

# Iterative Conditional Fitting for Estimation of a Covariance Matrix with Zeros\*

Mathias Drton  
*University of California, Berkeley*

Thomas S. Richardson  
*University of Washington*

October 5, 2004

## Abstract

We consider estimation of the covariance matrix of a random vector under the constraint that certain elements in the covariance matrix are zero. Assuming that the random vector follows a multivariate normal distribution, in which case the model is also known as covariance graph model, we present a new algorithm for maximum likelihood estimation of the covariance matrix with zero pattern. We give our new algorithm the name Iterative Conditional Fitting since in each step of the procedure, a conditional distribution is estimated, subject to constraints, while a marginal distribution is held fixed. This approach is in duality to the well-known iterative proportional fitting algorithm, in which marginal distributions are fitted while conditional distributions are held fixed. We show that Iterative Conditional Fitting can be implemented using least squares computations and we establish the convergence properties of the algorithm.

## 1 Introduction

Consider a random vector  $X = (X_1, X_2, X_3, X_4)' \in \mathbb{R}^4$  that is distributed according to the multivariate normal distribution  $\mathcal{N}_4(0, \Sigma)$  with a covariance matrix  $\Sigma$  that exhibits the zero pattern

$$\Sigma = \begin{pmatrix} \sigma_{11} & 0 & \sigma_{13} & 0 \\ 0 & \sigma_{22} & 0 & \sigma_{24} \\ \sigma_{13} & 0 & \sigma_{33} & \sigma_{34} \\ 0 & \sigma_{24} & \sigma_{34} & \sigma_{44} \end{pmatrix} \in \mathbb{R}^{4 \times 4}. \quad (1.1)$$

Especially for larger covariance matrices it is helpful to visualize the pattern of zeros by a so-called covariance graph (Cox and Wermuth, 1993, 1996). A covariance graph

---

\*Keywords and Phrases: covariance graphs, covariance matrix, graphical models, marginal independence, maximum likelihood estimation, multivariate normal distribution.

has one vertex for each one of the random variables in the random vector. In the above example, the vertex set is  $V = \{1, 2, 3, 4\}$ , where the random variable  $X_i$  is identified with its index  $i$ . Next, each pair of vertices  $(i, j) \in V^2$  is connected by an edge unless  $\sigma_{ij} = 0$ . Assuming that the covariance matrix in (1.1) has no zeros other than those indicated explicitly, its covariance graph is given in Figure 1. Here we use bi-directed edges in keeping with the path diagram notation used by Wright (1921); other authors have used dashed edges (see Cox and Wermuth (1993, 1996) and discussion below).

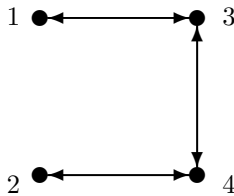


Figure 1: The covariance graph for the matrix in (1.1).

Associated with a covariance graph is the family of all multivariate normal distributions  $\mathcal{N}(0, \Sigma)$  such that  $\sigma_{ij} = 0$  whenever  $i \neq j$  and  $i \not\leftrightarrow j$ . Such a family is called a covariance graph model. Clearly,  $\sigma_{ij} = 0$  if and only if  $X_i$  and  $X_j$  are marginally independent; in symbols  $X_i \perp\!\!\!\perp X_j$ . Hence a covariance graph model is a graphical model based solely on marginal independence in contrast with graphical models based on undirected graphs (Markov random fields), directed acyclic graphs (DAGs, Bayesian networks), or chain graphs, where the absence of an edge between two vertices generally indicates some conditional independence between the associated variables (Edwards, 2000; Lauritzen, 1996; Whittaker, 1990).

Zero patterns in the covariance matrix are a special case of linear hypotheses on the covariance matrix, which were studied more generally in Anderson (1969, 1970, 1973). However, zero patterns form an important pattern of marginal independence and appear, for example, in the recent work by Grzebyk et al. (2004) and Mao et al. (2004). The use of bi-directed edges in covariance graphs is beneficial when the focus is on the independence interpretation since bi-directed edges make the connection to ancestral graph models (Richardson and Spirtes, 2002) explicit. In fact, a Gaussian ancestral graph model coincides with a covariance graph model if the ancestral graph exhibits only bi-directed edges. This establishes also a connection to path diagrams (e.g. Koster, 1999) and causality (Pearl, 2000; Pearl and Wermuth, 1994; Richardson and Spirtes, 2003; Spirtes et al., 2000). The fact that covariance graphs are a special

case of ancestral graphs can also be exploited to deduce all independence statements associated with a covariance graph via a graphical criterion,  $m$ -separation (Richardson and Spirtes, 2002), which naturally extends  $d$ -separation, (Pearl, 1988). Further details on the Markov, i.e. independence interpretation of bi-directed graphs can be found in Banerjee and Richardson (2003); Kauermann (1996); Richardson (2003).

For the traditional graphical models based on undirected graphs, DAGs and chain graphs, procedures for maximum likelihood (ML) estimation are well developed, and many methods are implemented, for example, in the software package MIM (Edwards, 2000). This is not the case, however, for covariance graph models. For instance, MIM does not permit ML estimation in covariance graph models but permits fitting only by a heuristic method due to Kauermann (1996), which is based on a “dual likelihood”; see also Edwards (2000, §7.4). There is, however, an algorithm due to Anderson (1969, 1970, 1973) that can be used to compute the ML estimate in models defined by linear hypotheses on covariance matrices, hence also in covariance graph models. However, this algorithm does not necessarily produce a positive (semi-) definite estimate and its convergence properties are unclear. In this paper, we introduce a new algorithm for ML estimation in covariance graph models, called Iterative Conditional Fitting (ICF), which does not suffer from the same problems as Anderson’s algorithm.

The paper is organized as follows. In Section 2 we give a formal definition of covariance graph models and derive the likelihood equations and the Fisher-information. In Section 3 we describe Anderson’s ML estimation algorithm and Kauermann’s dual estimation. In Section 4 we present our new ICF algorithm, which is based on univariate regressions. An alternate formulation based on multivariate regressions is described in Section 5. In Section 6 we compare ML and dual estimates for two example data sets. We conclude in Sections 7 and 8, where we discuss ICF and comment on related literature.

## 2 Covariance graph models

Suppose that we observe a set of variables  $V$  in the random vector  $X = (X_i \mid i \in V)' \in \mathbb{R}^V$  with multivariate normal joint distribution  $\mathcal{N}_V(0, \Sigma)$ , where  $\Sigma = (\sigma_{ij}) \in \mathbb{R}^{V \times V}$  is the unknown positive definite covariance matrix. Let  $G = (V, E)$  be a graph with the variable set  $V$  as vertex set and the edge set  $E \subseteq V \times V$  consisting exclusively of bi-directed edges  $(i, j)$ ,  $i \neq j$ ,  $i, j \in V$ , denoted by  $i \leftrightarrow j$ . Let  $\mathbf{P}(V)$  be the cone of all positive definite  $V \times V$  matrices and let  $\mathbf{P}(G)$  be the cone of all matrices  $\Sigma \in \mathbf{P}(V)$  which fulfill the linear restrictions

$$i \not\leftrightarrow j \implies \sigma_{ij} = 0. \tag{2.1}$$

The *covariance graph model* associated with the bi-directed graph  $G$  is the family of normal distributions

$$\mathbf{N}(G) = (\mathcal{N}_V(0, \Sigma) \mid \Sigma \in \mathbf{P}(G)). \quad (2.2)$$

Since for joint normal distributions the covariance  $\sigma_{ij} = 0$  if and only if the pairwise marginal independence  $X_i \perp\!\!\!\perp X_j$  holds, a covariance graph model is a graphical model for marginal independence.

This paper considers the estimation of the unknown parameter  $\Sigma$  based on a sample of i.i.d. observations  $X^{(k)} \in \mathbb{R}^V$ ,  $k \in N = \{1, \dots, n\}$ , from the covariance graph model (2.2). The set  $N$  can be interpreted as indexing the subjects on which we observe the variables in  $V$ . We group the vectors in the sample as columns in the  $V \times N$  random matrix  $Y$  which is distributed as

$$Y \in \mathbb{R}^{V \times N} \sim \mathcal{N}_{V \times N}(0, \Sigma \otimes I_N). \quad (2.3)$$

Here,  $I_N$  is the  $N \times N$  identity matrix,  $\Sigma \in \mathbf{P}(G)$  is the unknown positive definite covariance matrix, and  $\otimes$  is the Kronecker product. Thus the  $i$ -th row  $Y_i = Y_{i \cdot} \in \mathbb{R}^N$  of the matrix  $Y$  contains the i.i.d. observations for variable  $i \in V$  on all the subjects in  $N$  and the  $k$ -th column  $Y_{\cdot k} = X^{(k)}$  holds all the observations made on subject  $k \in N$ . Finally, the sample size is  $n = |N|$  and the number of variables is  $p = |V|$ .

