrow} 2 \underset{---}{\leftarrow} 3 \underset{---}{\leftarrow} 4$$

to prove that the two necessary conditions for parameter equivalence, an equal number of parameters and independence equivalence, taken together are not yet sufficient for parameter equivalence. As explained below, this recursive regression graph model has ten parameters and imposes no independence constraint, just like the saturated model, but it is nevertheless not parameter equivalent to the saturated model.

The matrix H of equation parameters for G^{N}_{rec} is obtained directly from the graph, and the matrix G^{-1} by the triangular decomposition of the residual covariance chain of Section 3.2:

$$H = \begin{pmatrix} 1 & -\alpha & 0 & 0 \\ 0 & 1 & -\gamma & 0 \\ 0 & 0 & 1 & -\delta \\ 0 & 0 & 0 & 1 \end{pmatrix}, \qquad G^{-1} = \begin{pmatrix} 1 & \phi & 0 & 0 \\ 0 & 1 & \psi & 0 \\ 0 & 0 & 1 & \xi \\ 0 & 0 & 0 & 1 \end{pmatrix}.$$

The matrices $B = H^{-1}G^{-1}$ and $A = GH$ of the triangular decompositions of Σ and Σ^{-1} are then

$$B = \begin{pmatrix} 1 & \alpha+\phi & \alpha(\gamma+\psi) & \alpha\gamma(\delta+\xi) \\ 0 & 1 & \gamma+\psi & \gamma(\delta+\xi) \\ 0 & 0 & 1 & \delta+\xi \\ 0 & 0 & 0 & 1 \end{pmatrix},$$

$$A = \begin{pmatrix} 1 & -(\alpha+\phi) & \phi(\gamma+\psi) & -\phi\psi(\delta+\xi) \\ 0 & 1 & -(\gamma+\psi) & \psi(\delta+\xi) \\ 0 & 0 & 1 & -(\delta+\xi) \\ 0 & 0 & 0 & 1 \end{pmatrix}.$$

These explicit forms make it transparent that the transformation from the parameters in H and G^{-1} to the linear least-squares coefficients in A or in B is for this model non-invertible: the confounding of $\beta_{3|4} = \delta + \xi$ cannot be resolved, i.e. not all elements of H, κ can be obtained given Σ. The recursive regression graph model has three equation parameters in H, and three residual covariances and four residual variances in κ. The

parameters are from A and B such that every marginal and every type of partial correlation is non-zero for each variable pair Y_i, Y_j. Therefore, the recursive regression graph model is independence but not parameter equivalent to the unconstrained covariance matrix with ten parameters. Thus, two models may have the same number of parameters and generate the same independencies but not be equivalent, since one implies an additional constraint, which in the above example does not correspond to an independence. The explicit form of the matrix B and the interpretation of its elements from (5) show that $\alpha = \beta_{1|3.4}/\beta_{2|3.4} = \beta_{1|4}/\beta_{2|4}$ may have two distinct representations and that $\gamma = \beta_{2|4}/\beta_{3|4}$. The non-unique representation poses a potential problem for instrumental variable estimation (Sargan 1958) of some of the dependencies.

4.2. Parameter equivalence in spite of confounding

The following case establishes parameter equivalence of a recursive regression graph with parameters (H, κ) to the triangular decomposition (B, Δ) of Σ with an independence constraint and a confounded least-squares regression coefficient. From the graph of Figure 2, the matrices H and G^{-1} and $B = H^{-1}G^{-1}$ are given by

$$H = \begin{pmatrix} 1 & -\alpha & 0 & -\gamma \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix}, \quad G^{-1} = \begin{pmatrix} 1 & \phi & 0 & 0 \\ 0 & 1 & \psi & 0 \\ 0 & 0 & 1 & \xi \\ 0 & 0 & 0 & 1 \end{pmatrix}, \quad B = \begin{pmatrix} 1 & \alpha+\phi & \alpha\psi & \gamma \\ 0 & 1 & \psi & 0 \\ 0 & 0 & 1 & \xi \\ 0 & 0 & 0 & 1 \end{pmatrix}.$$

The zero entry in B means $\beta_{2|4} = \rho_{24} = 0$. Furthermore, reversibility of the transformation results from the explicit form of B with $\gamma = \beta_{1|4}$, $\xi = \beta_{3|4}$, $\psi = \beta_{2|3.4}$, $\alpha = \beta_{1|3.4}/\beta_{2|3.4}$, and $\phi = \beta_{1|2.34} - \alpha$.

