Iterative Conditional Fitting for Gaussian Ancestral Graph Models

Mathias Drton
Department of Statistics
University of Washington
Seattle, WA 98195-4322

Thomas S. Richardson
Department of Statistics
University of Washington
Seattle, WA 98195-4322

Abstract

Ancestral graph models, introduced by Richardson and Spirtes (2002), generalize both Markov random fields and Bayesian networks. A key feature of ancestral graph models is that the global Markov property is closed under conditioning and marginalization. The conditional independence structures that can be encoded by ancestral graphs coincide with the structures that can arise from a Bayesian network with selection and unobserved variables. Thus, association structures learned via ancestral graph models may be interpreted causally. In this paper, we consider Gaussian ancestral graph models and present an algorithm for maximum likelihood estimation. We call this new algorithm 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.

1 INTRODUCTION

Markov random fields or equivalently undirected graph models as well as Bayesian networks or equivalently directed acyclic directed graph (DAG) models have found wide-spread application. Well-known generalizations of both undirected graph models and DAG models are the chain graph models, which can be equipped with two alternative Markov properties (Andersson et al. 2001). A different generalization is obtained from ancestral graphs, introduced by Richardson and Spirtes (2002) = RS (2002). Whereas chain graphs allow both undirected and directed edges, ancestral graphs have edges of three possible types: undirected and directed edges are complemented by bidirected edges.

In ancestral graphs, m-separation, a natural extension of d-separation, yields a global Markov property that is closed under conditioning and marginalization. Interpreted via this Markov property, ancestral graphs can encode any conditional independence structures that can arise from a Bayesian network with selection and unobserved variables (RS 2002). Marginalization (forming the marginal distribution of the observed variables) is associated with introducing bidirected edges; conditioning (on selection variables) is associated with introducing undirected edges. Due to this connection between ancestral graphs and underlying DAGs, ancestral graph models not only generalize undirected graph and DAG models, but also lead to conditional independence structures that could have arisen from DAGs and hence are causally interpretable.

This paper is a first step towards making ancestral graph methodology available for use in applications. The problem we consider is the problem of estimating or learning the parameters of a given ancestral graph model by maximum likelihood. We restrict ourselves to the Gaussian case, for which RS (2002) provided a parameterization, and propose an algorithm, which extends previous work by Drton and Richardson (2003). We give this new algorithm the name "iterative conditional fitting" (ICF) 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 (Whittaker 1990, pp. 182–185), in the steps of which a marginal distribution is fitted for a fixed conditional distribution.

The remainder of this paper is organized as follows. In Sections 2 and 3 we define ancestral graphs and their global Markov property, and in Section 4 we in-

troduce Gaussian ancestral graph models. In Section 5 we present the ICF algorithm. We implemented the algorithm in a function library for the statistical programming system R. Its use is demonstrated in an example session in Section 6. We conclude in Section 7.

2 ANCESTRAL GRAPHS

Consider a graph $G = (V, E)$ with the vertex set V and the edge set E containing three types of edge, *undirected* ($-$), *directed* ($\rightarrow$) and *bidirected* ($\leftrightarrow$). However, the graph G is not allowed to have an edge from a vertex to itself or more than one edge between a given pair of vertices. We use the following terminology to describe relations between two vertices $i, j \in V$ in G:

If $\left\{ \begin{array}{c} i - j \\ i \leftrightarrow j \\ i \rightarrow j \end{array} \right\}$ then i is a $\left\{ \begin{array}{c} \textit{neighbor} \\ \textit{spouse} \\ \textit{parent} \end{array} \right\}$ of j.

We denote the set of neighbors of a vertex i as $\mathrm{ne}(i)$, the set of spouses as $\mathrm{sp}(i)$ and the set of parents as $\mathrm{pa}(i)$. For vertex sets $A \subseteq V$, we define $\mathrm{ne}(A) = \cup(\mathrm{ne}(i) \mid i \in A)$ and similarly $\mathrm{sp}(A)$, $\mathrm{pa}(A)$.

