Estimation of a Covariance Matrix with Zeros
Author(s): Sanjay Chaudhuri, Mathias Drton and Thomas S. Richardson
Source: *Biometrika*, Vol. 94, No. 1 (Mar., 2007), pp. 199-216
Published by: Oxford University Press on behalf of Biometrika Trust
Stable URL: https://www.jstor.org/stable/20441363
Accessed: 26-02-2023 08:22 UTC

Biometrika (2007), **94**, 1, *pp.* 199–216
doi:10.1093/biomet/asm007

Printed in Great Britain

Estimation of a covariance matrix with zeros

BY SANJAY CHAUDHURI
Department of Statistics and Applied Probability, Faculty of Science, 6 Science Drive 2, National University of Singapore, 117546, Singapore
sanjay@stat.nus.edu.sg

MATHIAS DRTON
Department of Statistics, The University of Chicago, 5734 S. University Avenue, Chicago, Illinois 60637, U.S.A.
drton@galton.uchicago.edu

AND THOMAS S. RICHARDSON
Department of Statistics, University of Washington, Box 354322, Seattle, Washington 98105-4322, U.S.A.
tsr@stat.washington.edu

SUMMARY

We consider estimation of the covariance matrix of a multivariate random vector under the constraint that certain covariances are zero. We first present an algorithm, which we call iterative conditional fitting, for computing the maximum likelihood estimate of the constrained covariance matrix, under the assumption of multivariate normality. In contrast to previous approaches, this algorithm has guaranteed convergence properties. Dropping the assumption of multivariate normality, we show how to estimate the covariance matrix in an empirical likelihood approach. These approaches are then compared via simulation and on an example of gene expression.

Some key words: Covariance graph; Empirical likelihood; Graphical model; Marginal independence; Maximum likelihood estimation; Multivariate normal distribution.

1. INTRODUCTION

In this paper we consider estimation of the covariance matrix of a random vector, subject to certain entries being set to zero. Such restrictions appear, for example, in recent work by Grzebyk et al. (2004), Mao et al. (2004) and the influential paper of Butte et al. (2000). Suppose we have a random vector $Y = (Y_1, Y_2, Y_3, Y_4)' \in \mathbb{R}^4$ whose covariance matrix Σ exhibits the zero pattern

$$\Sigma = \begin{pmatrix} \sigma_{11} & 0 & \sigma_{13} & 0 \\ 0 & \sigma_{22} & 0 & \sigma_{24} \\ \sigma_{13} & 0 & \sigma_{33} & \sigma_{34} \\ 0 & \sigma_{24} & \sigma_{34} & \sigma_{44} \end{pmatrix} \in \mathbb{R}^{4\times 4}. \tag{1}$$

It is often helpful to visualize the pattern of zeros by a so-called covariance graph. A covariance graph has one vertex for each one of the random variables in the random vector.

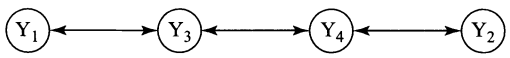

Fig. 1. The covariance graph for the matrix in (1).

In Fig. 1, the vertex set is $V = \{1, 2, 3, 4\}$, where random variable Y_i is identified with its index i. Next, each pair of vertices $(i, j) \in V \times V, i \neq j$, is connected by a bi-directed edge $i \leftrightarrow j$ unless $\sigma_{ij} = 0$. If we assume that the covariance matrix in (1) has no zero other than those indicated explicitly, its covariance graph is given in Fig. 1. Our use of bi-directed edges is in the tradition of path diagrams (Wright, 1921); other authors have used dashed edges (Cox & Wermuth, 1993, 1996).

We define a covariance graph model as the set of joint distributions in which the associated zero restrictions hold in the covariance matrix. The Gaussian covariance graph model comprises all multivariate normal distributions $\mathcal{N}(\mu, \Sigma)$ such that $\sigma_{ij} = 0$ whenever $i \neq j$ and $i \not\leftrightarrow j$. In a multivariate normal distribution it holds clearly that $\sigma_{ij} = 0$ if and only if Y_i and Y_j are marginally independent. Hence a Gaussian covariance graph model is a graphical model based on marginal independence (Edwards, 2000, §7·4).

Even in the Gaussian case, statistical inference is not well developed: the conceptual simplicity of covariance graph models belies the fact that they generally form curved and not regular exponential families. For instance, the graphical modelling software MIM (Edwards, 2000, §7·4) permits fitting of such models only by a 'dual likelihood' method (Kauermann, 1996). Wermuth et al. (2006) derived recently explicit approximations to maximum likelihood estimates that are asymptotically efficient in covariance graph models for exponential family distributions. Anderson (1969, 1970, 1973) proposed an algorithm that can be used to solve the likelihood equations of Gaussian models defined by linear hypotheses on covariance matrices, which include covariance graph models. However, it is unclear when this algorithm converges and when its limit points are positive semidefinite matrices.

In this paper, we introduce a new algorithm for maximum likelihood estimation in Gaussian covariance graph models, called iterative conditional fitting, which has clearer convergence properties than Anderson's algorithm. Moreover, we present an empirical likelihood approach to estimating the covariance matrix that provides consistent estimators even when multivariate normality does not hold.

2. COVARIANCE GRAPH MODELS

2·1. *Nonparametric model*

Suppose we observe a random vector $Y_V = (Y_i \mid i \in V)' \in \mathbb{R}^V$, with joint distribution P_V and covariance matrix $\Sigma(P_V) = (\sigma_{ij}) \in \mathbb{R}^{V \times V}$. Let $G = (V, E)$ be a graph with vertex set V and an edge set $E \subseteq V \times V \setminus \{(i, i) \mid i \in V\}$ consisting exclusively of bi-directed edges $(i, j), (j, i) \in E$. Let $\mathcal{P}(V)$ be the cone of positive definite $V \times V$ matrices, and $\mathcal{P}(G)$ the cone of matrices $\Sigma \in \mathcal{P}(V)$ which fulfil the linear restrictions

$$i \not\leftrightarrow j \quad \Longrightarrow \quad \sigma_{ij} = 0. \tag{2}$$

The covariance graph model $\mathcal{M}(G)$ associated with the bi-directed graph G is the family of joint distributions

$$\mathcal{M}(G) = \{P_V \mid \Sigma(P_V) \in \mathcal{P}(G)\}. \tag{3}$$

We consider estimation of the unknown parameter $\Sigma = \Sigma(P_V)$ based on a sample of independent and identically distributed observations $Y_V^{(k)} \in \mathbb{R}^V$, $k \in N = \{1, \ldots, n\}$, drawn from $P_V \in \mathcal{M}(G)$. The set N can be interpreted as indexing the subjects on which observations are made. We group the vectors $Y_V^{(k)}$ as columns in the $V \times N$ random matrix Y so that $\text{var}(Y) = \Sigma \otimes I_N$. Here, I_N is the $N \times N$ identity matrix and $\otimes$ is the Kronecker product. Thus the ith row $Y_i \in \mathbb{R}^N$ of the matrix Y contains the independent and identically distributed observations for variable $i \in V$ on all subjects in N and the kth column $Y_V^{(k)}$ holds all observations made on subject $k \in N$. Finally, the sample size is $n = |N|$ and the number of variables is $p = |V|$.

2·2. *Gaussian model*

We define a Gaussian covariance graph model as the multivariate normal submodel

$$\mathcal{N}(G) = \{\mathcal{N}_V(\mu, \Sigma) \mid \mu \in \mathbb{R}^V, \ \Sigma \in \mathcal{P}(G)\} \subset \mathcal{M}(G). \tag{4}$$

For observation matrix Y, the loglikelihood function ℓ of the model $\mathcal{N}(G)$ is a function from $\mathbb{R}^V \times \mathcal{P}(G)$ to $\mathbb{R}$ and can be expressed as

$$\ell(\mu, \Sigma) = -\frac{np}{2}\log(2\pi) - \frac{n}{2}\log|\Sigma| - \frac{n}{2}\text{tr}(\Sigma^{-1}\tilde{S}) \tag{5}$$

(Edwards, 2000, §3·1). Here

$$\tilde{S} = \frac{1}{n}(Y - \mu \otimes 1_N)(Y - \mu \otimes 1_N)' \in \mathbb{R}^{V \times V}, \tag{6}$$

where $1_N = (1, \ldots, 1)' \in \mathbb{R}^N$. For fixed Σ, (5) is maximized by setting $\mu = \bar{Y} \in \mathbb{R}^V$, i.e., the vector of the row means of Y. Hence, the profile loglikelihood $\ell(\Sigma)$ is obtained by replacing $\tilde{S}$ with

$$S = \frac{1}{n}(Y - \bar{Y} \otimes 1_N)(Y - \bar{Y} \otimes 1_N)' \in \mathbb{R}^{V \times V}, \tag{7}$$

in (5). The profile likelihood corresponds to the likelihood of the submodel of $\mathcal{N}(G)$ in which $\mu = 0$. However, whereas S has a Wishart distribution with $n - 1$ degrees of freedom, the empirical covariance matrix YY'/n in the submodel with $\mu = 0$ has a Wishart distribution with n degrees of freedom.