Here, by contrast with Section 4.1, we have two models that generate the same independence structure and that are parameter equivalent. Because in one of the models unique and simple estimation by least-squares is available, estimates in the other, seemingly more complex model, can be obtained directly due to the one-to-one correspondence of the sets of parameters. This type of parameter equivalence supplements results on identifiability of parameters in which there is renewed interest related to recursive regression graph models; see McDonald (2002), Pearl and Brito (2002) and Stanghellini and Wermuth (2005).

Figure 2. (a) A recursive regression graph and (b) an independence equivalent covariance graph.

4.3. The parameter constraint in an independence equivalent model

To have the same number of parameters is not a necessary condition for independence equivalence. For instance, the single factor analysis model generates an observed covariance

matrix constrained to be the sum of a rank-one and a diagonal matrix, but which is independence equivalent to a saturated model. We show here how the specific constraint by which two independence equivalent recursive regression graph models may differ, results directly from the triangular decomposition of the residual covariance matrix.

The graph in Figure 3(a) is an example of what Richardson and Spirtes (2002) call a maximal ancestral graph, and the graph in Figure 3(b) is a corresponding independence equivalent ancestral graph in which the missing edge does not correspond to any independence statement. Our form of the constraint shows that it captures the contribution of the residual covariance chain from node 1 to 2 via the parent nodes 3, 4. The authors suggest moving from ancestral graphs to corresponding maximal ancestral graphs so that graphs are obtained in which at least one independence statement is associated with every missing edge.

The two recursive regression graphs in Figure 3 coincide in the edges within $v = \{2, 3, 4\}$; therefore they have the same submatrices Σ_{vv}. By the symmetry of the graphs, nodes (1, 3) may be exchanged with (2, 4) and therefore the submatrices Σ_{uu} for $u = \{1, 3, 4\}$ coincide as well. As the constraint implied by the graph of Figure 3(b) we obtain from $\beta_{1|2.34}$ in the covariance chain (2, 3, 4, 1) in Example 1 of Section 3.2 that

$$\sigma_{12|34} = \kappa_{14}\kappa_{23}\kappa_{34}/D_{34},$$

where D now refers to the determinant of a submatrix of κ. The model with six edges in Figure 3(a) implies no constraints since it is parameter equivalent to the saturated model. Thus two independence equivalent models may differ appreciably with respect to one parameter as may be shown from a triangular decomposition of the residual covariance matrix.

Figure 3. Two independence equivalent recursive regression graphs with coinciding matrices Σ_{vv} for $v = \{2, 3, 4\}$ and Σ_{uu} for $u = \{1, 3, 4\}$, but different non-zero partial covariances for the pair (1,2) given (3,4): (a) Σ unconstrained; (b) Σ constrained by $\sigma_{12|34} = \kappa_{14}\kappa_{23}\kappa_{34}/D_{34}$.

5. Estimation of moment parameters of exponential families

5.1. The general case

If the observed random variable Y has a multivariate Gaussian distribution the estimation of Σ in a covariance chain model requires estimation of a covariance matrix with some elements constrained to be zero. We suppose without essential loss of generality that $E(Y) = 0$ and that n independent and identically distributed observations are available.

It is simplest to deal with the corresponding more general problem for a full exponential family, taking the log-likelihood in the form

$$n\{s^{\mathrm{T}}\phi - k(\phi)\},$$

where ϕ is the canonical parameter and s the sufficient statistic. The mean parameter η, a one-to-one function of ϕ, is defined by $\eta = \nabla k(\phi)$, where ∇ denotes a gradient, i.e. the vector of first derivatives, with respect to ϕ.

The maximum-likelihood estimate of η is $\hat{\eta} = s$, a possible value of the corresponding random variable S, and

$$n\,\mathrm{cov}(S) = \nabla\nabla^{\mathrm{T}} k(\phi) = i(\eta),$$

where $\nabla\nabla^T k(\phi)$ is the matrix of second derivatives of $k(\phi)$ with respect to ϕ. It is convenient to express this matrix as a function $i(\eta)$ of the mean parameter and to note that it is a gradient of η with respect to ϕ, i.e. $\nabla(\eta) = i(\eta)$. It can be shown that not only does $i(\eta)$ determine the covariance matrix of S but also, as the so-called Fisher information matrix for estimating ϕ, it gives the asymptotic concentration matrix of the maximum-likelihood estimate of the canonical parameter.