A sequence of edges between two vertices i and j in G is an ordered (multi)set of edges $< e_1, \ldots, e_m >$, such that there exists a sequence of vertices (not necessarily distinct) $< i = i_1, \ldots, i_{m+1} = j >$, where edge e_{i_k} has endpoints i_k, i_{k+1}. A sequence of edges for which the corresponding sequence of vertices contains no repetitions is called a *path*. A path of the form $i \rightarrow \cdots \rightarrow j$, on which every edge is of the form $\rightarrow$, with the arrowheads pointing toward j, is a *directed path from i to j*. A directed path from a vertex i to itself is called a (fully) *directed cycle*. A special case of the ancestral graphs defined below are directed acyclic graphs (DAG), in which all edges are directed, and there are no directed cycles.

A vertex i is said to be an *ancestor* of a vertex j, denoted $i \in \mathrm{an}(j)$, if either there is a directed path $i \rightarrow \cdots \rightarrow j$ from i to j, or $i = j$.

Definition 1 (Ancestral Graphs, RS 2002, §3) *A graph $G = (V, E)$ with undirected, directed and bidirected edges is an* ancestral graph *if for all $i \in V$ it holds that*

(i) if $\mathrm{ne}(i) \neq \emptyset$ *then* $\mathrm{pa}(i) \cup \mathrm{sp}(i) = \emptyset$*;*

(ii) $i \notin \mathrm{an}(\mathrm{pa}(i) \cup \mathrm{sp}(i))$.

In words, condition (i) states that if there is an undirected edge with endpoint i then there may not exist a directed or bidirected edge with an arrowhead at i, and condition (ii) states that there may not be a directed path from a vertex i to one of its parents or spouses. Condition (ii) may be restated equivalently as (1) there are no directed cycles, and (2) no spouses are ancestors.

An example of an ancestral graph with vertex set $V = \{0, 1, 2, 3, 4\}$ is given in Figure 1. Additional examples can be found in RS (2002, e.g. Fig. 3, 6, 7 and 12). Note also that any DAG and any undirected graph is an ancestral graph.

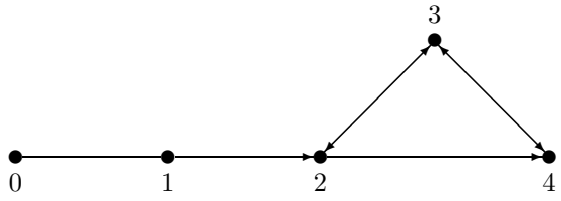


Figure 1: An ancestral graph.

Condition (i) implies that an ancestral graph can be decomposed into an undirected part and a part with only directed and bidirected edges (RS 2002, §3.2). Let $\mathrm{un}_G = \{i \in V \mid \mathrm{pa}(i) \cup \mathrm{sp}(i) = \emptyset\}$. Then the subgraph $G_{\mathrm{un}_G} = [\mathrm{un}_G, E \cap (\mathrm{un}_G \times \mathrm{un}_G)]$ induced by un_G is an undirected graph, and any edge between $i \in \mathrm{un}_G$ and $j \notin \mathrm{un}_G$ is directed as $i \rightarrow j$. Furthermore, the induced subgraph $G_{V \setminus \mathrm{un}_G}$ contains only directed and bidirected edges. In Figure 1, $\mathrm{un}_G = \{0, 1\}$.

3 GLOBAL MARKOV PROPERTY

Pearl's (1988) d-separation criterion for DAGs can be extended to ancestral graphs. A nonendpoint vertex i on a path is a *collider on the path* if the edges preceding and succeeding i on the path have an arrowhead at i, that is, $\rightarrow i \leftarrow$, $\rightarrow i \leftrightarrow$, $\leftrightarrow i \leftarrow$, $\leftrightarrow i \leftrightarrow$. A nonendpoint vertex i on a path which is not a collider is a *noncollider on the path*. A path between vertices i and j in an ancestral graph G is said to be *m-connecting given a set C* (possibly empty), with $i, j \notin C$, if:

(i) every noncollider on the path is not in C, and

(ii) every collider on the path is in $\mathrm{an}(C)$.