Since our model assumes a zero mean, the empirical covariance matrix is defined to be

$$S = \frac{1}{n} Y Y' \in \mathbb{R}^{V \times V}. \quad (2.4)$$

We shall assume that  $n \geq p$  such that  $S$  is positive definite with probability one.

The case where the model also includes an unknown mean vector  $\mu$  can be treated by estimating  $\mu$  by the empirical mean vector  $\bar{Y} \in \mathbb{R}^V$ , i.e., the vector of the row means of  $Y$ . The empirical covariance matrix would then be the matrix

$$\tilde{S} = \frac{1}{n} (Y - \bar{Y} \otimes \mathbf{1}_N)(Y - \bar{Y} \otimes \mathbf{1}_N)' \in \mathbb{R}^{V \times V}, \quad (2.5)$$

where  $\mathbf{1}_N = (1, \dots, 1) \in \mathbb{R}^N$ . However, we would have to assume that  $n \geq p + 1$  to ensure that  $\tilde{S}$  is positive definite with probability one.

## 2.1 The likelihood function

The *log-likelihood function*  $\ell$  of the covariance graph model  $\mathbf{N}(G)$  is a function from  $\mathbf{P}(G)$  to  $\mathbb{R}$  and can be expressed as

$$\ell(\Sigma) = -\frac{np}{2} \log(2\pi) - \frac{n}{2} \log |\Sigma| - \frac{n}{2} \text{tr}(\Sigma^{-1} S), \quad (2.6)$$

see e.g. Edwards (2000, §3.1). If  $S$  is positive definite, then the global maximum of  $\ell(\Sigma)$  over  $\mathbf{P}(G)$ , i.e. the ML estimator of  $\Sigma$ , exists. In general, the condition  $n \geq p$  is not necessary for almost sure existence of the ML estimator but we are not aware of any results in the literature which provide a necessary and sufficient condition. In the sequel we will assume  $S$  to be positive definite (compare Buhl, 1993).

Besides existence, there is also the issue of uniqueness of the ML estimates, i.e. does the likelihood function have a unique local maximum which is then global. The model  $\mathbf{N}(G)$  is a curved but not regular exponential family, thus, the log-likelihood function need not be concave. In fact, the likelihood function can have multiple local maxima, i.e. the likelihood equations below can have several solutions. This follows from the examples in Drton (2004); Drton and Richardson (2004c). For a large enough sample size, a multimodal likelihood function seems not to arise often in practice assuming the model assumptions hold but might still arise if the model assumptions do not hold (see also Cox and Wermuth, 1996, p. 102f).

## 2.2 The likelihood equations

Let

$$F = \{(i, i) \mid i \in V\} \cup \{(i, j) \in V^2 \mid i < j \wedge i \leftrightarrow j\} \quad (2.7)$$

be the pairs of vertices indexing unrestricted elements in the matrix  $\Sigma \in \mathbf{P}(G)$ . The cardinality of  $F$  is equal to the number of vertices plus the number of edges in the graph  $G$ , i.e.  $|F| = |V| + |E| = p + |E|$ . The unrestricted elements of  $\Sigma$  form the vector

$$\sigma = (\sigma_{ij} \mid (i, j) \in F) \in \mathbb{R}^F. \quad (2.8)$$

In order to write derivatives of the log-likelihood function in compact form we introduce the matrix

$$Q = \frac{\partial \Sigma}{\partial \sigma} \in \mathbb{R}^{V^2 \times F}, \quad (2.9)$$

which satisfies  $\text{vec}(\Sigma) = Q\sigma$ , where  $\text{vec}$  is the operator of column-wise matrix vectorization. The entries of the matrix  $Q$  are either equal to one or equal to zero. The columns of  $Q$  that are associated with a variance  $\sigma_{ii}$  contain exactly one entry equal to one, whereas a column of  $Q$  that is associated with a covariance  $\sigma_{ij}$ ,  $i \neq j$ ,  $i \leftrightarrow j$ , contains exactly two entries equal to one. If the graph  $G$  is complete, i.e. all possible edges are present in  $G$ , then  $Q$  is the duplication matrix described in Harville (1997, p.352).

The first derivative of the log-likelihood function, that is, the *score function* can then be written as

$$\frac{\partial \ell(\Sigma)}{\partial \sigma} = \frac{1}{2} Q' [\text{vec}(\Sigma^{-1} S \Sigma^{-1}) - \text{vec}(\Sigma^{-1})], \quad (2.10)$$

see Harville (1997, §15) for details on the necessary matrix differential calculus. It follows that the *likelihood equations*  $\partial\ell(\Sigma)/\partial\sigma = 0$  are

$$(\Sigma^{-1})_{ij} = (\Sigma^{-1}S\Sigma^{-1})_{ij}, \quad (i, j) \in F; \quad (2.11)$$

compare also Anderson and Olkin (1985, §2.1.1). The full matrix  $\Sigma$  is determined by  $\sigma_{ij} = 0$  for  $(i, j) \notin F$ , that is for  $i \neq j$  and  $i \not\leftrightarrow j$ .

### 2.3 The Fisher-information

The second derivative of  $\ell(\Sigma)$  can be computed using results from Harville (1997, §15.9), and we find that the Hessian matrix equals

$$\frac{\partial^2\ell(\Sigma)}{\partial\sigma^2} = \frac{1}{2}Q'\{[\Sigma^{-1} \otimes \Sigma^{-1}] - [(\Sigma^{-1}S\Sigma^{-1}) \otimes \Sigma^{-1}] - [\Sigma^{-1} \otimes (\Sigma^{-1}S\Sigma^{-1})]\}Q. \quad (2.12)$$

Its negated expectation under  $\mathcal{N}_V(0, \Sigma)$ , the *Fisher-information*, equals

$$\frac{\partial^2\ell(\Sigma)}{\partial\sigma^2} = \frac{1}{2}Q'(\Sigma^{-1} \otimes \Sigma^{-1})Q \quad (2.13)$$

and can be used for the standard normal approximation to the distribution of roots of the likelihood equations. The computations of such roots will be the focus of the remainder of the paper.

## 3 Existing estimation methods/algorithms

We are aware of only one specialized algorithm for ML estimation applicable to covariance graph models. This algorithm is due to Anderson (1973) and will fit any Gaussian model obtained from a linear hypothesis on the covariance matrix (Anderson, 1969, 1970). In this section, we describe the incarnation of *Anderson's algorithm* that fits covariance graph models. We also review a *dual estimation* method due to Kauermann (1996), which produces estimates that are, in general, not solutions to the likelihood equations, yet are also asymptotically efficient, and unique. Note that Cox et al. (2004) have recently proposed moment based estimators in the special case where the graph is a chain, or equivalently, the covariance matrix is tri-diagonal under a suitable ordering.

### 3.1 Anderson's algorithm for ML estimation

The iterations in Anderson's algorithm consist of solving a system of linear equations built from the current estimate of  $\Sigma$ . In the case of a covariance graph model, the linear equations are solved for the vector  $\sigma$  of unrestricted elements in  $\Sigma$ , compare (2.8), and

can be specified as follows. Let  $\sigma^{ij} = (\Sigma^{-1})_{ij}$  and  $A = A(\Sigma)$  be the  $F \times F$  matrix with entries

$$A_{(ij,k\ell)} = \begin{cases} \sigma^{ik}\sigma^{jk} & \text{if } k = \ell, \\ \sigma^{ik}\sigma^{j\ell} + \sigma^{jk}\sigma^{i\ell} & \text{if } k \neq \ell. \end{cases} \quad (3.1)$$

Here  $(i, j)$  and  $(k, \ell)$  are elements of  $F$ . Furthermore, let  $b = b_\Sigma$  be the  $F \times 1$  vector with components

$$b_{ij} = (\Sigma^{-1}S\Sigma^{-1})_{ij}, \quad (i, j) \in F. \quad (3.2)$$

From Anderson (1973), it follows that  $\Sigma \in \mathbf{P}(G)$  solves  $A_\Sigma\sigma = b_\Sigma$  if and only if  $\Sigma$  solves the likelihood equations (2.11).