We now introduce the constraints that $\eta_c = 0$ for some subset of elements of η, writing $\eta = (\eta_u, \eta_c)$, and consider the Lagrangian

$$s^{\mathrm{T}}\phi - k(\phi) - \lambda^{\mathrm{T}}\eta_c.$$

Differentiating with respect to ϕ gives the maximum-likelihood estimating equations as

$$\hat{\eta}_u = s_u - \hat{i}_{uc}\hat{i}_{cc}^{-1}s_c, \qquad \hat{\eta}_c = 0. \tag{23}$$

The (u, c) and (c, c) components of the matrix $i(\eta)$ are evaluated at the new maximum-likelihood estimate for which $\hat{\eta}_c = 0$. Usually iterative solution is required.

However, because in the second term s_c is $O_p(1/\sqrt{n})$, an asymptotically efficient estimate is obtained by replacing the matrices pre-multiplying s_c by any consistent estimate of them. That is, s_c differs from zero only by a random error of estimation, small for large n. The multiplying coefficients then do not have to be determined with high precision. For this the simplest procedure is to use the relevant portions of the matrix $i(s_u, 0)$, i.e. the information matrix evaluated at the unconstrained estimate s_u of η_u and at zero values of the constrained parameter η_c. The resulting explicit estimates are

$$\tilde{\eta}_u = s_u - \tilde{i}_{uc}\tilde{i}_{cc}^{-1}s_c, \qquad \tilde{\eta}_c = 0. \tag{24}$$

In a general context such estimates were called reduced model estimates by Cox and Wermuth (1990). There is a close connection with the Lagrange multiplier tests of Aitchison and Silvey (1958), here much simplified by the exponential family structure.

The adjustment to s_u in (24) means that the new estimate $\hat{\eta}_u$ is asymptotically orthogonal to s_c or, expressed differently, is obtained by adjusting s_u for linear regression on s_c. The attractive feature of this representation is that it shows for each parameter the precise modification of the corresponding unconstrained estimate. Though the explicit reduced model estimates of constrained moment parameters (24) for exponential families are asymptotically efficient, they are recommended only for observations which strongly support the reduced model.

5.2. The Gaussian case

Now consider the special case of a Gaussian distribution in which the moment parameter is σ, a vector obtained from the covariance matrix Σ by taking the variances σ_{ii} and the covariances σ_{ij} considered only once. Elements of a corresponding observed vector s and of a random variable S are

$$s_{ij} = \sum_t y_{ti} y_{tj}/n, \qquad S_{ij} = \sum_t Y_{ti} Y_{tj}/n.$$

Note that if an unknown mean were included in the model, the sum of products would be replaced by a sum of products of deviations from the sample mean; some of the following formulae would then be approximations, essentially from replacing $n-1$ by n.

Now because of the assumed Gaussian distribution

$$n \operatorname{cov}(S)_{ij;kl} = \sigma_{ik}\sigma_{jl} + \sigma_{il}\sigma_{jk},$$

the right-hand side being called the Isserlis matrix, after Isserlis (1918); see, for example, Roverato and Whittaker (1998). This Isserlis matrix, $\mathrm{Iss}(\sigma)$, of covariances has for q variables the determinant $2^q|\Sigma|^{q-1}$ (Press 1972: 79). Therefore, if the covariance matrix Σ is positive definite, so is the Isserlis matrix.

Equation (23) specializes to the maximum-likelihood equations for $\hat{\sigma}_u$ given that $\sigma_c = 0$ as

$$\hat{\sigma}_u = s_u - \hat{\mathrm{I}}\mathrm{ss}_{uc}\hat{\mathrm{I}}\mathrm{ss}_{cc}^{-1} s_c, \qquad \hat{\sigma}_c = 0, \tag{25}$$

where for $\hat{\mathrm{I}}\mathrm{ss}$ we first evaluate the matrix $\mathrm{Iss}(\sigma)$ at $\sigma_{\mathrm{red}} = (\sigma_u, 0)$ and then replace σ_u by $\hat{\sigma}_u$. Similarly, equation (24) specializes to the explicit estimate $\tilde{\sigma}_u$ given that $\sigma_c = 0$ as

$$\tilde{\sigma}_u = s_u - \tilde{\mathrm{I}}\mathrm{ss}_{uc}\tilde{\mathrm{I}}\mathrm{ss}_{cc}^{-1} s_c, \qquad \tilde{\sigma}_c = 0, \tag{26}$$

where for $\tilde{\mathrm{I}}\mathrm{ss}$ we first evaluate $\mathrm{Iss}(\sigma)$ again at $\sigma_{\mathrm{red}} = (\sigma_u, 0)$ but then replace σ_u by s_u.