If S is positive definite, as we assume from now on, then the global maximum of $\ell(\Sigma)$ over $\mathcal{P}(G)$, i.e., the maximum likelihood estimator $\hat{\Sigma}$, exists. A sufficient but in general not necessary condition for S being positive definite with probability one is that the sample size satisfies $n \geqslant p + 1$ (Eaton & Perlman, 1973). Note that the likelihood function of the model $\mathcal{N}(G)$ may, and in fact can, have multiple local maxima, although this appears to be a phenomenon occurring predominantly under model misspecification and with small samples (Drton, 2006; Drton & Richardson, 2004).

Let

$$F = \{(i, i) \mid i \in V\} \cup \{(i, j) \in V^2 \mid i < j \text{ and } i \leftrightarrow j\} \tag{8}$$

be the pairs of vertices indexing unrestricted elements in the matrix $\Sigma \in \mathcal{P}(G)$. The cardinality of F is equal to the number of vertices plus the number of edges in the graph G. The unrestricted elements of Σ form the vector

$$\sigma = \{\sigma_{ij} \mid (i, j) \in F\} \in \mathbb{R}^F. \tag{9}$$

In order to write derivatives of the loglikelihood function in compact form we introduce the matrix Q with entries in $\{0, 1\}$ that satisfies $\text{vec}(\Sigma) = Q\sigma$, where vec is the operator of columnwise matrix vectorization. The columns of Q that are associated with a variance σ_{ii} contain exactly one entry equal to one, whereas a column of Q that is associated with a covariance σ_{ij}, $i \neq j$, $i \leftrightarrow j$, contains exactly two entries equal to one.

The first derivative of the loglikelihood function, the score function, equals

$$\frac{\partial \ell(\Sigma)}{\partial \sigma} = \frac{n}{2} Q' \{\text{vec}(\Sigma^{-1} S \Sigma^{-1}) - \text{vec}(\Sigma^{-1})\}. \tag{10}$$

It follows that the likelihood equations $\partial \ell(\Sigma)/\partial \sigma = 0$ are

$$(\Sigma^{-1})_{ij} = (\Sigma^{-1} S \Sigma^{-1})_{ij}, \quad (i, j) \in F; \tag{11}$$

compare also with Anderson & Olkin (1985, § 2·1·1). The full matrix Σ is determined by $\sigma_{ij} = 0$ for $(i, j) \notin F$, that is for $i \neq j$ and $i \not\leftrightarrow j$. The Hessian matrix equals

$$\frac{\partial^2 \ell(\Sigma)}{\partial \sigma^2} = \frac{n}{2} Q' [\{\Sigma^{-1} \otimes \Sigma^{-1}\} - \{(\Sigma^{-1} S \Sigma^{-1}) \otimes \Sigma^{-1}\} - \{\Sigma^{-1} \otimes (\Sigma^{-1} S \Sigma^{-1})\}] Q. \tag{12}$$

Its negated expectation under $\mathcal{N}_V(0, \Sigma)$, the Fisher information matrix, is

$$-E\left(\frac{\partial^2 \ell(\Sigma)}{\partial \sigma^2}\right) = \frac{n}{2} Q' (\Sigma^{-1} \otimes \Sigma^{-1}) Q \tag{13}$$

and can be used for normal approximation to the distribution of roots of the likelihood equations. Sections 3 and 4 focus on the computation of such roots.

2·3. *Kauermann's dual estimation*

Dual estimation provides estimators in Gaussian covariance graph models that are unique and asymptotically efficient though, in general, not solutions to the likelihood equations (Kauermann, 1996). In the method one maximizes a dual likelihood function obtained by interchanging the roles of the parameter matrix Σ and the empirical covariance matrix S in (5). Procedurally, one finds $\hat{\Sigma}_D \in \mathcal{P}(G)$ that solves the equations

$$\left(\hat{\Sigma}_D^{-1}\right)_{ij} = (S^{-1})_{ij}, \quad \text{for all } (i, j) \in F, \tag{14}$$

while satisfying that $(\hat{\Sigma}_D)_{ij} = 0$ for all $(i, j) \notin F$. In contrast to (11), the equation system (14) always has a unique positive definite solution, provided S is positive definite as we assume here. This solution can be found by the iterative proportional fitting algorithm (Edwards, 2000, §7·4), which terminates in finitely many steps if the covariance graph is decomposable. In this case, the estimator $\hat{\Sigma}_D$ is available in closed form.

3. ITERATIVE CONDITIONAL FITTING FOR GAUSSIAN COVARIANCE GRAPHS

We now present the new iterative conditional fitting algorithm, which uses simple least squares computations to produce positive definite roots of the likelihood equations of covariance graph models. Given some initial estimate of the joint distribution, the idea behind the algorithm is to iterate repeatedly through all vertices $i \in V$, carrying out the following steps.

Step 1. Fix the marginal distribution for the variables different from i, i.e. the variables $-i = V \setminus \{i\}$.

Step 2. Estimate, by maximum likelihood, the conditional distribution of variable i given the variables $-i$ under the constraints implied by the model $\mathcal{N}(G)$.

Step 3. Find a new estimate of the joint distribution by multiplying together the fixed marginal and the estimated conditional distributions.

Since we fix the marginal distribution of variables $-i$ in the update for variable i, all marginal independences amongst the variables $-i$ still hold true after the update. Therefore, only marginal independences involving variable i lead to constraints in Step 2.

In order to make the idea more precise, let $\Sigma_{A,B}$ be the $A \times B$ submatrix of Σ, and let Y_A be the $A \times N$ submatrix of $Y \sim \mathcal{N}_V(0, \Sigma \otimes I_N)$, where $A, B \subseteq V$. Clearly,

$$Y_{-i} \sim \mathcal{N}_{-i \times N}(0, \Sigma_{-i,-i} \otimes I_N).$$

Hence, Step 1 simply fixes the value of $\Sigma_{-i,-i}$, i.e. everything but the ith row and column of Σ. As $\Sigma_{-i,-i}$ remains unchanged in the ith update, many of the zero constraints imposed on the covariance matrix trivially hold true even after the update.

The conditional distribution of Y_i given Y_{-i} is normal,

$$(Y_i \mid Y_{-i}) \sim \mathcal{N}_{\{i\} \times N}\{\Sigma_{i,-i}(\Sigma_{-i,-i})^{-1} Y_{-i}, \lambda_i I_N\}, \tag{15}$$

with conditional variance

$$\lambda_i = \sigma_{ii} - \Sigma_{i,-i}(\Sigma_{-i,-i})^{-1}\Sigma_{-i,i} > 0. \tag{16}$$

For the complete graph $\bar{G}$ in which an edge joins any pair of vertices the mapping

$$\begin{aligned} \mathcal{P}(\bar{G}) &\to (0, \infty) \times \mathbb{R}^{\{i\} \times -i} \times \mathcal{P}_{-i}(\bar{G}), \\ \Sigma &\mapsto (\lambda_i, \Sigma_{i,-i}, \Sigma_{-i,-i}) \end{aligned} \tag{17}$$

is bijective, and (15) presents a standard least squares regression. Here, $\mathcal{P}_A(G)$ is the set of all $A \times A$ submatrices of matrices in $\mathcal{P}(G)$, $A \subseteq V$. For a general graph G, (15) is not a standard regression because we need to respect the fact that $\Sigma \in \mathcal{P}(G)$, i.e. that $\sigma_{ij} = 0$ if $j \neq i$, $j \not\leftrightarrow i$. However, this can be circumvented using 'pseudo-variables' that are computed from the data Y_{-i} and the fixed matrix $\Sigma_{-i,-i}$.

Let $\mathrm{sp}(i) = \{j \mid i \leftrightarrow j\}$ be the set of 'spouses' of $i \in V$ and let $\mathrm{nsp}(i) = V \setminus (\mathrm{sp}(i) \cup \{i\})$ be the set of 'non-spouses,' yielding the partition $V = \{i\} \cup \mathrm{sp}(i) \cup \mathrm{nsp}(i)$. Then the conditional expectation of Y_i given Y_{-i} can be written as

$$E(Y_i \mid Y_{-i}) = \Sigma_{i,-i}\{(\Sigma_{-i,-i})^{-1} Y_{-i}\} = \Sigma_{i,\mathrm{sp}(i)} Z^{(i)}_{\mathrm{sp}(i)} = \sum_{j \in \mathrm{sp}(i)} \sigma_{ij} Z^{(i)}_j, \tag{18}$$

where the pseudo-variable $Z^{(i)}_j$ is equal to the jth row in

$$Z^{(i)}_{\mathrm{sp}(i)} = \{(\Sigma_{-i,-i})^{-1}\}_{\mathrm{sp}(i),-i} Y_{-i} \in \mathbb{R}^{\mathrm{sp}(i) \times N}. \tag{19}$$

In (18), we exploit the fact that $\sigma_{ij} = 0$ if $j \in \mathrm{nsp}(i)$. From (18), we obtain

$$(Y_i \mid Y_{-i}) \sim \mathcal{N}_{\{i\} \times N}\Big(\sum_{j \in \mathrm{sp}(i)} \sigma_{ij} Z^{(i)}_j, \lambda_i I_N\Big). \tag{20}$$

The bijectivity of the map

$$\begin{aligned} \mathcal{P}(G) &\to (0, \infty) \times \mathbb{R}^{\{i\} \times \mathrm{sp}(i)} \times \mathcal{P}_{-i}(G), \\ \Sigma &\mapsto (\lambda_i, \Sigma_{i,\mathrm{sp}(i)}, \Sigma_{-i,-i}) \end{aligned} \tag{21}$$

implies that σ_{ij}, $j \in \mathrm{sp}(i)$, and λ_i are variation independent in (20). Hence, if $\Sigma_{-i,-i} \in \mathcal{P}_{-i}(G)$ is fixed, then (20) constitutes a standard normal regression whose

parameters σ_{ij}, $j \in \mathrm{sp}(i)$, and λ_i can be estimated by the usual least squares formulae. By (16), the estimate of λ_i determines an estimator of σ_{ii}. Thus, we obtain the maximum likelihood estimator of the ith row and column of Σ when $\Sigma_{-i,-i}$ is fixed. In particular, after updating the ith row and column we are still left with a matrix $\Sigma \in \mathcal{P}(G)$.