This motivates the following iterative scheme. Start with some  $\Sigma^{(0)} \in \mathbf{P}(G)$ . Iteratively, update the current estimate  $\Sigma^{(r)}$  to  $\Sigma^{(r+1)}$  determined by the linear equations

$$A_{\Sigma^{(r)}}\sigma^{(r+1)} = b_{\Sigma^{(r)}}. \quad (3.3)$$

A fixed point of this algorithm solves the likelihood equations (2.11). As starting value, Anderson suggests the identity matrix, i.e.  $\Sigma^{(0)} = I_V$ . In the first step, his algorithm constructs the empirical estimate  $\Sigma^{(1)}$  with  $\sigma_{ij}^{(1)} = S_{ij}$ ,  $(i, j) \in F$ . However, neither  $\Sigma^{(1)}$  nor any subsequent estimate of  $\Sigma$  has to be positive (semi-) definite and thus may not be a valid covariance matrix. Moreover, at any given stage, the likelihood function may decrease, and convergence of Anderson's algorithm cannot be guaranteed.

### 3.2 Kauermann's dual estimation

Dual estimation is based on the maximization of a dual likelihood function, which is motivated by interchanging the role of the parameter matrix  $\Sigma$  and the empirical covariance matrix (Kauermann, 1996, §4). Procedurally dual estimation can be performed in two steps Edwards (see also 2000, §7.4). First determine the ML estimator  $\hat{\Omega}$  of the covariance matrix in an undirected graphical model based on the inverted empirical covariance matrix  $S^{-1}$ , where the undirected graph has the same adjacencies as the covariance graph we are given. In other words we find the matrix  $\hat{\Omega}$  solves the equations

$$\hat{\Omega}_{ij} = (S^{-1})_{ij}, \quad \forall (i, j) \in F, \quad (3.4)$$

while satisfying that  $(\hat{\Omega}^{-1})_{ij} = 0$  for all  $(i, j) \notin F$ . Second, invert the estimate obtained,  $\hat{\Omega}$ , to give the dual estimator  $\hat{\Sigma}_{\text{dual}} = \hat{\Omega}^{-1} \in \mathbf{P}(G)$ .

Contrary to (2.11), the equation system (3.4) has a unique solution that can be found by the iterative proportional fitting algorithm (Speed and Kiiveri, 1986; Whittaker, 1990, pp.182–185). In particular, if the covariance graph is decomposable, then iterative proportional fitting will terminate in finitely many steps, and the dual estimator  $\hat{\Sigma}_{\text{dual}}$  is available in closed form.

## 4 Iterative conditional fitting

In this section, we present the new *Iterative Conditional Fitting* (ICF) algorithm for ML estimation, which is guaranteed to produce positive definite roots of the likelihood equations of covariance graph models, and for which convergence guarantees can be given. We begin by explaining the idea of iteratively fitting conditional distributions that stands behind ICF, and then show how the algorithm can be implemented using least squares computations. The algorithm, prior to being baptized ICF, was presented in Drton and Richardson (2003).

### 4.1 The idea of ICF

Starting with some initial estimate of the joint distribution, the idea of ICF is to repeatedly iterate through all vertices  $i \in V$ , and

- (i) Fix the marginal distribution for the variables different from  $i$ , i.e. the variables  $-i = V \setminus \{i\}$ ;
- (ii) Estimate, by maximum likelihood, the conditional distribution of variable  $i$  given the variables  $-i$  under the constraints implied by the covariance graph model  $\mathbf{N}(G)$ ;
- (iii) Find a new estimate of the joint distribution by multiplying together the fixed marginal and the estimated conditional distribution.

The fact that we fix the marginal distribution of variables  $-i$  in the update for variable  $i$  yields that all marginal independences amongst the variables  $-i$  still hold true after the update. Therefore, only the marginal independences involving variable  $i$  lead to constraints for the estimation in step (ii).

In order to make the idea more precise, let  $\Sigma_{A,B}$  denote the  $A \times B$  submatrix of  $\Sigma$  and  $Y_A$  denote the  $A \times N$  submatrix of  $Y$ , where  $A, B \subseteq V$ . Clearly,

$$Y_{-i} \sim \mathcal{N}_{-i \times N}(0, \Sigma_{-i, -i}). \quad (4.1)$$

Hence, step (i) consists of nothing but fixing the value of  $\Sigma_{-i, -i}$ , i.e. fixing everything but the  $i$ -th row and column of  $\Sigma$ . Not changing  $\Sigma_{-i, -i}$  in the  $i$ -th update means that many of the zero constraints imposed on the covariance matrix trivially hold true also after the update.

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

$$(Y_i \mid Y_{-i}) \sim \mathcal{N}_{\{i\} \times N}(B_i Y_{-i}, \lambda_i I_N), \quad (4.2)$$

where

$$B_i = \Sigma_{i, -i}(\Sigma_{-i, -i})^{-1} \in \mathbb{R}^{\{i\} \times -i} \quad (4.3)$$

is the  $\{i\} \times -i$  matrix of regression coefficients, and

$$\lambda_i = \sigma_{ii} - \Sigma_{i,-i}(\Sigma_{-i,-i})^{-1}\Sigma_{-i,i} \in (0, \infty) \quad (4.4)$$

is the conditional variance. If the graph  $G$  was the complete graph  $\bar{G}$  in which an edge joins any pair of vertices then the mapping

$$\begin{aligned} \mathbf{P}(\bar{G}) = \mathbf{P}(V) &\rightarrow (0, \infty) \times \mathbb{R}^{\{i\} \times -i} \times \mathbf{P}_{-i}(\bar{G}), \\ \Sigma &\mapsto (\lambda_i, B_i, \Sigma_{-i,-i}) \end{aligned} \quad (4.5)$$

would be bijective and the regression in (4.2) a standard least squares regression. Here,  $\mathbf{P}_A(G)$  is the set of all  $A \times A$  submatrices of matrices in  $\mathbf{P}(G)$ ,  $A \subseteq V$ . For a general graph  $G$ , (4.5) is no longer true and (4.2) is not a standard regression because we need to respect the restriction  $\Sigma \in \mathbf{P}(G)$ , i.e. the restrictions  $\sigma_{ij} = 0$  if  $j \in -i$ ,  $j \not\leftrightarrow i$ . In Drton and Richardson (2003) we translate the restrictions of covariances being zero into linear restrictions on the entries of  $B_i$ . The fact that these restrictions are linear allow for estimation of  $B_i$  and  $\lambda_i$  via least squares computations. However, these computations involve synthetic *pseudo-variables* that are computed from the data  $Y_{-i}$  and the fixed matrix  $\Sigma_{-i,-i}$ . In this paper, we take a slightly different approach in which it becomes more transparent where sparsity of the graph can be employed; see (4.7) where a possibly sparse matrix has to be inverted. However, it should be noted that the difference simply consists in a different method for computing the same quantity, namely the new estimate of the  $i$ -th row and column of  $\Sigma$ , is computed in the  $i$ -th update step of the ICF algorithm.

## 4.2 Pseudo-variable regressions

Instead of working with the regressions coefficients  $B_i$ , we exploit the fact that  $B_i$  equals  $\Sigma_{i,-i}$  multiplied by the inverse of the fixed submatrix  $\Sigma_{-i,-i}$ . Let  $\text{sp}(i) = \{j \mid i \leftrightarrow j\}$  be the set of *spouses* of  $i \in V$  and let  $\text{nsp}(i) = V \setminus (\text{sp}(i) \cup \{i\})$  be the set of *non-spouses*, yielding the partition  $V = \{i\} \cup \text{sp}(i) \cup \text{nsp}(i)$ . Then the conditional expectation of  $(Y_i \mid Y_{-i})$  can be written as

$$\mathbb{E}[Y_i \mid Y_{-i}] = \Sigma_{i,-i} [(\Sigma_{-i,-i})^{-1} Y_{-i}] = \Sigma_{i,\text{sp}(i)} Z_{\text{sp}(i)}^{(i)} = \sum_{j \in \text{sp}(i)} \sigma_{ij} Z_j^{(i)}, \quad (4.6)$$

where the *pseudo-variable*  $Z_j^{(i)}$  is equal to the  $j$ -th row in

$$Z_{\text{sp}(i)}^{(i)} = [(\Sigma_{-i,-i})^{-1}]_{\text{sp}(i),-i} Y_{-i} \in \mathbb{R}^{\text{sp}(i) \times N}. \quad (4.7)$$