The non-zero estimates of σ_u are thus obtained by an adjustment to the standard unconstrained estimates s_u, the adjustment having the form of a typically small correction for regression on s_c. Provided the sample size is large and the assumed model is correct, s_c differs from zero only by small sampling fluctuations. It can be shown that after two steps of the iterative algorithm suggested by Anderson (1973), to solve his maximum-likelihood equations in Section 1.4, one also obtains the reduced model estimate of equation (26). Kauermann (1996) has applied to Gaussian covariance graph models a method which maximizes the dual likelihood function in exponential families (Christensen 1989). This gives estimates which tend to underestimate some of the variances.

The notion of a regression-based correction has direct appeal without strong distributional assumptions. The matrix of regression coefficients used in the method does, however, involve a theoretical calculation based initially on a Gaussian distribution of the original observations. A generalization of the Isserlis matrix to arbitrary distributions (McCullagh 1987: 118) involves fourth cumulants. Thus the method has a strong justification if the fourth cumulants are close to zero or if the effect of non-Gaussian form is merely to inflate the whole covariance matrix of S by a constant factor.

While we have not explored the issue in detail, it seems likely that some forms of departure, for example long-tailed marginal distributions for particular components, would inflate the diagonal elements of cov(S) more strongly than the off-diagonal elements and that this would tend to reduce the magnitude of the regression correction. Since the uncorrected estimate is already often of quite high efficiency, this suggests that the assumption of multivariate Gaussian form may not be critical.

6. Applications to covariance chains

6.1. The covariance chain of length 3

For the covariance chain (1, 2, 3) the independence constraint is $\rho_{13} = 0$ and the moment parameters may be written in vector form as

$$\sigma_u = (\sigma_{11}, \sigma_{22}, \sigma_{33}, \sigma_{12}, \sigma_{23}), \qquad \sigma_c = (\sigma_{13}) = 0.$$

The Isserlis matrix for the constrained model given $\sigma_{\text{red}} = (\sigma_u, 0)$, with elements ordered as above, is

$$\text{Iss}(\sigma_{\text{red}}) = \begin{pmatrix} 2\sigma_{11}^2 & 2\sigma_{12}^2 & 0 & 2\sigma_{12}\sigma_{11} & 0 & 0 \\ . & 2\sigma_{22}^2 & 2\sigma_{23}^2 & 2\sigma_{12}\sigma_{22} & 2\sigma_{23}\sigma_{22} & 2\sigma_{12}\sigma_{23} \\ . & . & 2\sigma_{33}^2 & 0 & 2\sigma_{23}\sigma_{33} & 0 \\ . & . & . & \sigma_{11}\sigma_{22} + \sigma_{12}^2 & \sigma_{12}\sigma_{23} & \sigma_{23}\sigma_{11} \\ . & . & . & . & \sigma_{22}\sigma_{33} + \sigma_{23}^2 & \sigma_{12}\sigma_{33} \\ . & . & . & . & . & \sigma_{11}\sigma_{33} \end{pmatrix}.$$

From the last column of this matrix equations (25) become $\hat{\sigma}_{11} = s_{11}$, $\hat{\sigma}_{33} = s_{33}$ and

$$\hat{\sigma}_{22} = s_{22} - 2\frac{\hat{\sigma}_{12}\hat{\sigma}_{23}}{\hat{\sigma}_{11}\hat{\sigma}_{33}}s_{13}, \qquad \hat{\sigma}_{12} = s_{12} - \frac{\hat{\sigma}_{23}}{\hat{\sigma}_{33}}s_{13}, \qquad \hat{\sigma}_{23} = s_{23} - \frac{\hat{\sigma}_{12}}{\hat{\sigma}_{11}}s_{13}.$$

The three equations can be solved to give the maximum-likelihood estimates

$$\hat{\sigma}_{12} = \hat{\beta}_{2|1.3}s_{11}, \qquad \hat{\sigma}_{23} = \hat{\beta}_{2|3.1}s_{33}, \qquad \hat{\sigma}_{22} = s_{22} - 2\hat{\beta}_{2|1.3}\hat{\beta}_{2|3.1}s_{13}, \tag{27}$$

where $\hat{\beta}_{i|j.k} = s_{ij|k}/s_{jj|k} = -s^{ij}/s^{ii}$ denotes the least-squares estimate of $\beta_{i|j.k}$.