For the precise formulation of the algorithm, let $\hat{\Sigma}^{(r)}$ be the estimate of Σ after the rth iteration and $\hat{\Sigma}^{(r,i)}$ the estimate of Σ after the ith update step of the algorithm's rth iteration in iterative conditional fitting, i.e., after estimating $(Y_i \mid Y_{-i})$.

ALGORITHM 1. *Iterative conditional fitting can be implemented as follows.*

Step 1. *Set the iteration counter* $r = 0$, *and choose a starting value* $\hat{\Sigma}^{(0)} \in \mathcal{P}(G)$, *such as the identity matrix* $\hat{\Sigma}^{(0)} = I_V$ *or the dual estimate* $\hat{\Sigma}^{(0)} = \hat{\Sigma}_{\mathrm{D}}$.

Step 2. *Order the variables as* $V = \{1, \ldots, p\}$, *set* $\hat{\Sigma}^{(r,0)} = \hat{\Sigma}^{(r)}$, *and repeat the following steps for all* $i = 1, \ldots, p$:

(i) *let* $\hat{\Sigma}^{(r,i)}_{-i,-i} = \hat{\Sigma}^{(r,i-1)}_{-i,-i}$ *and calculate from this submatrix the pseudo-variables* $Z^{(i)}_{sp(i)}$ *according to* (19);

(ii) *compute the maximum likelihood estimates*

$$\begin{aligned} \hat{\Sigma}^{(r,i)}_{i,sp(i)} &= Y_i \, (Z^{(i)}_{sp(i)})' \Big\{ Z^{(i)}_{sp(i)} (Z^{(i)}_{sp(i)})' \Big\}^{-1}, \\ \hat{\lambda}_i &= \frac{1}{n} \Big(Y_i - \hat{\Sigma}^{(r,i)}_{i,sp(i)} Z^{(i)}_{sp(i)} \Big) \Big(Y_i - \hat{\Sigma}^{(r,i)}_{i,sp(i)} Z^{(i)}_{sp(i)} \Big)', \end{aligned} \tag{22}$$

for the linear regression (20), *the existence of the matrix inverse following from the assumed nonsingularity of the empirical covariance matrix* S;

(iii) *complete* $\hat{\Sigma}^{(r,i)}$ *by solving for* σ_{ii} *in* (16) *and thus setting*

$$\hat{\sigma}^{(r,i)}_{ii} = \hat{\lambda}_i + \hat{\Sigma}^{(r,i)}_{i,sp(i)} \{ (\hat{\Sigma}^{(r,i)}_{-i,-i})^{-1} \}_{sp(i),sp(i)} \hat{\Sigma}^{(r,i)}_{sp(i),i}. \tag{23}$$

Step 3. *Set* $\hat{\Sigma}^{(r+1)} = \hat{\Sigma}^{(r,p)}$, *increment the counter* r *to* $r + 1$ *and return to Step* 2.

The iterations can be stopped according to a criterion such as 'the norm of the score function (10) at the estimate is smaller than a predetermined tolerance.' A less transparent version of Algorithm 1 is described in Drton & Richardson (2003); it is implemented in the 'fitCovGraph' function in the 'ggm' R package.

For illustration, consider the graph G from Fig. 1. Iterative conditional fitting of the model $\mathcal{N}(G)$ cycles, in arbitrary order, through the four regressions $(Y_i \mid Y_{-i})$, $i = 1, 2, 3, 4$. The regressions are illustrated in Fig. 2, where unfilled circles represent the conditioning set $-i$, and a filled circle stands for the response variable i. The vertices that are joined to vertex i by a directed edge are labelled with the pseudo-variables that act as covariates. The covariances to be estimated label these directed edges.

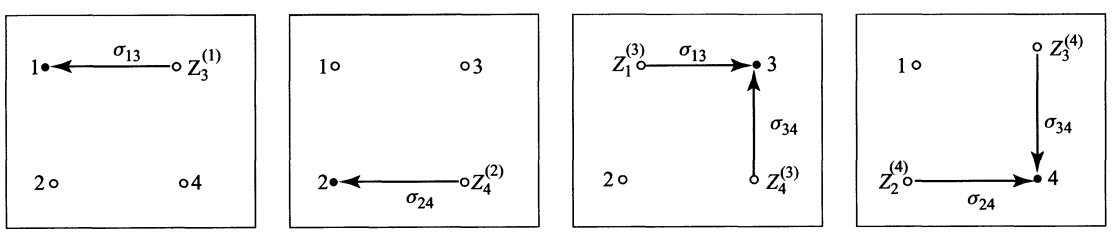

Fig. 2. Illustration of the pseudo-variable regressions in iterative conditional fitting.

THEOREM 1. *Suppose the empirical covariance matrix S is positive definite, $\hat{\Sigma}^{(0)}$ is an arbitrary starting value in $\mathcal{P}(G)$, and $(\hat{\Sigma}^{(r)})$ is the sequence constructed by iterative conditional fitting as described in Algorithm* 1. *Then all accumulation points of $(\hat{\Sigma}^{(r)})$ are positive definite saddlepoints or local maxima of the likelihood function, and all accumulation points have the same value of the likelihood function. It follows that the sequence $(\hat{\Sigma}^{(r)})$ converges if, for all $c \in \mathbb{R}$, the contour set $\{\Sigma \in \mathcal{P}(G) \mid \ell(\Sigma) = c\}$ contains at most finitely many solutions to the likelihood equations.*

It remains an open question whether there can ever be an infinite number of solutions to the likelihood equations with a positive definite sample covariance matrix; we have never encountered such an example, real or artificial.

Proof of Theorem 1. The key to the proof of convergence is to recognize that the algorithm consists of iterated maximizations over sections of the parameter space $\mathcal{P}(G)$. In Algorithm 1, we repeatedly maximize the likelihood function of the covariance graph model partially by allowing only the entries in the ith row and column of Σ to vary. The remaining entries are fixed.

More formally, consider the parameter space

$$\Theta = \{\Sigma \in \mathcal{P}(G) \mid \ell(\Sigma) \geqslant \ell(\hat{\Sigma}^{(0)})\}, \tag{24}$$

which is compact, though not necessarily connected. Recall that we assume the empirical covariance matrix S to be positive definite, in which case the loglikelihood function $\ell(\Sigma)$ tends to minus infinity if Σ approaches a singular matrix or if one or more of its elements tend to infinity (Anderson, 1958, p. 47). If we define

$$\Theta_i(\bar{\Sigma}) = \{\Sigma \in \Theta \mid \Sigma_{-i,-i} = \bar{\Sigma}_{-i,-i}\} \subsetneq \Theta, \tag{25}$$

then computing the maximum likelihood estimates for the linear regression with pseudo-variables in (20) yields the maximizer of the conditional likelihood function for $(Y_i \mid Y_{-i})$ under the constraint $\Sigma \in \Theta_i(\hat{\Sigma}^{(r,i-1)})$. Since under this constraint the marginal likelihood function for Y_{-i} is fixed, the algorithm Steps 2(i)-2(iii) maximize the joint loglikelihood function over the section $\Theta_i(\hat{\Sigma}^{(r,i-1)})$, i.e.

$$\hat{\Sigma}^{(r,i)} = \arg\max\{\ell(\Sigma) \mid \Sigma \in \Theta_i(\hat{\Sigma}^{(r,i-1)})\}. \tag{26}$$

This local and global maximizer over the section is unique because the maximization over the section corresponds to maximum likelihood fitting of a standard regression model with log-concave likelihood function. Since the least squares computations in regression involve only rational expressions, and there are no poles, the map taking $\hat{\Sigma}^{(r,i-1)}$ to $\hat{\Sigma}^{(r,i)}$ is, in particular, continuous. Note that clearly $\ell(\hat{\Sigma}^{(r,i)}) \geqslant \ell(\hat{\Sigma}^{(r,i-1)})$.

If $\Sigma \in \mathcal{P}(G)$ maximizes the loglikelihood function over all sections $\Theta_i(\Sigma)$, $i \in V$, simultaneously, then it solves the likelihood equations. This holds because such a section maximizer Σ is, in particular, a maximizer of $\ell(\Sigma)$ if only an individual entry σ_{ij} is varied.