If there is no path m-connecting i and j given C, then i and j are said to be m-separated given C. Sets A and B are m-separated given C, if for every pair i, j, with $i \in A$ and $j \in B$, i and j are m-separated given C (A, B, C are disjoint sets; A, B are nonempty). This is an extension of Pearl's d-separation criterion to ancestral graphs in that in a DAG, a path is d-connecting if and only if it is m-connecting.

Let $G = (V, E)$ be an ancestral graph whose vertices V are identified with random variables $(Y_i \mid i \in V)$

and let P be the joint probability distribution of $(Y_i \mid i \in V)$. If $Y_A \perp\!\!\!\perp Y_B \mid Y_C$ whenever A and B are m-separated given C, then P is said to satisfy the *global Markov property for* G, or to be *globally Markov with respect to* G. Here $Y_A = (Y_i \mid i \in A)$ for $A \subseteq V$. For the joint distribution P to be globally Markov with respect to the graph G in Figure 1, the conditional independences $Y_0 \perp\!\!\!\perp Y_{234} \mid Y_1$, $Y_1 \perp\!\!\!\perp Y_3$, $Y_1 \perp\!\!\!\perp Y_4 \mid Y_2$ must hold. The global Markov property in addition implies for example $Y_1 \perp\!\!\!\perp Y_3 \mid Y_0$ and $Y_1 \perp\!\!\!\perp Y_4 \mid Y_{02}$ but these are consequences of the previous conditional independences.

In this example, if two vertices i and j are not adjacent, then a conditional independence $Y_i \perp\!\!\!\perp Y_j \mid Y_C$, $C \subseteq V$, holds. Ancestral graphs for which this is true are called *maximal* (RS 2002, §3.7). In fact, by adding bidirected edges, any non-maximal ancestral graph can be converted into a unique maximal ancestral graph without changing the independence model implied by the global Markov property.

The main motivation for considering ancestral graphs is that they can encode conditional independence structures arising from DAGs with selection and unobserved variables. We illustrate this by an example. Consider the DAG in Figure 2. Assume that the variables u_{23} and u_{34} are unobserved and that variable s_{01} is a selection variable. If we form the conditional distribution $(0\ 1\ 2\ 3\ 4 \mid s_{01})$, with unobserved variables marginalized out and selection variables conditioned on, then the conditional independences holding in $(0\ 1\ 2\ 3\ 4 \mid s_{01})$ are exactly those implied by the global Markov property of the graph G from Figure 1. For details on this connection between DAGs and ancestral graphs see RS (2002).

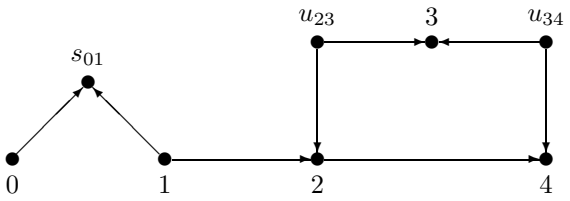


Figure 2: DAG with selection variable s_{01} and the unobserved variables u_{23} and u_{34}.

4 GAUSSIAN ANCESTRAL GRAPH MODELS

Suppose now that the variables $(Y_i \mid i \in V)$ jointly follow a centered Gaussian $\equiv$ normal distribution $\mathcal{N}_V(0, \Sigma) \in \mathbb{R}^V$, where $\Sigma = (\sigma_{ij}) \in \mathbb{R}^{V \times V}$ is the unknown positive definite covariance matrix. Let $\mathbf{P}(V)$ be the cone of all positive definite $V \times V$ matrices and let $\mathbf{P}(G)$ be the subcone of all matrices $\Sigma \in \mathbf{P}(V)$ such that $\mathcal{N}_V(0, \Sigma)$ is globally Markov with respect to the given ancestral graph G. The *Gaussian ancestral graph model* based on G is the family of all normal distributions

$$\mathbf{N}(G) = \{\mathcal{N}_V(0, \Sigma) \mid \Sigma \in \mathbf{P}(G)\}. \tag{1}$$

As shown in RS (2002, §8.4), the model $\mathbf{N}(G)$ forms a curved exponential family.