Alternatively, the triangular decomposition $(\hat{A}, \hat{\Delta}^{-1})$, defined as in Section 2.2 for the estimated concentration matrix given in the ordering (2, 1, 3), can be used to obtain the same estimates

$$\hat{A} = \begin{pmatrix} 1 & -\hat{\beta}_{2|1.3} & -\hat{\beta}_{2|3.1} \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix}, \qquad \hat{\Delta} = \begin{pmatrix} s_{22|13} & 0 & 0 \\ . & s_{11} & 0 \\ . & . & s_{33} \end{pmatrix}.$$

The estimates in (27) are a one-to-one transformation of the maximum-likelihood

estimates for a Gaussian distribution of Y in the corresponding system of linear regression equations given by

$$Y_2 = \beta_{2|1.3} Y_1 + \beta_{2|3.1} Y_3 + \varepsilon_2, \qquad Y_1 = \varepsilon_1, \qquad Y_3 = \varepsilon_3, \tag{28}$$

with diagonal residual covariance matrix $\mathrm{cov}(\varepsilon) = \Delta$.

The linear least-squares estimates for the parameters in (28) are written compactly in terms of concentrations by using (9) as

$$\hat{\beta}_{2|1.3} = -\frac{s^{12}}{s^{22}}, \qquad \hat{\beta}_{2|3.1} = -\frac{s^{23}}{s^{22}}, \qquad \hat{\sigma}_{22.13} = \frac{1}{s^{22}}, \qquad \hat{\sigma}_{11} = s_{11}, \qquad \hat{\sigma}_{33} = s_{33}. \tag{29}$$

The covariance chain (1, 2, 3) and this system of equations both specify marginal orthogonality for the pair Y_1, Y_3, i.e. $\rho_{13} = 0$. The corresponding independence equivalent graphs are

$$G^N_{\mathrm{cov}}\colon 1 \text{ - - - } 2 \text{ - - - } 3, \qquad G^N_{\mathrm{par}}\colon 1 \to 2 \leftarrow 3.$$

6.2. The covariance chain of length 4

For the covariance chain (1, 2, 3, 4), the independence constraints are the three zero marginal correlations $\rho_{13} = \rho_{14} = \rho_{24} = 0$ and there are two implied constraints, (see (3)), on partial correlations: $\rho_{24|1} = \rho_{13|4} = 0$. The chain has no independence equivalent parent graph in the observed nodes and the moment parameters may be written in vector form as

$$\sigma_u = (\sigma_{11}, \sigma_{22}, \sigma_{33}, \sigma_{44}, \sigma_{12}, \sigma_{23}, \sigma_{34}), \qquad \sigma_c = (\sigma_{13}, \sigma_{14}, \sigma_{24}) = 0.$$

The maximum-likelihood equations (25) do not appear to have a explicit solution, but the reduced model estimates of (26) can be written as $\tilde{\sigma}_{11} = s_{11}$, $\tilde{\sigma}_{44} = s_{44}$ and

$$\begin{aligned}
\tilde{\sigma}_{22} &= s_{22} - 2 s_{23} \hat{\beta}_{2|1} \hat{\beta}_{1|3.4}, \\
\tilde{\sigma}_{33} &= s_{33} - 2 s_{23} \hat{\beta}_{3|4} \hat{\beta}_{4|2.1}, \\
\tilde{\sigma}_{12} &= s_{12} - s_{23} \hat{\beta}_{1|3.4}, \\
\tilde{\sigma}_{23} &= s_{23} - \hat{\beta}_{3|4}\, s_{24} - \hat{\beta}_{2|1}\, s_{13} + \hat{\beta}_{3|4} \hat{\beta}_{2|1}\, s_{14}, \\
\tilde{\sigma}_{34} &= s_{34} - s_{23} \hat{\beta}_{4|2.1}.
\end{aligned} \tag{30}$$

When the data are a sample of moderate size from a covariance chain (1, 2, 3, 4), then the constrained observed correlations will be close to zero, as well as s_{13}, s_{14}, s_{24}, $\hat{\beta}_{1|3.4}$ and $\hat{\beta}_{4|2.1}$. Thus, the above corrections to the observed second moments will be small.