At this stage we can invoke general results on iterative partial maximization algorithms. Lauritzen (1996, Appendix A.4) provides convergence results if there is only one solution to the likelihood equations, which need not be the case here. However, the results in the Appendix of Drton & Eichler (2006) encompass the situation with multiple roots to the likelihood equations and imply the claim. □

Algorithm 1 can be restated purely in terms of the empirical covariance matrix S defined in (7). For example, in (22),

$$\begin{aligned} Y_i(Z^{(i)}_{\mathrm{sp}(i)})' &= S_{i,-i}\{(\Sigma_{-i,-i})^{-1}\}_{-i,\mathrm{sp}(i)}, \\ Z^{(i)}_{\mathrm{sp}(i)}(Z^{(i)}_{\mathrm{sp}(i)})' &= \{(\Sigma_{-i,-i})^{-1}\}_{\mathrm{sp}(i),-i} S_{-i,-i}\{(\Sigma_{-i,-i})^{-1}\}_{-i,\mathrm{sp}(i)}. \end{aligned} \tag{27}$$

Thus, sample size does not affect the complexity of the algorithm. The complexity of one of the algorithm's pseudo-variable regression steps is dominated by the computation of the inverse of $\Sigma_{-i,-i}$ in (19), and the inversion of a matrix of size $\mathrm{sp}(i) \times \mathrm{sp}(i)$ in (22). If $\Sigma_{-i,-i}$ is sparse, special methods for inversion of sparse matrices may be employed. In particular, if the induced subgraph G_{-i} has disconnected components then only the submatrices of Σ over connected components containing spouses of i have to be inverted.

4. ITERATIVE CONDITIONAL FITTING WITH MULTIVARIATE UPDATES

Algorithm 1 updates an estimate of the covariance matrix $\Sigma \in \mathcal{P}(G)$ one row and column at a time. A natural modification of this approach is to update jointly several rows and columns of the estimate $\Sigma \in \mathcal{P}(G)$. For a subset $C \subseteq V$, let $-C = V \setminus C$. In order to estimate simultaneously all rows and columns of Σ that are indexed by the vertices in C we would like to consider one multivariate regression of the form $(Y_C \mid Y_{-C})$. The conditional distribution of Y_C given Y_{-C} is multivariate normal:

$$(Y_C \mid Y_{-C}) \sim \mathcal{N}_{C\times N}\{\Sigma_{C,-C}(\Sigma_{-C,-C})^{-1}Y_{-C}, \Lambda_C \otimes I_N\} \tag{28}$$

with conditional covariance matrix

$$\Lambda_C = \Sigma_{C,C} - \Sigma_{C,-C}(\Sigma_{-C,-C})^{-1}\Sigma_{-C,C} \in \mathcal{P}(C). \tag{29}$$

For the conditional distribution in (28) to be of simple structure, Λ_C should be unconstrained. This is achieved if C is chosen to be a complete set such that $i \leftrightarrow j$ whenever $i, j \in C$ and $i \neq j$. Then all constraints are of the form $\sigma_{ij} = 0$ for $i \in C$, $j \notin C$ and $j \not\leftrightarrow i$, and affect only the regression coefficients in $\Sigma_{C,-C}(\Sigma_{-C,-C})^{-1}$. Hence, the submatrix $\Sigma_{C,C}$, and thus Λ_C, remains unconstrained.

Let $\mathrm{sp}(C) = \cup_{i\in C}\mathrm{sp}(i) \setminus C$ be the spouses of C, that is, the vertices that are not in C but adjacent to some vertex in C, and let $\mathrm{nsp}(C) = V \setminus \{\mathrm{sp}(C) \cup C\}$ be the non-spouses of C, yielding the partition $V = C \cup \mathrm{sp}(C) \cup \mathrm{nsp}(C)$. If we define the pseudo-variables

$$Z^{(C)}_{\mathrm{sp}(C)} = \{(\Sigma_{-C,-C})^{-1}\}_{\mathrm{sp}(C),-C}\, Y_{-C} \in \mathbb{R}^{\mathrm{sp}(C)\times N}, \tag{30}$$

then we can rewrite (28) as

$$(Y_C \mid Y_{-C}) \sim \mathcal{N}_{C\times N}\left(\Sigma_{C,\mathrm{sp}(C)}\, Z^{(C)}_{\mathrm{sp}(C)}, \Lambda_C \otimes I_N\right), \tag{31}$$

because $\Sigma_{C,\mathrm{nsp}(C)} = 0$. As Σ ranges through $\mathcal{P}(G)$, the submatrix $\Sigma_{C,\mathrm{sp}(C)}$ playing the role of regression coefficients in (31) ranges through the linear space

$$\mathcal{P}_{C,\mathrm{sp}(C)}(G) = \left\{A \in \mathbb{R}^{C\times\mathrm{sp}(C)} \mid A_{ij} = 0 \text{ if } i \not\leftrightarrow j\right\}. \tag{32}$$

Hence, (31) constitutes seemingly unrelated regressions (Zellner, 1962).

Maximum likelihood estimation in seemingly unrelated regressions itself generally requires iterative algorithms, such as iterating the two-step estimator of Zellner (1962). In the case of (31), the two-step estimator consists of first estimating $\Sigma_{C,\mathrm{sp}(C)}$ for some fixed Λ_C by generalized least squares, and then estimating Λ_C as the empirical covariance matrix

of the residuals $Y_i - \Sigma_{C,\mathrm{sp}(C)} Z^{(C)}_{\mathrm{sp}(C)}$ computed with the estimate of $\Sigma_{C,\mathrm{sp}(C)}$ obtained in the first step. However, if the current estimate of Σ is used to obtain starting values $\Sigma_{C,\mathrm{sp}(C)}$ and Λ_C, then the two-step method does not have to be iterated in order to obtain a convergent iterative conditional fitting algorithm with multivariate updates. For specification of the estimator of $\Sigma_{C,\mathrm{sp}(C)}$ we introduce the matrix P_C of the linear map that sends the vector of unrestricted elements in $\Sigma_{C,\mathrm{sp}(C)}$, i.e. the vector $\sigma_C = \{\sigma_{ij} \mid i \in C,\ j \in \mathrm{sp}(C),\ i \leftrightarrow j\}$, to the matrix $\Sigma_{C,\mathrm{sp}(C)} \in \mathcal{P}_{C,\mathrm{sp}(C)}(G)$. The matrix P_C has exactly one entry equal to one in each column, the other entries being zero, and it satisfies $\mathrm{vec}(\Sigma_{C,\mathrm{sp}(C)}) = P_C \sigma_C$ for $\Sigma \in \mathcal{P}(G)$; compare the definition of the matrix Q in §2.

In order to run iterative conditional fitting with multivariate updates, we have to choose a family of complete but not necessarily disjoint sets $(C \mid C \in \mathcal{C})$ such that

$$\cup(C \mid C \in \mathcal{C}) = V. \tag{33}$$

For example, the sets C could be chosen as edges, but the largest possible choice for the sets C would be the cliques, i.e. the maximal complete sets, in G.

ALGORITHM 2. *For a given choice of the family of subsets $\mathcal{C}$, iterative conditional fitting with multivariate updates can be implemented as follows.*

Step 1. *Set the iteration counter $r = 0$, and choose a starting value $\hat{\Sigma}^{(0)} \in \mathcal{P}(G)$, such as the identity matrix $\hat{\Sigma}^{(0)} = I_V$ or the dual estimate $\hat{\Sigma}^{(0)} = \hat{\Sigma}_{\mathrm{D}}$.*

Step 2. *Order the sets in the family $\mathcal{C}$ as $\mathcal{C} = \{C_1, \ldots, C_q\}$, set $\hat{\Sigma}^{(r,0)} = \hat{\Sigma}^{(r)}$, and repeat the following steps for all $C_k \in \mathcal{C}$:*

(i) *let $\hat{\Sigma}^{(r,k)}_{-C_k,-C_k} = \hat{\Sigma}^{(r,k-1)}_{-C_k,-C_k}$, compute from this submatrix the conditional covariance matrix $\hat{\Lambda}_{C_k}$ according to (29) and the pseudo-variables $Z^{(C_k)}_{sp(C_k)}$ according to (30), and calculate $\hat{\Omega}_{C_k} = (\hat{\Lambda}_{C_k})^{-1}$;*

(ii) *compute the generalized least squares matrix that satisfies $\mathrm{vec}(\hat{\Sigma}^{(r,k)}_{C_k,sp(C_k)}) = P_{C_k}\hat{\sigma}_{C_k}$, where*

$$\hat{\sigma}_{C_k} = \Big(P'_{C_k}\Big[\Big\{Z^{(C_k)}_{sp(C_k)}\Big(Z^{(C_k)}_{sp(C_k)}\Big)'\Big\} \otimes \hat{\Omega}_{C_k}\Big] P_{C_k}\Big)^{-1} \times \Big[P'_{C_k}\,\mathrm{vec}\Big\{\hat{\Omega}_{C_k} Y_{C_k}\Big(Z^{(C_k)}_{sp(C_k)}\Big)'\Big\}\Big]; \tag{34}$$

(iii) *compute the empirical covariance matrix of residuals,*

$$\hat{\Lambda}_{C_k} = \frac{1}{n}\Big(Y_{C_k} - \hat{\Sigma}^{(r,k)}_{C_k,sp(C_k)} Z^{(C_k)}_{sp(C_k)}\Big)\Big(Y_{C_k} - \hat{\Sigma}^{(r,k)}_{C_k,sp(C_k)} Z^{(C_k)}_{sp(C_k)}\Big)'; \tag{35}$$

(iv) *complete $\hat{\Sigma}^{(r,k)}$ by solving for Σ_{C_k,C_k} via (29) and thus setting*

$$\hat{\Sigma}^{(r,k)}_{C_k,C_k} = \hat{\Lambda}_{C_k} + \hat{\Sigma}^{(r,k)}_{C_k,sp(C_k)}\Big\{\Big(\hat{\Sigma}^{(r,k)}_{-C_k,-C_k}\Big)^{-1}\Big\}_{sp(C_k),sp(C_k)} \hat{\Sigma}^{(r,k)}_{sp(C_k),C_k}. \tag{36}$$

Step 3. *Set $\hat{\Sigma}^{(r+1)} = \hat{\Sigma}^{(r,q)}$, increment the counter r to $r+1$, and return to Step* 2.