In (4.6), we exploit that  $\sigma_{ij} = 0$  if  $j \in \text{nsp}(i)$ . From (4.6), we obtain that when fixing  $\Sigma_{-i,-i}$ ,

$$(Y_i \mid Y_{-i}) \sim \mathcal{N}_{\{i\} \times N} \left( \Sigma_{i,\text{sp}(i)} Z_{\text{sp}(i)}^{(i)}, \lambda_i I_N \right) = \mathcal{N}_{\{i\} \times N} \left( \sum_{j \in \text{sp}(i)} \sigma_{ij} Z_j^{(i)}, \lambda_i I_N \right). \quad (4.8)$$

Let  $\mathbf{P}_{-i}(G)$  be the set of  $-i \times -i$  submatrices of the matrices in  $\mathbf{P}(G)$ . Then the mapping

$$\begin{aligned} \mathbf{P}(G) &\rightarrow (0, \infty) \times \mathbb{R}^{\{i\} \times \text{sp}(i)} \times \mathbf{P}_{-i}(G) \\ \Sigma &\mapsto (\lambda_i, \Sigma_{i, \text{sp}(i)}, \Sigma_{-i, -i}) \end{aligned} \quad (4.9)$$

is a bijection, which implies that the parameters  $\sigma_{ij}$ ,  $j \in \text{sp}(i)$ , and  $\lambda_i$  can vary freely in (4.8). Therefore, (4.8) constitutes a standard normal regression model whose parameters  $\sigma_{ij}$ ,  $j \in \text{sp}(i)$ , and  $\lambda_i$  can be estimated by the usual least squares formula. The estimate of  $\lambda_i$  yields an estimate of  $\sigma_{ii}$  by solving (4.4) for  $\sigma_{ii}$ . Thus, we obtain the ML estimator of the  $i$ -th row and column of  $\Sigma$  when  $\Sigma_{-i, -i}$  is fixed to equal some given matrix in  $\mathbf{P}_{-i}(G)$ .

### 4.3 The ICF algorithm

Let  $\hat{\Sigma}^{(r)}$  be the estimate of  $\Sigma$  after the  $r$ -th iteration and  $\hat{\Sigma}^{(r,i)}$  the estimate of  $\Sigma$  after the  $i$ -th update step of the  $r$ -th iteration in ICF, i.e. after estimating  $(Y_i | Y_{-i})$ .

**Algorithm 1.** *The ICF algorithm can be implemented as:*

1. (Initialization) *Set the iteration counter  $r = 0$ , and choose a starting value  $\hat{\Sigma}^{(0)} \in \mathbf{P}(G)$ , e.g. the identity matrix  $\hat{\Sigma}^{(0)} = I_V$ .*
2. (Updates) *Order the variables as  $V = \{1, \dots, p\}$ , set  $\hat{\Sigma}^{(r,0)} = \hat{\Sigma}^{(r)}$ , and repeat the following steps for all  $i = 1, \dots, p$ :*
  - (i) *Let  $\hat{\Sigma}_{-i, -i}^{(r,i)} = \hat{\Sigma}_{-i, -i}^{(r,i-1)}$  and calculate from this submatrix the pseudo-variables  $Z_{\text{sp}(i)}^{(i)}$  according to (4.7).*
  - (ii) *Compute the ML estimators*

$$\begin{aligned} \hat{\Sigma}_{i, \text{sp}(i)}^{(r,i)} &= Y_i (Z_{\text{sp}(i)}^{(i)})' [Z_{\text{sp}(i)}^{(i)} (Z_{\text{sp}(i)}^{(i)})']^{-1}, \\ \hat{\lambda}_i &= \frac{1}{n} (Y_i - \hat{\Sigma}_{i, \text{sp}(i)}^{(r,i)} Z_{\text{sp}(i)}^{(i)}) (Y_i - \hat{\Sigma}_{i, \text{sp}(i)}^{(r,i)} Z_{\text{sp}(i)}^{(i)})'. \end{aligned} \quad (4.10)$$

*for the linear regression (4.8).*

- (iii) *Complete  $\hat{\Sigma}^{(r,i)}$  by setting*

$$\hat{\sigma}_{ii}^{(r,i)} = \hat{\lambda}_i + \hat{\Sigma}_{i, \text{sp}(i)}^{(r,i)} [(\hat{\Sigma}_{-i, -i}^{(r,i)})^{-1}]_{\text{sp}(i), \text{sp}(i)} \hat{\Sigma}_{\text{sp}(i), i}^{(r,i)}; \quad (4.11)$$

*compare (4.4).*

3. (Repeat) *Set  $\hat{\Sigma}^{(r+1)} = \hat{\Sigma}^{(r,p)}$ . Increment the counter  $r$  to  $r + 1$ . Go to 2.*

The iterations can be stopped according to a criterion such as “the estimate of  $\Sigma$  is not changed” (in some pre-determined accuracy).

**Remark 2 (Complexity).** The new algorithm can be restated only in terms of the empirical covariance matrix  $S$  defined in (2.4). For example in (4.10),

$$\begin{aligned} Y_i(Z_{\text{sp}(i)}^{(i)})' &= S_{i,-i}[(\Sigma_{-i,-i})^{-1}]_{-i,\text{sp}(i)}, \\ Z_{\text{sp}(i)}^{(i)}(Z_{\text{sp}(i)}^{(i)})' &= [(\Sigma_{-i,-i})^{-1}]_{\text{sp}(i),-i} S_{-i,-i}[(\Sigma_{-i,-i})^{-1}]_{-i,\text{sp}(i)}. \end{aligned} \quad (4.12)$$

Other products between data matrices appearing in the sequel can be similarly expressed in terms of the empirical covariance matrix  $S$ . Thus, the 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 (4.7), and the inversion of a matrix of size  $\text{sp}(i) \times \text{sp}(i)$  (4.10). Note that  $\Sigma_{-i,-i}$  may be sparse and special methods for inversion of sparse matrices might be useful. In particular, if the induced subgraph  $G_{-i}$  has disconnected components then only the submatrices of  $\Sigma$  over connected components containing spouses of  $i$  have to be inverted.

#### 4.4 Convergence

The key to prove convergence of ICF is to recognize that the algorithm consists of iterated partial maximizations over sections of the parameter space  $\mathbf{P}(G)$ . In ICF we repeatedly maximize the likelihood function of the covariance graph model partially by allowing only the entries in the  $i$ -th row and column of  $\Sigma$  to vary. The remaining entries are fixed. A bit more formally, we consider the parameter space

$$\Theta = \{\Sigma \in \mathbf{P}(G) \mid \ell(\Sigma) \geq \ell(\hat{\Sigma}^{(0)})\}, \quad (4.13)$$

which is compact, though not necessarily connected, and contains the global maximizer of  $\ell(\Sigma)$ . Recall that we assume the empirical covariance matrix  $S$  to be positive definite. Defining the section  $\Theta_i(\bar{\Sigma}) \subseteq \Theta$  as

$$\Theta_i(\bar{\Sigma}) = \{\Sigma \in \Theta \mid \Sigma_{-i,-i} = \bar{\Sigma}_{-i,-i}\}, \quad (4.14)$$

it becomes clear that the algorithm steps 2(i)-2(iii) maximize the log-likelihood function partially over the section  $\Theta_i(\hat{\Sigma}^{(r,i-1)})$ , i.e.

$$\hat{\Sigma}^{(r,i)} = \arg \max \{\ell(\Sigma) \mid \Sigma \in \Theta_i(\hat{\Sigma}^{(r,i-1)})\}. \quad (4.15)$$

This local and global maximizer over the section is unique. If a matrix  $\Sigma \in \mathbf{P}(G)$  maximizes the log-likelihood function over all sections  $\Theta_i(\Sigma)$ ,  $i \in V$ , simultaneously, then it solves the likelihood equations. Hence, the following theorem follows from results in Drton and Eichler (2004, Appendix).

**Theorem 3.** *Suppose the sequence  $(\hat{\Sigma}^{(r)})$  is constructed by the ICF algorithm. Then all accumulation points of  $(\hat{\Sigma}^{(r)})$  are saddle points or local maxima of the log-likelihood function. Moreover, all accumulation points have the same likelihood value. In particular, if the likelihood equations have only finitely many solutions, then  $(\hat{\Sigma}^{(r)})$  converges.*

In practice, the finite accuracy in computer calculations seems to prevent convergence to a saddle point.