There is the distinction between chains of length 3 and chains of length greater than 3 in that the former, but not the latter, have explicit expressions for maximum-likelihood estimates in terms of least-squares regression coefficients. The explanation is that only the former model is parameter equivalent to a parent graph model for which the regression

coefficients arise directly. But for both types of estimates, the estimated correlation structure remains unchanged when the units of measurement are changed for some of the variables.

6.3. Some simulation results

The following simulation results show that the estimates in (30) give a close approximation to the population parameters also in the case of moderate-size samples from a covariance chain (1, 2, 3, 4) for a Gaussian distribution. The log-likelihood ratio, evaluated at estimates of the unconstrained, i.e. saturated, covariance matrix Σ_{sat} and of the constrained covariance matrix having a reduced number of parameters Σ_{red}, provide a goodness-of-fit statistic. For q variables and k constraints these statistics are, for the two types of estimates,

$$-n \log (|\hat{\Sigma}_{\text{sat}}|/|\hat{\Sigma}_{\text{red}}|), \qquad -n \{\log (|\hat{\Sigma}_{\text{sat}}|/|\tilde{\Sigma}_{\text{red}}|) + \text{trace}(\hat{\Sigma}_{\text{sat}}\tilde{\Sigma}_{\text{red}}^{-1}) - q\}.$$

where the first has for large n approximately a central chi-squared distribution with k degrees of freedom.

The curves in Figure 4 show how closely the smoothed observed distributions approximate the corresponding chi-squared distribution for moderate sample sizes. The population covariance chain chosen for these simulations is

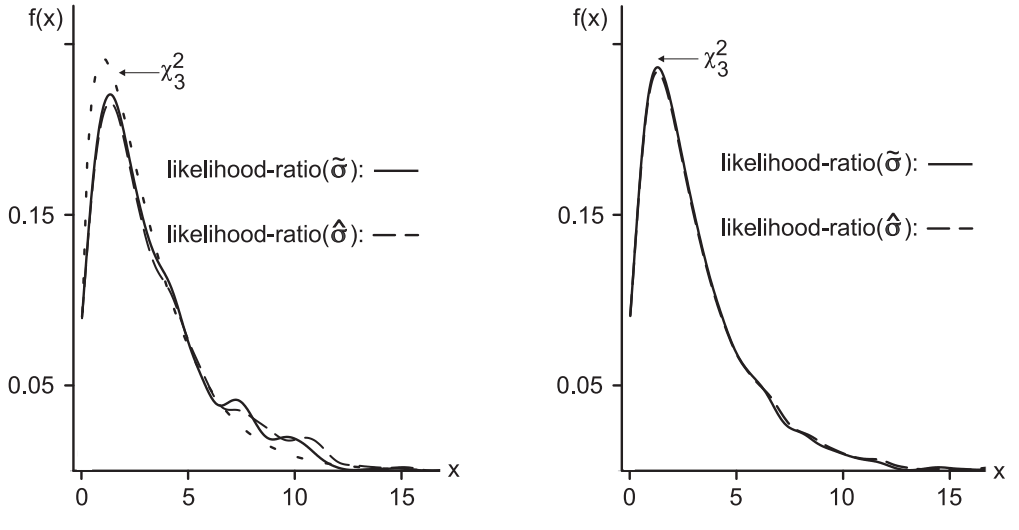

Figure 4. Density plots of the log-likelihood ratio for the maximum-likelihood estimate (dashed) and for the reduced model estimates (solid); chi-squared density with 3 degrees of freedom (dotted); 1000 draws each for a sample size n= *50* (left) and for n= *100* (right).

$$\Sigma = \begin{pmatrix} 1 & \alpha & 0 & 0 \\ . & 1+\alpha^2+\gamma^2 & \gamma\theta & 0 \\ . & . & 1+\beta^2+\theta^2 & \beta \\ . & . & . & 1 \end{pmatrix} = \begin{pmatrix} 1 & 0.5 & 0 & 0 \\ . & 2.25 & -3 & 0 \\ . & . & 10.64 & -0.8 \\ . & . & . & 1 \end{pmatrix}.$$

In addition, equally correlated variables ($\rho = 0.5$) having equal variances ($\sigma_{ii} = 2$) were chosen to exemplify strong deviations from a covariance chain.