Clearly, Algorithm 2 reduces to Algorithm 1 if the family $\mathcal{C}$ comprises only singletons. A stopping rule can again be based on the norm of the score function (10).

For illustration, we take up the covariance graph shown in Fig. 1. For the family $\mathcal{C}$ of complete vertex sets, several choices are possible. If the cliques $\mathcal{C} = \{13, 34, 24\}$ are chosen, then all considered conditional distributions are bivariate, whereas for $\mathcal{C} = \{1, 2, 34\}$ two univariate distributions are estimated in conjunction with a bivariate distribution. For

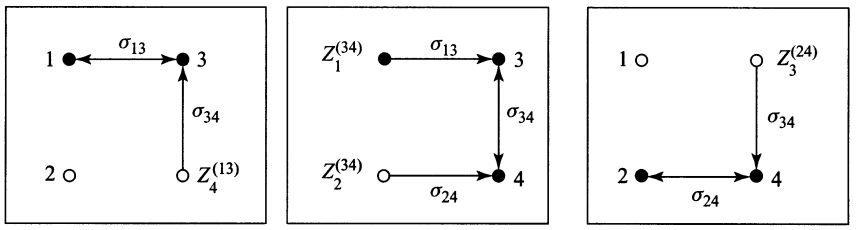

Fig. 3. Illustration of the seemingly unrelated pseudo-variable regressions in iterative conditional fitting with multivariate updates using $\mathcal{C} = \{13, 34, 24\}$.

the clique choice $\mathcal{C} = \{13, 34, 24\}$, we illustrate the seemingly unrelated pseudo-variable regressions to be estimated in Fig. 3, which is to be interpreted similarly to Fig. 2. An additional feature are the bi-directed edges that connect the vertices in the sets $C \in \mathcal{C}$ and which can be interpreted as in path diagrams (Wright, 1921).

THEOREM 2. *The claim of Theorem* 1 *holds for the sequence* $(\hat{\Sigma}^{(r)})$ *constructed by iterative conditional fitting with multivariate updates as described in Algorithm* 2.

Proof. Iterative conditional fitting with multivariate updates is still an iterative partial maximization algorithm but the maximizations are performed over parameter space sections different from those described in the proof of Theorem 1. Steps 2(ii) and 2(iii) of Algorithm 2 do not jointly maximize the loglikelihood function ℓ over the sections

$$\Theta_C(\bar{\Sigma}) = \{\Sigma \in \Theta \mid \Sigma_{-C,-C} = \bar{\Sigma}_{-C,-C}\}, \quad C \in \mathcal{C}. \tag{37}$$

Instead Step 2(ii) maximizes ℓ over sections of the form

$$\Theta_{1,C}(\bar{\Sigma}) = \{\Sigma \in \Theta \mid \Sigma_{-C,-C} = \bar{\Sigma}_{-C,-C},\ \Lambda_C = \bar{\Lambda}_C\}, \tag{38}$$

where Λ_C is again the conditional covariance matrix from (29). The subsequent Step 2(iii) maximizes ℓ over sections of the form

$$\Theta_{2,C}(\bar{\Sigma}) = \{\Sigma \in \Theta \mid \Sigma_{-C,-C} = \bar{\Sigma}_{-C,-C},\ \Sigma_{C,-C} = \bar{\Sigma}_{C,-C}\}. \tag{39}$$

Nevertheless it holds under condition (33) that, if Σ maximizes the loglikelihood function ℓ over both $\Theta_{1,C}(\bar{\Sigma})$ and $\Theta_{2,C}(\bar{\Sigma})$ simultaneously for all $C \in \mathcal{C}$, then Σ solves the likelihood equations. Thus, the claim follows from the Appendix of Drton & Eichler (2006). □

5. EMPIRICAL LIKELIHOOD ESTIMATION

If it is not appropriate to assume multivariate normality, we may still wish to estimate a covariance matrix subject to zero restrictions. Here we present an approach based on empirical likelihood (Owen, 2001), in which an estimate of the underlying distribution is obtained by maximising a nonparametric likelihood under constraints that include the desired zero covariance restrictions; see Hellerstein & Imbens (1999) and a University of Washington technical report by S. Chaudhuri, M. S. Handcock and M. Rendall for similar applications of empirical likelihood.

We associate a weight w_k with the kth sample observation $Y_V^{(k)}$, $k \in N$. Estimating the mean vector and covariance matrix simultaneously, we solve the nested constrained

maximization problem

$$\max_{\mu}\Big(\max_{w=(w_1,\ldots,w_n)}\prod_{k\in N} nw_k\Big) \tag{40}$$

subject to

$$w_k \geqslant 0, k \in N, \tag{41}$$

$$\sum_{k\in N} w_k = 1, \tag{42}$$

$$\sum_{k\in N} w_k\Big(Y_i^{(k)} - \mu_i\Big) = 0, \quad \text{for all } i \in V, \tag{43}$$

$$\sum_{k\in N} w_k\Big(Y_i^{(k)} - \mu_i\Big)\Big(Y_j^{(k)} - \mu_j\Big) = 0, \quad \text{for all } i, j \in V \text{ such that } i \not\leftrightarrow j. \tag{44}$$

Without the constraints (43) and (44), the empirical likelihood ratio $\prod_{k\in N} nw_k$ is maximized for $w_k = 1/n, k \in N$. The additional constraint (43) forces the mean of the reweighted rows of Y to be equal to μ, and (44) ensures that the estimated weights $\hat{w}_k$ are such that the empirical covariance matrix of the reweighted sample satisfies the zero constraints specified by the graph G. More precisely, let $\hat{\mu}$ and $\hat{w}$, respectively, be the mean and weight vectors maximising (40) under the constraints (41)–(44). Denote by $\mathrm{diag}(\hat{w})$ the $N \times N$ diagonal matrix with $\hat{w}$ along its diagonal. Then

$$\hat{\Sigma}_{\mathrm{E}} = \big(Y - \hat{\mu} \otimes 1_N\big)\ \mathrm{diag}(\hat{w})\ \big(Y - \hat{\mu} \otimes 1_N\big)' \tag{45}$$

is a covariance matrix estimator that exhibits the desired pattern of zeros and is positive semidefinite. In order to avoid obvious problems with feasibility of the optimization problem, the sample size, i.e. the number of weights, must be strictly larger than the number of constraints in (42), (43) and (44). Note that the number of constraints in (44) may grow quadratically as the number of variables increases. It is our experience that $\hat{\Sigma}_E$ is positive definite for generic data Y, but additional work is required to provide rigorous guarantees for existence and positive definiteness of $\hat{\Sigma}_{\mathrm{E}}$. For sample size tending to infinity, however, $\hat{\mu}$ and $\hat{\Sigma}_{\mathrm{E}}$ are asymptotically consistent, as can be shown following Owen (2001) and Qin & Lawless (1994).

In order to compute $\hat{\mu}$ and $\hat{w}$, we proceed as follows. We use a standard optimization algorithm, such as that implemented in the 'optim' function in R, to solve the unconstrained outer maximization problem over μ in (40). Following Owen (2001, Ch. 3), we solve the linearly constrained inner problem via the corresponding dual optimization problem, which involves as many unknowns as there are constraints in (43) and (44). For large sample size n the dual is thus of much lower dimension than the original problem.

The objective function of the dual problem is obtained by maximising the Lagrangian

$$\begin{aligned} L(w, \alpha, \lambda, \gamma) = \sum_{k\in N} \log(nw_k) - \alpha\Big(\sum_{k\in N} w_k - 1\Big) - n\sum_{i\in V}\lambda^{(i)}\sum_{k\in N} w_k\Big(Y_i^{(k)} - \mu_i\Big) \\ - n\sum_{i,j\in V, i\not\leftrightarrow j}\gamma^{(i,j)}\sum_{k\in N} w_k\Big(Y_i^{(k)} - \mu_i\Big)\Big(Y_j^{(k)} - \mu_j\Big) \end{aligned} \tag{46}$$

with respect to the weights w_k. The quantities α, $\lambda^{(i)}$ and $\gamma^{(i,j)}$ in (46) are Lagrange multipliers. By differentiating (46) with respect to w_k and then noting that

$\sum_{k\in N} w_k \partial L/\partial w_k = 0$, we obtain that $\alpha = n$ and, for all $k \in N$, $w_k = (nL_k)^{-1}$, where

$$L_k = \Big\{1 + \sum_{i\in V} \lambda^{(i)}\Big(Y_i^{(k)} - \mu_i\Big) + \sum_{i,j\in V, i\not\leftrightarrow j} \gamma^{(i,j)}\Big(Y_i^{(k)} - \mu_i\Big)\Big(Y_j^{(k)} - \mu_j\Big)\Big\}. \qquad (47)$$

Thus, in the dual problem we minimize $-\sum_{k\in N}\log(L_k)$ over $\lambda^{(i)}$, $i \in V$ and $\gamma^{(i,j)}$, $i, j \in V, i \not\leftrightarrow j$, subject to the constraints $L_k \geqslant 1/n$ for all $k \in N$. The R package 'emplik' developed by Owen provides routines to solve this dual problem. In §6, we compare the maximum, the dual and the empirical likelihood approach to estimation of a covariance matrix with zeros in a data example and in simulations.