### 4.5 Example

Figure 2 illustrates ICF for the covariance graph model based on the graph shown in Figure 1. The algorithm cycles in arbitrary order through the four regressions  $(Y_i | Y_{-i})$ ,  $i = 1, 2, 3, 4$ . In Figure 2, a filled circle represents variables in the conditioning set  $-i$ , and an unfilled circle stands for the variable  $i$  forming the response variable in the considered regression. The directed edges coincide with bi-directed edges in the original graph in Figure 1 and indicate the pseudo-variable regressions to be carried out. The vertices that are joined to vertex  $i$  by a directed edges are labelled with the pseudo-variables that act as covariates. The directed edges are labelled with the covariances that are estimated.

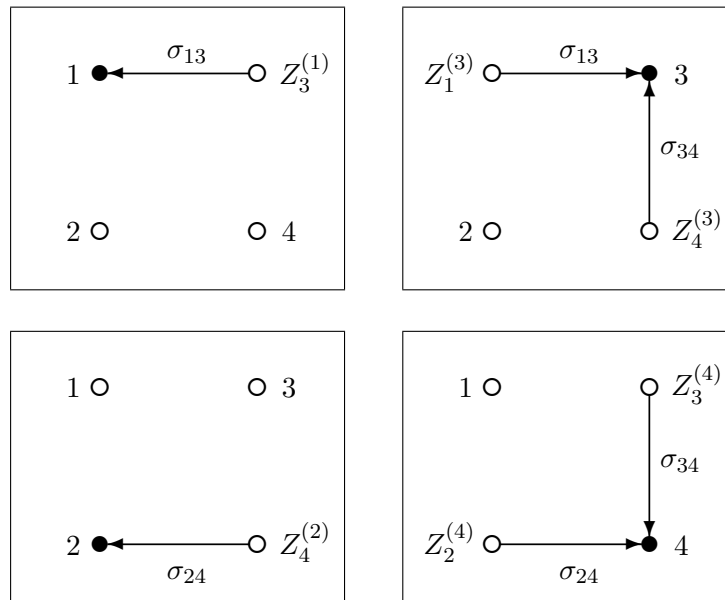


Figure 2: Illustration of the pseudo-variable regressions in ICF.

## 5 Iterative conditional fitting with multivariate updates

ICF, as presented in Section 4, is based on updating one row and column of an estimate of the covariance matrix  $\Sigma \in \mathbf{P}(G)$  by carrying out a univariate regression. A natural modification of this approach is to try to update several rows and columns of the estimate  $\Sigma \in \mathbf{P}(G)$  simultaneously using multivariate regression.

### 5.1 Seemingly unrelated pseudo-variable regressions

Let  $C \subseteq V$  be a subset of the vertices. In order to estimate all rows and columns of  $\Sigma$  that are indexed by a vertex in  $C$  in the ICF algorithm presented in Section 4, we have to carry out several univariate pseudo-variable regressions for  $(Y_i | Y_{-i})$ ,  $i \in C$ . Instead, we would like to consider only one multivariate regression of the form  $(Y_C | Y_{-C})$ , where  $-C = V \setminus C$ . The conditional distribution

$$(Y_C | Y_{-C}) \sim \mathcal{N}_{C \times N}(B_C Y_{-C}, \Lambda_C \otimes I_N) \quad (5.1)$$

is specified by the matrix of regression coefficients

$$B_C = \Sigma_{C,-C}(\Sigma_{-C,-C})^{-1} \in \mathbb{R}^{C \times -C}, \quad (5.2)$$

and the conditional covariance matrix

$$\Lambda_C = \Sigma_{C,C} - \Sigma_{C,-C}(\Sigma_{-C,-C})^{-1}\Sigma_{-C,C} \in \mathbf{P}(C). \quad (5.3)$$

In order for the conditional distribution (5.1) to be of a simple structure, there should be no constraints on the  $\Lambda_C$ , in which case  $\mathbf{P}(C) = \mathbf{P}_C(G)$ . This holds if there are no constraints on the submatrix  $\Sigma_{C,C}$ , which in turn holds if the set  $C$  is complete, i.e. if  $i \leftrightarrow j$  whenever  $i, j \in C$  and  $i \neq j$ . Then the only constraints on the conditional distribution (5.1) are on the matrix of regression coefficients  $B_C$  and stem from restrictions that  $\sigma_{ij} = 0$ , if  $i \in C$ ,  $j \neq i$  and  $j \not\leftrightarrow i$ .

Let

$$\text{sp}(C) = [\cup(\text{sp}(i) | i \in C)] \setminus C \quad (5.4)$$

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

$$Z_{\text{sp}(C)}^{(C)} = [(\Sigma_{-C,-C})^{-1}]_{\text{sp}(C),-C} Y_{-C} \in \mathbb{R}^{\text{sp}(C) \times N}, \quad (5.5)$$

then we can rewrite (5.1) as

$$(Y_C | Y_{-C}) \sim \mathcal{N}_{C \times N}(\Sigma_{C,\text{sp}(C)} Z_{\text{sp}(C)}^{(C)}, \Lambda_C \otimes I_N), \quad (5.6)$$

because  $\Sigma_{C, \text{ns}p(C)} = 0$ . As  $\Sigma$  ranges through  $\mathbf{P}(G)$ , the submatrix  $\Sigma_{C, \text{sp}(C)}$  playing the role of regression coefficients in (5.6) ranges through the linear space

$$\mathbf{P}_{C, \text{sp}(C)}(G) = \{A \in \mathbb{R}^{C \times \text{sp}(C)} \mid A_{ij} = 0 \text{ if } i \not\leftrightarrow j\}. \quad (5.7)$$

Hence, (5.6) constitutes seemingly unrelated regressions (Zellner, 1962).

## 5.2 The ICF with multivariate updates

ML estimation in seemingly unrelated regressions itself generally requires iterative algorithms, such as iterating the two-step estimator of Zellner (1962). In the case of (5.6), the two-step estimator consists of first estimating  $\Sigma_{C, \text{sp}(C)}$  for some fixed  $\Lambda_C$  by generalized least squares, and then estimating  $\Lambda_C$  as the empirical covariance matrix of the residuals  $Y_i - \Sigma_{C, \text{sp}(C)} Z_{\text{sp}(C)}^{(C)}$  computed with the estimate of  $\Sigma_{C, \text{sp}(C)}$  obtained in the first step. However, if the current estimate of  $\Sigma$  is used to obtain starting values  $\Sigma_{C, \text{sp}(C)}$  and  $\Lambda_C$ , then the two-step method does not have to be iterated in order to obtain estimates for the seemingly unrelated pseudo-regressions (5.6) that yield a convergent ICF algorithm with multivariate updates. For specification of the estimator of  $\Sigma_{C, \text{sp}(C)}$  we need to introduce the matrix  $P_C$  of the linear map that maps the vector of unrestricted elements in  $\Sigma_{C, \text{sp}(C)}$  to the matrix  $\Sigma_{C, \text{sp}(C)} \in \mathbf{P}_{C, \text{sp}(C)}(G)$ . The vector of unrestricted elements of  $\Sigma_{C, \text{sp}(C)}$  is the vector  $\sigma_C = (\sigma_{ij} \mid i \in C, j \in \text{sp}(C), i \leftrightarrow j)$ . The matrix  $P_C$  has exactly one entry equal to one in each column, the other entries are zero, and it satisfies  $\text{vec}(\Sigma_{C, \text{sp}(C)}) = P_C \sigma_C$ ; compare (2.9).