Table 1 contains some detailed results for 1000 draws at sample size $n = 50$ (first row) and $n = 100$ (second row). The simulations started with seed $-26\,634$, using the graphical Gausssian model package of R (Marchetti 2006). The identity matrix was used as the starting value for the iterative conditional fitting algorithm.

The differences between maximum likelihood estimates, satisfying (25), and the explicit reduced model estimates, obtained with (26), are small compared with the standard error of estimation when the draws are from covariance chains (left-hand part of the table). For population parameters far away from a covariance chain this is no longer the case (right-hand part of the table), illustrating a quite general feature of such iterative procedures applied to ill-fitting models. In this case about 0.5% of the matrices $\tilde{\Sigma}$ are not positive definite.

Unimodality of the likelihood was established by finding just one real root to equations (25); for a related discussion, see Drton and Richardson (2004). Since the likelihood function turned out to be unimodal in all 2000 samples, the maximum-likelihood estimates agree with those by the E(xpectation) M(aximization) algorithm, as specialized by Kiiveri (1987) to path analysis with some variables unobserved. The simulation results suggest that even in moderate-sized samples the explicit form estimates differ from the population

Table 1. Comparison of maximum-likelihood estimates $\hat{\sigma}$ and the explicit reduced model estimates $\tilde{\sigma}$ and for sample size $n = 50$ (first row) and $n = 100$ (second row); 1000 draws from two models

	Covariance chain model						Equal-correlation model				
	$\hat{\sigma}$		$\tilde{\sigma}$				$\hat{\sigma}$		$\tilde{\sigma}$		
	Mean	St.d.	Mean	St.d.	rms*		Mean	St.d.	Mean	St.d.	rms*
$\sigma_{22} = 2.25$	2.19	0.42	2.15	0.42	0.08	$\sigma_{22} = 2$	1.89	0.38	1.62	0.33	0.35
	2.22	0.32	2.20	0.31	0.04		1.87	0.26	1.63	0.23	0.28
$\sigma_{33} = 10.64$	10.31	2.06	10.10	2.04	0.38	$\sigma_{33} = 2$	1.88	0.38	1.62	0.34	0.35
	10.51	1.45	10.40	1.42	0.20		1.89	0.27	1.65	0.24	0.28
$\sigma_{12} = 0.50$	0.49	0.18	0.48	0.17	0.02	$\sigma_{12} = 1$	0.86	0.33	0.64	0.26	0.28
	0.50	0.13	0.49	0.13	0.01		0.87	0.23	0.66	0.19	0.24
$\sigma_{23} = -3.00$	−2.92	0.76	−2.81	0.75	0.16	$\sigma_{23} = 1$	0.31	0.26	0.24	0.20	0.11
	−2.96	0.55	−2.90	0.54	0.09		0.31	0.18	0.25	0.14	0.08
$\sigma_{34} = -0.80$	−0.79	0.37	−0.77	0.37	0.05	$\sigma_{34} = 1$	0.86	0.33	0.64	0.27	0.28
	−0.79	0.26	−0.78	0.26	0.02		0.88	0.23	0.66	0.19	0.25

*rms: root mean square difference between the two estimates.

parameters by only small sampling fluctuations, provided the constraints are closely reflected in the observed correlations.

Acknowledgement

The first author thanks the German Research Society and the University of Mainz for a research grant, also the Swedish Research Society for supporting her cooperation with D.R. Cox. Computations were carried out using MATLAB and the ggm package for R available at http://cran.r-project.org/

References

Aitchison, J. and Silvey, S.D. (1958) Maximum-likelihood estimation of parameters subject to restraints. *Ann. Math. Statist*. **29**, 813–828.

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

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

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

Anderson, T.W. and Olkin, I. (1986) Maximum-likelihood estimation of the parameters of a multivariate normal distribution. *Linear Algebra Appl.*, **70**, 147–171.

Christensen, S. (1989) Statistical properties of I-projections within exponential families. *Scand. J. Statist.*, **16**, 307–318.

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

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. (1993) Linear dependencies represented by chain graphs (with discussion). *Statist. Sci.*, **8**, 204–218; 247–277.