6. DATA AND SIMULATIONS

6·1. *Gene expression in yeast*

Gasch et al. (2000) present gene expression data from microarray experiments with yeast strands. We focus on $p = 8$ genes related to galactose use. The gene GAL11 is responsible for transcription. The genes GAL4 and GAL80 are involved in galactose regulation. Gene GAL2 is related to transport and the remaining four genes, GAL1, GAL3, GAL7 and GAL10, are involved in galactose metabolism. There are $n = 134$ experiments with gene expression measurements for all eight genes. Observed marginal correlations and standard deviations are shown in Table 1, where, in an obvious index correspondence, the variables Y_i represent the gene expression measurements.

By multiple testing of correlations as described in Drton & Perlman (2004) and implemented in the R package 'SIN', we selected the two covariance graphs $G_s \subset G_d$ shown in Fig. 4. The larger graph G_d contains both the solid and the dashed edges, whereas the sub-graph G_s includes only the solid edges. In G_s the set $\{1, 2, 3, 7, 10\}$ forms a clique, whereas in G_d the clique is enlarged to include vertex 80. With the R package 'ggm' and additional code, we computed the three different estimates of the covariance matrix under the zero constraints specified by the graphs; see Table 2. We remark that checking the Hessian (12) confirmed that the 'maximum likelihood estimates' are local maxima of the likelihood. Using different starting values for iterative conditional fitting, including the identity matrix and the dual estimate $\hat{\Sigma}_D$, yielded identical results, but there is no theoretical guarantee that global maxima were found.

Inspection of Table 2 shows that the three estimates are in better agreement for the graph G_d, which yields the better fitting covariance graph model. The deviance of the model

Table 1. *Yeast data. Observed marginal correlations and standard deviations,* SD

	Y_{11}	Y_4	Y_{80}	Y_2	Y_1	Y_3	Y_7	Y_{10}
Y_{11}								
Y_4	0·24							
Y_{80}	0·08	0·23						
Y_2	−0·18	−0·03	0·26					
Y_1	−0·10	−0·10	0·28	0·87				
Y_3	−0·18	0·12	0·20	0·44	0·39			
Y_7	−0·07	−0·08	0·21	0·81	0·88	0·50		
Y_{10}	−0·08	−0·07	0·26	0·87	0·92	0·46	0·91	
SD	0·39	0·36	0·47	1·70	1·70	0·78	1·85	1·54

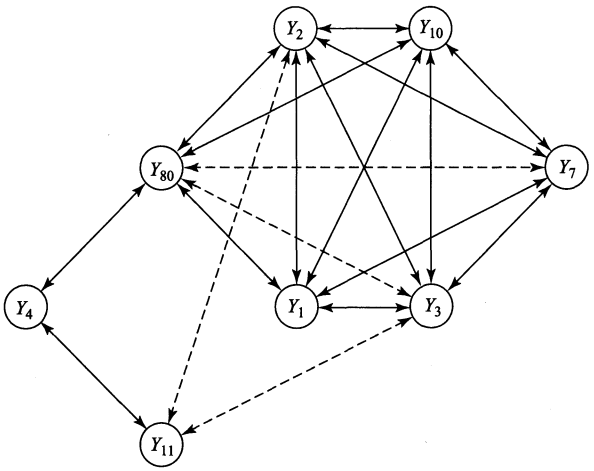

Fig. 4. Covariance graph for data in Table 1.

Table 2. *Yeast data. Marginal correlations and standard deviations,* SD, *from maximum,* M, *dual,* D, *and empirical likelihood,* E, *estimates for graph* G_s, *lower half, and graph* G_d, *upper italicized half*

	Y_{11}	Y_4	Y_{80}	Y_2	Y_1	Y_3	Y_7	Y_{10}	SD	
Y_{11}		*0·28*	*0*	*−0·12*	*0*	*−0·21*	*0*	*0*	*0·40*	M
		0·26	*0*	*−0·11*	*0*	*−0·20*	*0*	*0*	*0·39*	D
		0·25	*0*	*−0·11*	*0*	*−0·20*	*0*	*0*	*0·39*	E
Y_4	0·22		*0·20*	*0*	*0*	*0*	*0*	*0*	*0·36*	M
	0·27		*0·21*	*0*	*0*	*0*	*0*	*0*	*0·35*	D
	0·28		*0·27*	*0*	*0*	*0*	*0*	*0*	*0·36*	E
Y_{80}	0	0·22		*0·27*	*0·29*	*0·19*	*0·22*	*0·27*	*0·47*	M
	0	0·20		*0·28*	*0·31*	*0·19*	*0·23*	*0·28*	*0·47*	D
	0	0·18		*0·26*	*0·31*	*0·16*	*0·21*	*0·27*	*0·48*	E
Y_2	0	0	0·08		*0·86*	*0·43*	*0·81*	*0·87*	*1·69*	M
	0	0	0·09		*0·86*	*0·43*	*0·81*	*0·87*	*1·68*	D
	0	0	0·17		*0·83*	*0·43*	*0·79*	*0·85*	*1·48*	E
Y_1	0	0	0·11	0·86		*0·38*	*0·88*	*0·92*	*1·70*	M
	0	0	0·12	0·86		*0·39*	*0·88*	*0·91*	*1·69*	D
	0	0	0·10	0·83		*0·34*	*0·85*	*0·88*	*1·48*	E
Y_3	0	0	0	0·43	0·38		*0·49*	*0·44*	*0·78*	M
	0	0	0	0·39	0·37		*0·51*	*0·46*	*0·78*	D
	0	0	0	0·39	0·31		*0·49*	*0·46*	*0·78*	E
Y_7	0	0	0	0·81	0·88	0·50		*0·91*	*1·85*	M
	0	0	0	0·80	0·87	0·50		*0·91*	*1·84*	D
	0	0	0	0·77	0·83	0·38		*0·90*	*1·68*	E
Y_{10}	0	0	0·08	0·87	0·91	0·45	0·91		*1·54*	M
	0	0	0·08	0·86	0·91	0·44	0·90		*1·53*	D
	0	0	0·13	0·86	0·87	0·36	0·88		*1·36*	E
SD	0·39	0·36	0·47	1·70	1·70	0·78	1·85	1·54		M
	0·37	0·35	0·45	1·61	1·61	0·75	1·79	1·47		D
	0·38	0·33	0·47	1·41	1·37	0·74	1·57	1·22		E

$\mathcal{N}(G_d)$ under comparison to the model based on the complete graph equals 9·98 over 9 degrees of freedom, whereas the deviance of $\mathcal{N}(G_s)$ equals 33·07 over 13 degrees of freedom. This indicates a good fit of $\mathcal{N}(G_d)$ and a poor fit of the more restrictive model $\mathcal{N}(G_s)$. The difference in loglikelihood between maximum likelihood and dual estimates equals 4·29 in $\mathcal{N}(G_s)$ and 0·51 in $\mathcal{N}(G_d)$. The difference in loglikelihood between maximum and empirical likelihood estimates equals 20·54 in $\mathcal{N}(G_s)$ and 5·67 in $\mathcal{N}(G_d)$.

6·2. *Simulations*

Since the maximum likelihood estimator $\hat{\Sigma}_M$ and Kauermann's dual estimator $\hat{\Sigma}_D$ are based on a normality assumption, whereas the empirical-likelihood-based estimator $\hat{\Sigma}_E$ is not, it is interesting to compare their performance, both when the underlying distribution is, and is not, Gaussian. We simulated $M = 10\,000$ datasets for sample sizes $n = 10, 20, 25, 30, 50, 100$ from a multivariate normal distribution, a multivariate t distribution with 5 degrees of freedom, t_5, a noncentral distribution with 5 degrees of freedom, nct_5, and a standardized multivariate lognormal distribution. For each of these four distributions we chose parameter values resulting in the covariance matrix

$$\Sigma = (\sigma_{ij}) = \begin{pmatrix} 1 & 0 & 0{\cdot}375 & 0 \\ 0 & 1 & 0 & 0{\cdot}165 \\ 0{\cdot}375 & 0 & 1 & 0{\cdot}65 \\ 0 & 0{\cdot}165 & 0{\cdot}65 & 1 \end{pmatrix}, \tag{48}$$

which corresponds to the graph shown in Fig. 1. We remark that the skew-normal and skew-t distributions of Azzalini & Capitanio (1999, 2003) lead to sample covariance matrices that have the same distribution as their respective non-skewed versions (Gupta & Chang, 2003) and thus do not provide any additional insight for our problem.