In order to run ICF with multivariate updates, we have to choose a family of complete sets  $(C \mid C \in \mathcal{C})$  such that

$$\cup(C \mid C \in \mathcal{C}) = V, \quad (5.8)$$

where the sets  $C$  do not have to be disjoint. 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 4.** *For a given choice of  $\mathcal{C}$ , the ICF algorithm with multivariate updates can be implemented as:*

1. (Initialization) *Set the iteration counter  $r = 0$ , and choose a starting value  $\hat{\Sigma}^{(0)} \in \mathbf{P}(G)$ , e.g. the identity matrix  $\hat{\Sigma}^{(0)} = I_V$ .*
2. (Updates) *Order the sets in the family  $\mathcal{C}$  as  $\mathcal{C} = \{C_1, \dots, C_q\}$ , set  $\hat{\Sigma}^{(r,0)} = \hat{\Sigma}^{(r)}$ , and repeat the following steps for all  $C_k \in \mathcal{C}$ :*
  - (i) *Let  $\hat{\Sigma}_{-C_k, -C_k}^{(r,k)} = \hat{\Sigma}_{-C_k, -C_k}^{(r,k-1)}$ . From this submatrix, compute the conditional covariance matrix  $\hat{\Lambda}_{C_k}$  according to (5.3) and the pseudo-variables  $Z_{\text{sp}(C_k)}^{(k)}$  according to (5.5). Calculate  $\hat{\Omega}_{C_k} = (\hat{\Lambda}_{C_k})^{-1}$ .*

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

$$\hat{\sigma}_{C_k} = \left\{ P_{C_k}' \left\{ [Z_{\text{sp}(C_k)}^{(k)} (Z_{\text{sp}(C_k)}^{(k)})'] \otimes \hat{\Omega}_{C_k} \right\} P_{C_k} \right\}^{-1} \times \left\{ P_{C_k}' \text{vec}[\hat{\Omega}_{C_k} Y_C (Z_{\text{sp}(C_k)}^{(k)})'] \right\}. \quad (5.9)$$

(iii) Compute the empirical covariance matrix of residuals

$$\hat{\Lambda}_{C_k} = \frac{1}{n} (Y_{C_k} - \hat{\Sigma}_{C_k, \text{sp}(C_k)} Z_{\text{sp}(C_k)}^{(k)}) (Y_{C_k} - \hat{\Sigma}_{C_k, \text{sp}(C_k)} Z_{\text{sp}(C_k)}^{(k)})'. \quad (5.10)$$

(iii) Complete  $\hat{\Sigma}^{(r,k)}$  by setting

$$\hat{\Sigma}_{C_k, C_k}^{(r,k)} = \hat{\Lambda}_{C_k} + \hat{\Sigma}_{C_k, \text{sp}(C_k)}^{(r,k)} [(\hat{\Sigma}_{-C_k, -C_k}^{(r,k)})^{-1}]_{\text{sp}(C_k), \text{sp}(C_k)} \hat{\Sigma}_{\text{sp}(C_k), C_k}^{(r,k)}; \quad (5.11)$$

compare (5.3).

3. (Repeat) Set  $\hat{\Sigma}^{(r+1)} = \hat{\Sigma}^{(r,q)}$ . Increment the counter  $r$  to  $r + 1$ . Go to 2.

### 5.3 Convergence

The ICF algorithm with multivariate updates is still an iterative partial maximization algorithm. However, the sections in the parameter space over which maximizations are performed are not quite as simple as the sections described in Section 4.4. Steps 2(ii) and 2(iii) of Algorithm 4 do not jointly maximize the log-likelihood function  $\ell$  over sections of the form

$$\Theta_C(\bar{\Sigma}) = \{\Sigma \in \Theta \mid \Sigma_{-C, -C} = \bar{\Sigma}_{-C, -C}\}. \quad (5.12)$$

Instead step 2(ii) maximizes  $\ell$  over sections of the form

$$\Theta_{1,C}(\bar{\Sigma}) = \{\Sigma \in \Theta \mid \Sigma_{-C, -C} = \bar{\Sigma}_{-C, -C}, \Lambda_C = \bar{\Lambda}_C\}, \quad (5.13)$$

where  $\Lambda_C$  is again the conditional covariance matrix from (5.3). The subsequent step 2(iii) maximizes  $\ell$  over sections of the form

$$\Theta_{2,C}(\bar{\Sigma}) = \{\Sigma \in \Theta \mid \Sigma_{-C, -C} = \bar{\Sigma}_{-C, -C}, \Sigma_{C, -C} = \bar{\Sigma}_{C, -C}\}. \quad (5.14)$$

Nevertheless it holds under condition (5.8) that if  $\Sigma$  maximizes the log-likelihood function  $\ell$  over both section  $\Theta_{1,C}(\bar{\Sigma})$  and  $\Theta_{2,C}(\bar{\Sigma})$  simultaneously for all  $C \in \mathcal{C}$ , then  $\Sigma$  is a solution to the likelihood equations. Thus, Theorem 3 holds also for ICF with multivariate updates as stated in Algorithm 4.

## 5.4 Example

Take up the covariance graph shown in Figure 1. For the family  $\mathcal{C}$  of complete vertex sets, several choices are possible. If we choose all singletons  $\mathcal{C} = \{1, 2, 3, 4\}$ , then Algorithm 4 reduces to Algorithm 1. If the cliques  $\mathcal{C} = \{13, 34, 24\}$  are chosen, then all conditional distributions considered in ICF are bivariate, whereas for  $\mathcal{C} = \{1, 2, 34\}$  two univariate distributions are estimated in conjunction with a bivariate distribution. For the clique choice  $\mathcal{C} = \{13, 34, 24\}$ , we illustrate the seemingly unrelated pseudo-variable regressions to be estimated in Figure 3, which is to be interpreted similarly as Figure 2. An additional feature are the bi-directed edges that connect the vertices in the sets  $C \in \mathcal{C}$  (see Richardson and Spirtes (2002) for a formal definition of these graphs.)

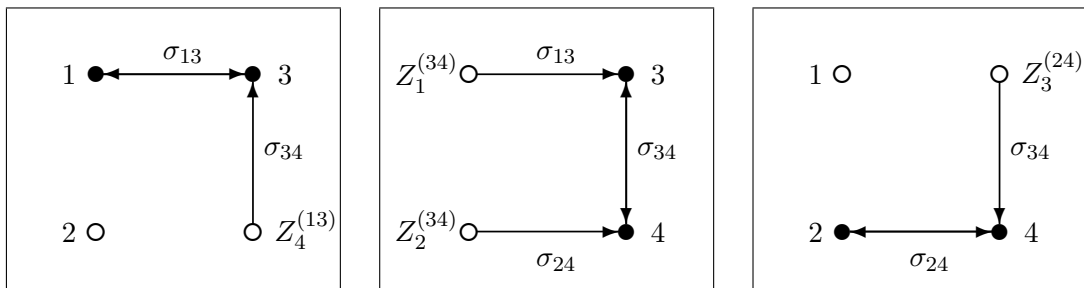


Figure 3: Illustration of the seemingly unrelated pseudo-variable regressions in ICF with multivariate updates.

## 6 Examples: dual versus maximum likelihood estimation

We now use ICF to compute ML estimates in covariance graph models using two example data sets from the literature. The estimates are compared to those from the dual estimation procedure described in Section 3.2.

### 6.1 Diabetes

Table 1 presents data on  $p = 4$  variables measured on  $n = 39$  diabetes patients; see Cox and Wermuth (1993, Table 7) and Kauermann (1996, Table 1). If we index the variables in this data set by  $W = 1$ ,  $V = 2$ ,  $X = 3$ , and  $Y = 4$  then the covariance graph model fitted by Kauermann is the one illustrated in Figure 1. We compute ML estimates for the covariance graph model  $\mathbf{N}(G)$  with  $G$  from Figure 1 by ICF started at the identity matrix. Anderson's algorithm converged to the same estimate. Results from Drton and Richardson (2003) can guarantee for this particular covariance graph  $G$

Table 1: Observed marginal correlations and standard deviations.

	$W$	$V$	$X$	$Y$
$V$	0.060			
$X$	-0.460	0.042		
$Y$	-0.071	-0.404	-0.334	
SD	5.72	92.00	7.86	2.07

that the likelihood function of  $\mathbf{N}(G)$  is unimodal for the considered dataset, and thus that the solution to the likelihood equation found by ICF is the global maximum

Table 2 shows that the ML and the dual estimates are very similar in this example. The difference in log-likelihood between the ML and the dual estimator is negligible, equaling 0.005. The deviance under comparison to the covariance graph model based

Table 2: Marginal correlations and standard deviations from ML (lower half & 6th row) and dual estimation (upper half & 5th row).

ML\dual	$W$	$V$	$X$	$Y$
$W$		0	-0.478	0
$V$	0		0	-0.374
$X$	-0.475	0		-0.341
$Y$	0	-0.378	-0.342	
SD <sub>dual</sub>	5.70	91.6	7.92	2.04
SD <sub>ML</sub>	5.72	92.0	7.93	2.05

on the complete graph equals 0.49, which is very small compared to the three degrees of freedom and indicates a very good fit. Note that the dual estimates we give in Table 2, which were computed using the R package ‘ggm’ and confirmed using MIM Edwards (2000), differ slightly from those stated in Kauermann (1996, Table 1(b)).