Cox, D.R. and Wermuth, N. (2000) On the generation of the chordless 4-cycle. *Biometrika*, **87**, 206–212.

Cramér, H. (1946) *Mathematical Methods of Statistics.* Princeton, NJ: Princeton University Press.

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

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

Drton, M. and Richardson, T.S. (2003) A new algorithm for maximum-likelihood estimation in Gaussian graphical models for marginal independence. In U. Kjærulff and C. Meek (eds), *Uncertainty in Artificial Intelligence: Proceedings of the Nineteenth Conference*, pp. 181–191.

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

Fisher, R.A. (1922) On the mathematical foundations of theoretical statistics. *Philos. Trans. Roy. Soc. Lond. Ser. A*, **222**, 309–368.

Frydenberg, M. (1990) The chain graph Markov property. *Scand. J. Statist.*, **17**, 333–353.

Goldberger, A.S. (1964) *Econometric Theory.* New York: Wiley.

Gram, J.P. (1883) Über die Entwicklung reeller Funktionen in Reihen mittelst der Methode der kleinsten Quadrate. *J. Reine Angew. Math.*, **94**, 41–73.

Heywood, H.B. (1931) On finite sequences of real numbers. *Proc. Roy. Soc. Lond. Ser. A*, **134**, 486–501.

Isserlis, L. (1918) Formulae for determining the near values of products of deviations of mixed moment coefficients. *Biometrika*, **12**, 183–184.

Kauermann, G. (1996) On a dualization of graphical Gaussian models. *Scand. J. Statist.*, **23**, 105–116.

Kiiveri, H.T. (1987) An incomplete data approach to the analysis of covariance structures. *Psychometrika*, **52**, 539–554.

Koster, J. (1999) On the validity of the Markov interpretation of path diagrams of Gaussian structural equation systems of simultaneous equations. *Scand. J. Statist.*, **26**, 413–431.

Marchetti, G.M. (2006) Independencies induced from a graphical Markov model after marginalization and conditioning: the R package ggm. *J. Statist. Software*, **15**(6).

Markov, A.A. (1912) *Wahrscheinlichkeitsrechnung* (German translation of 2nd Russian edition). Leipzig: Teubner.

McDonald, R.P. (2002) What can we learn from the path equations? Identifiability, constraints, equivalence. *Psychometrika*, **67**, 225–249.

McCullagh, P. (1987) *Tensor Methods in Statistics.* London: Chapman & Hall.

Pearl, J. and Brito, C. (2002) A new identification condition for recursive models with correlated errors. *Structural Equation Modeling*, **9**, 459–474.

Pearl, J. and Wermuth, N. (1994) When can association graphs admit a causal interpretation? In P. Cheeseman and W. Oldford (eds.), Selecting Models from Data: Artificial Intelligence and Statistics IV, Lecture Notes in Statist. 89, pp. 205–214. New York: Springer-Verlag.

Press, S.J. (1972) *Applied Multivariate Analysis: Using Bayesian and Frequentist Methods of Inference.* New York: Holt, Rinehart and Winston.

Richardson, T.S. and Spirtes, P. (2002) Ancestral Markov graphical models. *Ann. Statist.*, **30**, 962–1030.

Roverato, A. and Whittaker, J. (1998) The Isserlis matrix and its application to non-decomposable graphical models. *Biometrika*, **85**, 711–725.

Sargan, J.D. (1958) The estimation of economic relationships using instrumental variables. *Econometrica*, **26**, 393–415.

Schmidt, E. (1907) Zur Theorie der linearen und nichtlinearen Integralgleichungen. I. Teil: Entwicklungen willkürlicher Funktionen nach Systemen vorgeschriebener, *Math. Ann.*, **63**, 433–476.

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

Stanghellini, E. and Wermuth, N. (2005) On the identification of path analysis models with one hidden variable. *Biometrika*, **92**, 332–350.

Wermuth, N. (1980) Linear recursive equations, covariance selection, and path analysis. *J. Amer. Statist. Assoc.*, **75**, 963–997.

Wermuth, N. and Cox, D.R. (1998) On association models defined over independence graphs. *Bernoulli*, **4**, 477–495.

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.

Wilks, S.S. (1932) Certain generalisations in the analysis of variance. *Biometrika*, **24**, 471–494.

Received February 2005 and revised December 2005