For the multivariate normal distribution we used mean zero and the covariance matrix Σ from (48). For the t_5 distribution we chose mean zero and dispersion matrix $D = (3/5)\Sigma$ which yields covariance matrix Σ. Draws from a noncentral t_5 distibution were generated as $Y = \sqrt{5}Z/s$, where $Z \sim \mathcal{N}_4(\mu, D)$ and s^2 follows a χ^2_5 distribution (Kotz & Nadarajah, 2004). Note that all components of Z are divided by the same scalar s. The covariance matrix of the random vector Y can be shown to be equal to

$$\mathrm{var}(Y) = \frac{5}{3}D + \left(\frac{5}{3} - \frac{40}{9\pi}\right)\mu\mu'.$$

To obtain $\mathrm{var}(Y) = \Sigma$, we chose $\mu = (1, 1, 1, 1)'$ and the positive definite dispersion matrix

$$D = \frac{3}{5}\left\{\Sigma - \left(\frac{5}{3} - \frac{40}{9\pi}\right)\mu\mu'\right\}. \tag{49}$$

All the above distributions have support on $\mathbb{R}^4$. In order to investigate the performance of the estimators on a skewed distribution with restricted support we used a standardized multivariate lognormal distribution as described in Johnson (1987, Ch. 5). To be more precise, we considered

$$Y = \frac{e^Z - \sqrt{e}}{\sqrt{\{e(e-1)\}}}, \tag{50}$$

where $Z \sim \mathcal{N}_4(0, D)$ and the dispersion matrix D has entries $d_{ij} = \log\{1 + (e-1)\sigma_{ij}\}$, $1 \leqslant i, j \leqslant 4$. It then holds that Y has mean zero and covariance matrix Σ.

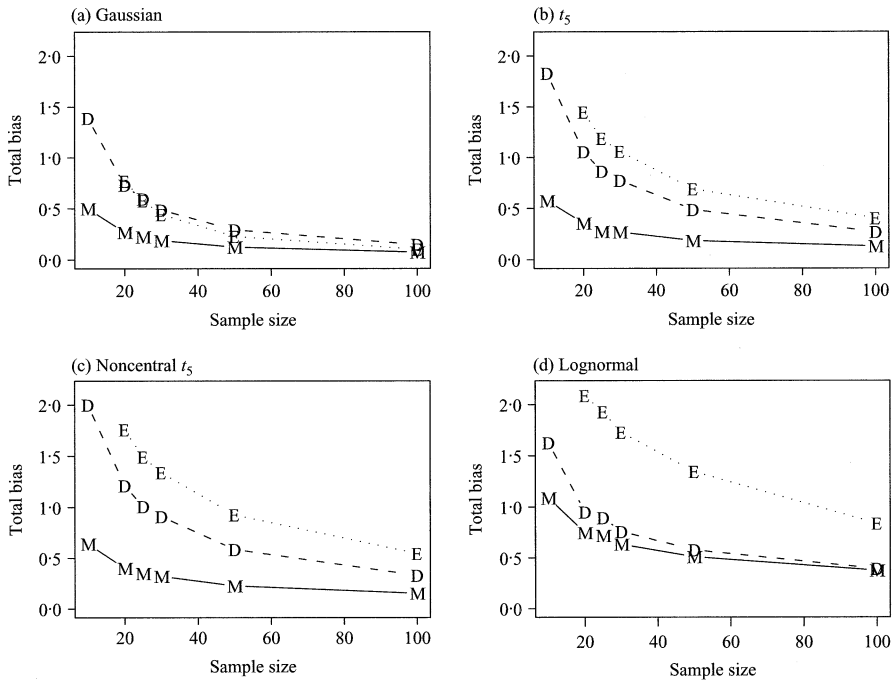

Fig. 5. Simulation study. Biases of maximum (M —), dual (D – –) and empirical (E. . .) likelihood estimators for various sample sizes, for cases of (a) Gaussian, (b) t_5, (c) noncentral t_5, (d) lognormal.

Figures 5 and 6 present our simulation results on bias and root mean squared error, respectively, for the three estimation methods. Note that, for both the t_5 and the nct_5 distributions, moments of up to fourth order exist. Let $m \in \{1, \ldots, M\}$ index the current simulation for sample size n, and let $\hat{\sigma}_{ij,\star}^{(m,n)}$ be the estimate of σ_{ij} using method $\star \in \{\mathrm{M},\mathrm{D},\mathrm{E}\}$. Then we measure the total bias,

$$\mathrm{bias}(n, \star) = \sum_{i \geqslant j} \left| \frac{1}{M} \sum_{m=1}^{M} \left(\hat{\sigma}_{ij,\star}^{(m,n)} - \sigma_{ij} \right) \right|, \tag{51}$$

and the total root mean squared error,

$$\mathrm{RMSE}(n, \star) = \sqrt{\left\{ \sum_{i \geqslant j} \frac{1}{M} \sum_{m=1}^{M} \left(\hat{\sigma}_{ij,\star}^{(m,n)} - \sigma_{ij} \right)^2 \right\}}. \tag{52}$$

The maximum likelihood estimates are taken to be the result of iterative conditional fitting started at the identity matrix. For the particular model considered here, the results in Drton & Richardson (2004) allow us to check for multimodality of the likelihood, but since it occurred only four times, all at sample size 10, there were at most four occasions on which iterative conditional fitting found a local, but not global, maximum.

For sample size $n = 10$ we experienced problems with the empirical likelihood procedure, resulting from an inability to find feasible starting values. Consequently, we do not report results for $\hat{\Sigma}_{\mathrm{E}}$ if $n = 10$. In Fig. 5, the biases of $\hat{\Sigma}_{\mathrm{E}}$ and of $\hat{\Sigma}_{\mathrm{D}}$ are larger than that of $\hat{\Sigma}_{\mathrm{M}}$ for all sample sizes. Moreover, the bias of $\hat{\Sigma}_{\mathrm{D}}$ is smaller than that of $\hat{\Sigma}_{\mathrm{E}}$, with the exception of the normal case and for $n \geqslant 25$. Under normality, as would be expected, the root mean squared error of the maximum likelihood estimator $\hat{\Sigma}_{\mathrm{M}}$ is lower than that of the

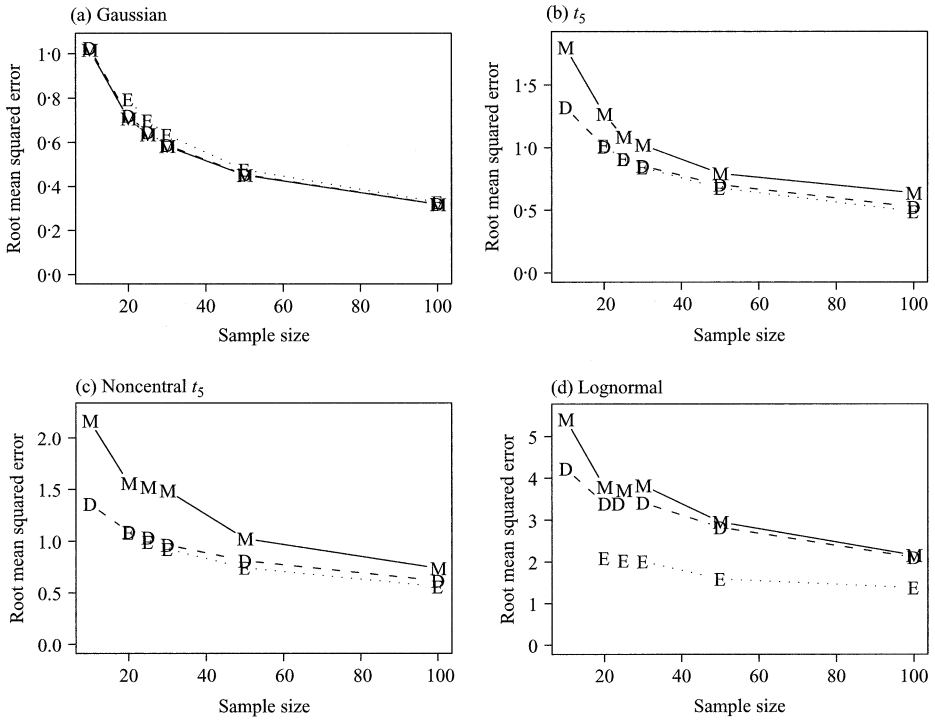

Fig. 6. Simulation study. Root mean squared errors of maximum (M —), dual (D – –) and empirical (E. . .) likelihood estimators for various sample sizes, for cases of (a) Gaussian, (b) t_5, (c) noncentral t_5, (d) lognormal. Note the different scales for the vertical axes.

also asymptotically efficient dual estimator $\hat{\Sigma}_D$ and the empirical likelihood estimator $\hat{\Sigma}_E$, but it is somewhat surprising how little efficiency is lost in this example when using $\hat{\Sigma}_E$; the three curves in Fig. 6(a) essentially overlap. Under the two t distributions, $\hat{\Sigma}_E$ and $\hat{\Sigma}_D$ outperform $\hat{\Sigma}_M$, the empirical likelihood approach faring best; see Fig. 6(b)–(c). Under the lognormal distribution, see Fig. 6(d), the root mean squared error of $\hat{\Sigma}_E$ is considerably smaller than the rather similar values for $\hat{\Sigma}_M$ and $\hat{\Sigma}_D$.