## 6.2 HIV

Roverato and Whittaker (1996) present data on six blood measurements relevant to HIV that were made on  $n = 107$  babies. The variables are immunoglobulin G ( $G$ ), immunoglobulin A ( $A$ ), lymphocyte B ( $B$ ), platelet count ( $P$ ), lymphocyte T4 ( $T$ ), T4/T8 lymphocyte ratio ( $R$ ). We state the observed marginal correlations and standard deviations in Table 3.

Using SIN model selection for covariance graphs Drton and Perlman (2004, 2005) as implemented in the R package ‘SIN’, which also contains the data from Table 3, we select, using the simultaneous confidence level 0.1, the covariance graph  $G_a$  shown in

Table 3: Observed marginal correlations and standard deviations.

	$G$	$A$	$B$	$P$	$T$	$R$
$A$	0.483					
$B$	0.220	0.057				
$P$	-0.034	-0.133	0.149			
$T$	0.253	-0.124	0.523	0.179		
$R$	-0.276	-0.314	-0.183	0.064	0.213	
SD	2.97	0.44	2987.35	142.80	1397.42	1.17

Figure 4(a). The ML and the dual estimates for  $\Sigma$  in the model  $\mathbf{N}(G_a)$  are stated in Table 4. We computed the ML estimates with ICF starting from the identity matrix; Anderson’s algorithm gave the same result. The difference in log-likelihood between the ML and the dual estimator is now considerable, equaling 4.81. The deviance under comparison to the covariance graph model based on the complete graph equals 28.87 over 10 degrees of freedom, indicating a rather poor fit.

If we use a less strict simultaneous confidence level in SIN model selection, such as 0.25, then we select the graph  $G_b$  shown in Figure 4(b). This graph has two additional edges. The model  $\mathbf{N}(G_b)$  fits the data better with a deviance of 13.15 over 8 degrees of freedom. Just as in the diabetes example from Section 6.1, where the model also fit well, the ML and the dual estimates are much more similar; compare Table 5. The difference in log-likelihood between ML and dual estimates is reduced to 0.72. Finally, we emphasize that although we refer to “ML estimates”, ICF is not guaranteed to find the global maximizer of the likelihood.

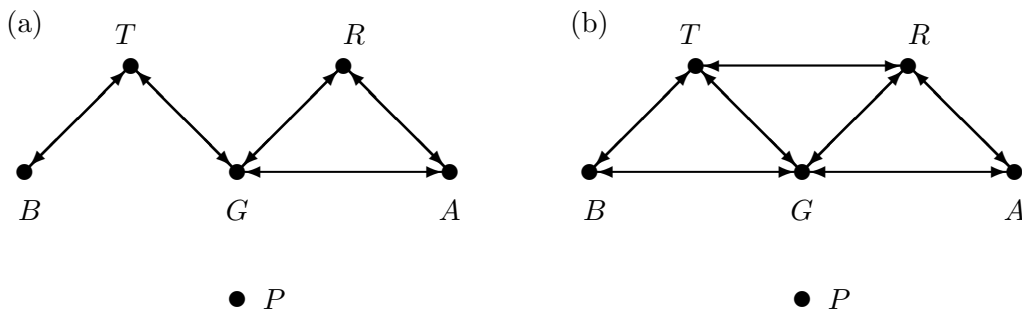


Figure 4: Covariance graph selected by SIN for (a) simultaneous level 0.1 and (b) simultaneous level 0.25.

Table 4: Marginal correlations and standard deviations from ML (lower half & 8th row) and dual estimation (upper half & 7th row).

	$G$	$A$	$B$	$P$	$T$	$R$
$G$		0.499	0	0	0.256	-0.316
$A$	0.515		0	0	0	-0.261
$B$	0	0		0	0.526	0
$P$	0	0	0		0	0
$T$	0.287	0	0.479	0		0
$R$	-0.375	-0.314	0	0	0	
$SD_{\text{dual}}$	2.98	0.43	2839.89	138.98	1293.67	1.07
$SD_{\text{ML}}$	3.14	0.44	2987.35	142.80	1359.93	1.17

Table 5: Marginal correlations and standard deviations from ML (lower half & 8th row) and dual estimation (upper half & 7th row).

	$G$	$A$	$B$	$P$	$T$	$R$
$G$		0.499	0.169	0	0.303	-0.218
$A$	0.512		0	0	0	-0.248
$B$	0.170	0		0	0.552	0
$P$	0	0	0		0	0
$T$	0.302	0	0.558	0		0.267
$R$	-0.225	-0.259	0	0	0.274	
$SD_{\text{dual}}$	2.98	0.43	2896.41	138.98	1398.54	1.13
$SD_{\text{ML}}$	3.02	0.44	2987.35	142.80	1438.47	1.15

## 7 Discussion

The new Iterative Conditional Fitting (ICF) algorithm finds solutions to the likelihood equations in covariance graph models using only tools from least squares regression. The idea to iteratively fit conditional distributions while fixing marginal distributions is not limited in any way to the Gaussian case of covariance graph models discussed in this paper. In fact, initial work by the authors on applying the ICF idea in binary graphical models for marginal independence is very promising.

As shown by Kauermann (1996), covariance graph models are a dualization of undirected graph models (Markov random fields). It is a very nice feature that ICF extends this duality to the level of fitting algorithms. The commonly used method for fitting undirected graph models is the iterative proportional fitting (IPF) algorithm (Speed and Kiiveri, 1986; Whittaker, 1990, pp.182–185). Whereas IPF fits marginal distributions while fixing conditionals, ICF does exactly the converse as said above.

We have described two incarnations of ICF, the first in Algorithm 1 uses solely univariate regression, whereas the second in Algorithm 4 is based on seemingly unrelated regressions. In order to run the second version a family of complete subsets with union equal to the vertex set has to be chosen. In this choice there is a trade off. The benefit of choosing large complete sets is that in the course of ICF the likelihood function is maximized over larger sections of the parameter space; the disadvantage is the computational overhead in carrying out generalized least squares computations as opposed to standard univariate least squares. Future practical experience will show whether general recommendations in this trade off can be given, but the structure of the particular covariance graph considered will certainly be important.

The examples in Section 6 suggest that if the model fits the data well, then ML and Kauermann’s (1996) dual estimates will be quite similar. However, the example in Section 6.2 also reveals that the dual estimate can be quite different from ML estimates when the model describes the data less well. Thus, it would be dangerous to use dual estimates in model selection procedures based on maximum likelihood methodology. It should be noted, however, that if all hypothesized models come from the class of covariance graph models, then maximum dual likelihood methods can be used (Christensen, 1989; Kauermann, 1996).

In this paper we have presented ICF in a way that does not exploit the structure of the specific covariance graph being fitted. To make this concrete let us take up the example of the covariance graph shown in Figure 1. This graph implies that  $1 \perp\!\!\!\perp (2, 4)$  and  $2 \perp\!\!\!\perp (1, 3)$ . This marginal independences are equivalent to the conditional independences  $1 \perp\!\!\!\perp 2$ ,  $3 \perp\!\!\!\perp 2 \mid 1$ , and  $4 \perp\!\!\!\perp 1 \mid 2$ . From these it can be shown that the likelihood function based on the joint distribution of variables  $(1, 2, 3, 4)$  can be factored as:

$$f_{\Sigma}(y_1, y_2, y_3, y_4) = f(y_1)f(y_2)f(y_3, y_4 \mid y_1, y_2).$$

Since the parameter space also factorizes into Cartesian products in agreement with the factorization of the likelihood function, it follows that the components  $\hat{\sigma}_{11}$  and  $\hat{\sigma}_{22}$  of a solution to the (joint) likelihood equations must coincide with the corresponding empirical quantities  $S_{11}$  and  $S_{22}$ , respectively. Thus the pseudo-variable regressions  $(Y_1 \mid Y_2, Y_3, Y_4)$  and  $(Y_2 \mid Y_1, Y_2, Y_4)$  need not be carried out, which speeds up the algorithm. In Drton and Richardson (2004a), we describe in general when such factorizations are possible. In exploiting the factorizations the extension of ICF to ancestral graphs described in Drton and Richardson (2004b) is useful.