Figures 5 and 6 show averages over $M = 10\,000$ simulations. In order to assess the Monte Carlo error, we computed standard deviations of bias$(n, \star)$ and RMSE$(n, \star)$ by splitting the simulation runs into 100 batches of 100 simulations each. We divided the result by 10 to rescale the standard deviation estimate to the true simulation size M. The resulting standard deviations decreased with sample size, and for the bias they were always smaller than 0·04. For the root mean squared error the standard deviations were at most 0·05, with the exception of the lognormal distribution and sample size $n < 50$, for which the standard deviations of RMSE$(n, \star)$ were roughly 0·1.

7. DISCUSSION

It is very appealing that iterative conditional fitting extends the duality between covariance graph and undirected, concentration, graph models (Kauermann, 1996) to the level of fitting algorithms. The commonly used method for fitting undirected graph models, the iterative proportional fitting algorithm, fits marginal distributions while fixing conditionals (Whittaker, 1990, pp. 182–5); iterative conditional fitting does exactly the converse. When the idea behind iterative conditional fitting is expressed in terms of marginal and conditional

distributions, it is not limited in any way to Gaussian covariance graph models. In fact, work by the authors on applying iterative conditional fitting in binary graphical models for marginal independence appears promising.

The iterative conditional fitting algorithm resembles the Iterative Conditional Modes algorithm of Besag (1986), which, however, addresses a very different problem, namely maximum a posteriori estimation in Bayesian modelling. Another related algorithm is the Conditional Iterative Proportional Fitting algorithm of Cramer (1998, 2000), which can be used to maximize the likelihood function of a model that comprises joint distributions with prescribed conditional distributions. However, this algorithm differs from iterative conditional fitting because the update steps of iterative conditional fitting do not simply equate a conditional distribution to a prescribed conditional, but rather maximize a conditional likelihood function that will generally not be the same in two different iterations.

Algorithmically, empirical likelihood estimation is more complicated than maximum likelihood and dual estimation. While we used general empirical likelihood algorithms in this paper, we believe that attempting to devise special-case algorithms for empirical likelihood estimation of a covariance matrix with zeros constitutes a worthwhile topic for future research. Focusing on smaller sample sizes in such attempts may be particularly valuable, as in our simulations we had difficulties in obtaining empirical likelihood estimates for sample size $n = 10$ in a model for four variables. The issues with small sample size are, however, also related to a fundamental difference between empirical likelihood estimation and the normality-based methodology. Both maximum and dual likelihood estimation are possible if the empirical covariance matrix is positive definite, which occurs with probability one if the sample size is larger than the number of variables, and may occur for smaller sample sizes if the covariance graph is disconnected. In contrast, the optimization problem to be solved for empirical likelihood estimation may become infeasible if the sample size is small compared to the number of constraints imposed. The number of constraints depends on the covariance graph, and seemingly simpler sparser structures impose more constraints and render the empirical likelihood approach more sample-size-demanding.

ACKNOWLEDGEMENT

We thank Steffen Lauritzen for pointing out the duality between iterative conditional and iterative proportional fitting, and Art Owen for suggesting use of empirical likelihood. Sanjay Chaudhuri thanks Mark Handcock for helpful discussions. We also thank the editor and two anonymous referees for comments leading to improved presentation of the paper. This work was supported by grants from the U.S. National Science Foundation, the U.S. National Institutes of Health, the University of Washington Royalty Research Fund, the William and Flora Hewlett Foundation, the U.S. National Institute of Child Health and Human Development and the National University of Singapore.

REFERENCES

ANDERSON, T. W. (1958). *An Introduction to Multivariate Statistical Analysis*, 1st ed. New York: Wiley.

ANDERSON, T. W. (1969). Statistical inference for covariance matrices with linear structure. In *Multivariate Analysis, II*, Ed. P. R. Krishnaiah, pp. 55–56, New York: Academic Press.

ANDERSON, T. W. (1970). Estimation of covariance matrices which are linear combinations or whose inverses are linear combinations of given matrices. In *Essays in Probability and Statistics*, Eds. R. C. Bose & S. N. Roy, pp. 1–24, Chapel Hill: University of North Carolina Press.

ANDERSON, T. W. (1973). Asymptotically efficient estimation of covariance matrices with linear structure. *Ann. Statist.* **1**, 135–41.

ANDERSON, T. W. & OLKIN, I. (1985). Maximum-likelihood estimation of the parameters of a multivariate normal distribution. *Lin. Algeb. Applic.* **70**, 147–71.

AZZALINI, A. & CAPITANIO, A. (1999). Statistical applications of the multivariate skew normal distribution. *J. R. Statist. Soc.* B **61**, 579–602.

AZZALINI, A. & CAPITANIO, A. (2003). Distributions generated by perturbation of symmetry with emphasis on a multivariate skew t-distribution. *J. R. Statist. Soc.* B **65**, 367–89.

BESAG, J. (1986). On the statistical analysis of dirty pictures (with Discussion). *J. R. Statist. Soc.* B **48**, 259–302.

BUTTE, A. J., TAMAYO, P., SLONIM, D., GOLUB, T. R. & KOHANE, I. S. (2000). Discovering functional relationships between RNA expression and chemotherapeutic susceptibility using relevance networks. *Proc. Nat. Acad. Sci.* **97**, 12182–6.

COX, D. R. & WERMUTH, N. (1993). Linear dependencies represented by chain graphs (with Discussion). *Statist. Sci.* **8**, 204–18, 247–77.

COX, D. R. & WERMUTH, N. (1996). *Multivariate Dependencies: Models, Analysis and Interpretation*. London: Chapman and Hall.

CRAMER, E. (1998). Conditional iterative proportional fitting for Gaussian distributions. *J. Mult. Anal.* **65**, 261–76.

CRAMER, E. (2000). Probability measures with given marginals and conditionals: I-projections and conditional iterative proportional fitting. *Statist. Decis.* **18**, 311–29.

DRTON, M. (2006). Computing all roots of the likelihood equations of seemingly unrelated regressions. *J. Symb. Comp.* **41**, 245–54.

DRTON, M. & EICHLER, M. (2006). Maximum likelihood estimation in Gaussian chain graph models under the alternative Markov property. *Scand. J. Statist.* **33**, 247–57.

DRTON, M. & PERLMAN, M. D. (2004). Model selection for Gaussian concentration graphs. *Biometrika* **91**, 591–602.

DRTON, M. & RICHARDSON, T. S. (2003). A new algorithm for maximum likelihood estimation in Gaussian graphical models for marginal independence. In *Proceedings of the 19th Conference on Uncertainty in Artificial Intelligence*, Eds. U. Kjærulff & C. Meek, pp. 184–91, San Francisco, CA: Morgan Kaufmann.

DRTON, M. & RICHARDSON, T. S. (2004). Multimodality of the likelihood in the bivariate seemingly unrelated regressions model. *Biometrika* **91**, 383–92.

EATON, M. L. & PERLMAN, M. D. (1973). The non-singularity of generalized sample covariance matrices. *Ann. Statist.* **1**, 710–7.

EDWARDS, D. M. (2000). *Introduction to Graphical Modelling*, 2nd ed. New York: Springer-Verlag.

GASCH, A. P., SPELLMAN, P. T., KAO, C. M., CARMEL-HAREL, O., EISEN, M. B., STORZ, G., BOTSTEIN, D. & BROWN, P. O. (2000). Genomic expression programs in the response of yeast cells to environmental changes. *Mol. Biol. Cell* **11**, 4241–57.

GRZEBYK, M., WILD, P. & CHOUANIÈRE, D. (2004). On identification of multi-factor models with correlated residuals. *Biometrika* **91**, 141–51.

GUPTA, A. K. & CHANG, F.-C. (2003). Multivariate skew-symmetric distributions. *Appl. Math. Lett.* **16**, 643–6.

HELLERSTEIN, J. & IMBENS, G. W. (1999). Imposing moment restrictions from auxiliary data by weighting. *Rev. Econ. Statist.* **LXXXI**, 1–14.

JOHNSON, M. E. (1987). *Multivariate Statistical Simulation*. New York: Wiley.

KAUERMANN, G. (1996). On a dualization of graphical Gaussian models. *Scand. J. Statist.* **23**, 105–16.

KOTZ, S. & NADARAJAH, S. (2004). *Multivariate t Distributions and their Applications*. Cambridge: Cambridge University Press.

LAURITZEN, S. L. (1996). *Graphical Models*. Oxford: Clarendon Press.

MAO, Y., KSCHISCHANG, F. R. & FREY, B. J. (2004). Convolutional factor graphs as probabilistic models. In *Proceedings of the 20th Conference on Uncertainty in Artificial Intelligence*, Eds. M. Chickering & J. Halpern, pp. 374–81, Arlington, MA: AUAI Press.

OWEN, A. B. (2001). *Empirical Likelihood*. Boca Raton, FL: Chapman and Hall.

QIN, J. & LAWLESS, J. (1994). Empirical likelihood and general estimating equations. *Ann. Statist.* **22**, 300–25.

WERMUTH, N., COX, D. R. & MARCHETTI, G. M. (2006). Covariance Chains. *Bernoulli* **12**, 841–62.

WHITTAKER, J. (1990). *Graphical Models in Applied Multivariate Statistics*. Chichester: Wiley.

WRIGHT, S. (1921). Correlation and causation. *J. Agric. Res.* **20**, 557–85.

ZELLNER, A. (1962). An efficient method of estimating seemingly unrelated regression equations and tests for aggregation bias. *J. Am. Statist. Assoc.* **57**, 348–68.

[*Received August 2005. Revised June 2006*]