## 8 Related work

The ICF algorithm has similarities with the Iterative Conditional Modes (ICM) algorithm of Besag (1986). However, ICM obtains maximum *a posteriori* estimates in a

Bayesian framework, whereas our ICF maximizes a likelihood function. The difference in the update steps is that in the updates of ICM conditional density functions are maximized, whereas in the updates of ICF one maximizes conditional likelihood functions.

Another related algorithm is the Conditional Iterative Proportional Fitting (CIPF) algorithm of Cramer (1998, 2000). CIPF can be used to maximize the likelihood function of a model that comprises joint distributions for which a set of conditional distributions are set equal to prescribed conditionals. However, CIPF differs from ICF for covariance graphs because the update steps of ICF do not simply equate a conditional distribution with a prescribed conditional, but rather find a conditional distribution by maximizing a conditional likelihood function. In particular, since ICF maximizes over sections in the parameter space that depend on the current estimate, these sections and hence the maximizer over the section, i.e. the fitted conditional distribution, will generally not be the same in two different iterations of ICF.

## Acknowledgment

We thank Steffen Lauritzen for pointing out the duality between ICF and IPF. This work was supported by the U.S. National Science Foundation (DMS-9972008), the University of Washington Royalty Research Fund and the William and Flora Hewlett Foundation.

## References

- Anderson, T. W. (1969). Statistical inference for covariance matrices with linear structure. In *Multivariate Analysis, II (Proc. Second Internat. Sympos., Dayton, Ohio, 1968)*, pp. 55–66. 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*, pp. 1–24. University of North Carolina Press, Chapel Hill, N.C.
- Anderson, T. W. (1973). Asymptotically efficient estimation of covariance matrices with linear structure. *Ann. Statist.* 1, 135–141.
- Anderson, T. W. and I. Olkin (1985). Maximum-likelihood estimation of the parameters of a multivariate normal distribution. *Linear Algebra Appl.* 70, 147–171.
- Banerjee, M. and T. S. Richardson (2003). On a dualization of graphical Gaussian models: A correction note. *Scand. J. Statist.* 30, 817–820.
- Besag, J. (1986). On the statistical analysis of dirty pictures. *J. Roy. Statist. Soc. Ser. B* 48, 259–302.

- Buhl, S. (1993). On the existence of maximum likelihood estimators for graphical gaussian models. *Scand. J. Statist.* 20, 263–270.
- Christensen, E. S. (1989). Statistical properties of  $I$ -projections within exponential families. *Scand. J. Statist.* 16, 307–318.
- Cox, D. R. and N. Wermuth (1993). Linear dependencies represented by chain graphs (with discussion). *Statist. Sci.* 8, 204–218, 247–277.
- Cox, D. R. and N. Wermuth (1996). *Multivariate Dependencies: Models, Analysis and Interpretation*. London: Chapman and Hall.
- Cox, D. R., N. Wermuth, and G. Marchetti (2004). Decompositions and estimation of a chain of covariances. Technical report, Department of Mathematical Statistics, Chalmers Göteborgs Universitet.
- Cramer, E. (1998). Conditional iterative proportional fitting for Gaussian distributions. *J. Multivariate Anal.* 65, 261–276.
- Cramer, E. (2000). Probability measures with given marginals and conditionals:  $I$ -projections and conditional iterative proportional fitting. *Statist. Decisions* 18, 311–329.
- Drton, M. (2004). Computing all roots of the likelihood equations of seemingly unrelated regressions. *Journal of Symbolic Computation*. Submitted.
- Drton, M. and M. Eichler (2004). Maximum likelihood estimation in Gaussian chain graph models under the alternative Markov property. Manuscript.
- Drton, M. and M. D. Perlman (2004). Model selection for Gaussian concentration graphs. *Biometrika* 91, 591–602.
- Drton, M. and M. D. Perlman (2005). A SINful approach to Gaussian graphical model selection. *Statist. Sci.*. Submitted.
- Drton, M. and T. S. Richardson (2003). A new algorithm for maximum likelihood estimation in Gaussian graphical models for marginal independence. In U. Kjærulff and C. Meek (Eds.), *Proceedings of the 19th Conference on Uncertainty in Artificial Intelligence*, pp. 184–191. San Francisco: Morgan Kaufmann.
- Drton, M. and T. S. Richardson (2004a). Graphical answers to questions about likelihood inference for Gaussian covariance models. Technical Report 467, Department of Statistics, University of Washington.

- Drton, M. and T. S. Richardson (2004b). Iterative conditional fitting for Gaussian ancestral graph models. In M. Chickering and J. Halpern (Eds.), *Proceedings of the 20th Conference on Uncertainty in Artificial Intelligence*, pp. 130–137. San Francisco: Morgan Kaufmann.
- Drton, M. and T. S. Richardson (2004c). Multimodality of the likelihood in the bivariate seemingly unrelated regressions model. *Biometrika* *91*, 383–392.
- Edwards, D. M. (2000). *Introduction to Graphical Modelling* (Second ed.). New York: Springer-Verlag.
- Grzebyk, M., P. Wild, and D. Chouanière (2004). On identification of multi-factor models with correlated residuals. *Biometrika* *91*, 141–151.
- Harville, D. A. (1997). *Matrix Algebra from a Statistician's Perspective*. New York: Springer-Verlag.
- Kauermann, G. (1996). On a dualization of graphical Gaussian models. *Scand. J. Statist.* *23*, 105–116.
- Koster, J. T. A. (1999). On the validity of the Markov interpretation of path diagrams of Gaussian structural equation systems with correlated errors. *Scand. J. Statist.* *26*, 413–431.
- Lauritzen, S. L. (1996). *Graphical Models*. Oxford Statistical Science Series. Oxford, UK: Clarendon Press.
- Mao, Y., F. R. Kschischang, and B. J. Frey (2004). Convolutional factor graphs as probabilistic models. In U. Kjærulff and C. Meek (Eds.), *Proceeding of the 20th Conference on Uncertainty in Artificial Intelligence*, pp. 374–381. San Francisco: Morgan Kaufmann.
- Pearl, J. (1988). *Probabilistic Reasoning in Intelligent Systems*. Morgan Kaufmann.
- Pearl, J. (2000). *Causality*. Cambridge, UK: Cambridge University Press.
- Pearl, J. and N. Wermuth (1994). When can association graphs admit a causal interpretation? In *Selecting Models from Data: Artificial Intelligence and Statistics IV*, Volume 89 of *Lecture Notes in Statistics*, pp. 205–214. New York: Springer.
- Richardson, T. S. (2003). Markov properties for acyclic directed mixed graphs. *Scand. J. Statist.* *30*, 145–157.
- Richardson, T. S. and P. Spirtes (2002). Ancestral graph Markov models. *Ann. Statist.* *30*, 962–1030.

- Richardson, T. S. and P. Spirtes (2003). Causal inference via ancestral graph models. In P. Green, N. Hjort, and S. Richardson (Eds.), *Highly Structured Stochastic Systems*, Chapter 3, pp. 83–105. Oxford, UK: Oxford University Press.
- Roverato, A. and J. Whittaker (1996). A hyper normal prior distribution for approximate Bayes factor calculations on non-decomposable graphical Gaussian models. Unpublished manuscript.
- Speed, T. P. and H. T. Kiiveri (1986). Gaussian Markov distributions over finite graphs. *Ann. Statist.* *14*, 138–150.
- Spirtes, P., C. Glymour, and R. Scheines (2000). *Causation, Prediction, and Search* (Second ed.). Cambridge, MA: MIT Press.
- Whittaker, J. (1990). *Graphical Models in Applied Multivariate Statistics*. Chichester: Wiley.
- Wright, S. (1921). Correlation and Causation. *J. Agricultural Research* *20*, 557–585.
- Zellner, A. (1962). An efficient method of estimating seemingly unrelated regression equations and tests for aggregation bias. *J. Amer. Statist. Assoc.* *57*, 348–368.

DEPARTMENT OF MATHEMATICS  
UNIVERSITY OF CALIFORNIA  
1073 EVANS HALL #3840  
BERKELEY, CA, 94720-3840  
U.S.A.  
E-MAIL: [drton@math.berkeley.edu](mailto:drton@math.berkeley.edu)

DEPARTMENT OF STATISTICS  
UNIVERSITY OF WASHINGTON  
BOX 354322  
SEATTLE, WA 98105-4322  
U.S.A.  
E-MAIL: [tsr@stat.washington.edu](mailto:tsr@stat.washington.edu)