4.1 PARAMETERIZATION

RS (2002, §8) provide a parameterization of the Gaussian ancestral graph model $\mathbf{N}(G)$. This parameterization associates one parameter with each vertex in V and each edge in E. Let $\Lambda = (\lambda_{ij})$ be a positive definite $\mathrm{un}_G \times \mathrm{un}_G$ matrix such that $\lambda_{ij} \neq 0$ only if $i = j$ or $i - j$. Recall that $i, j \in \mathrm{un}_G$ can only be adjacent by an undirected edge. Let $\Omega = (\omega_{ij})$ be a positive definite $(V \setminus \mathrm{un}_G) \times (V \setminus \mathrm{un}_G)$ matrix such that $\omega_{ij} \neq 0$ only if $i = j$ or $i \leftrightarrow j$. Finally, let $B = (\beta_{ij})$ be a $V \times V$ matrix such that $\beta_{ij} \neq 0$ only if $j \to i$. Note that the $\mathrm{un}_G \times V$ submatrix of B must be zero, i.e. $B_{\mathrm{un}_G, V} = 0$, because no vertex in un_G has a directed edge pointing towards it. With the parameter matrices Λ, B, Ω, we can define the covariance matrix

$$\Sigma = (I_V - B)^{-1} \begin{pmatrix} \Lambda^{-1} & 0 \\ 0 & \Omega \end{pmatrix} (I_V - B)^{-\mathrm{T}}, \tag{2}$$

which satisfies $\Sigma \in \mathbf{P}(G)$. Equivalently said, the normal distribution $\mathcal{N}_V(0, \Sigma)$ is globally Markov with respect to the considered ancestral graph G. If G is maximal, then for any $\Sigma \in \mathbf{P}(G)$ there exist unique Λ, Ω, B of the above type such that (2) holds.

The population interpretation of the parameters is the following: First, the parameter matrix Λ clearly forms an inverse covariance matrix for the undirected graph G_{un_G}. Second, just as for Gaussian DAG models, the parameter β_{ij}, associated with a directed edge $j \to i$ is the regression coefficient for variable j in the regression of variable i on its parents $\mathrm{pa}(i)$. Third, the parameter ω_{ii} is the conditional variance of the conditional distribution $(Y_i \mid Y_{\mathrm{pa}(i)})$, or equivalently $\omega_{ii} = \mathrm{Var}[\varepsilon_i]$, where

$$\varepsilon_i = Y_i - \sum_{j \in \mathrm{pa}(i)} \beta_{ij} Y_j. \tag{3}$$

The parameter ω_{ij} for $i \leftrightarrow j$ is the covariance between the residuals ε_i and ε_j. We illustrate the parameterization by showing in Figure 3 the parameters for the graph from Figure 1.

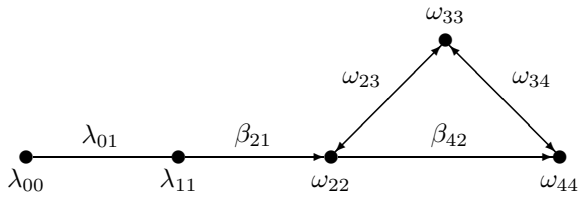


Figure 3: Parameters of a Gaussian ancestral graph model.

4.2 MAXIMUM LIKELIHOOD ESTIMATION

This article considers the estimation of the unknown parameter Σ, or equivalently Λ, Ω, B, of a Gaussian ancestral graph model $\mathbf{N}(G)$ based on a sample of i.i.d. observations from $\mathbf{N}(G)$. We assume that G is maximal. Let the i.i.d. copies in the sample be indexed by the set N, which can be interpreted as indexing the subjects on which we observed the variables in V. Then we can group the observed random vectors in the sample as columns in the $V \times N$ matrix Y, which means that Y_{im} represents the observation of the i-th variable on the m-th subject. 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

$$S = \frac{1}{n} YY' \in \mathbb{R}^{V \times V}. \tag{4}$$

We shall assume that $n \geq p$ such that S is positive definite with probability one. Note that the case where the model also includes an unknown mean vector $\mu \in \mathbb{R}^V$ can be treated by estimating μ 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

$$\tilde{S} = \frac{1}{n}(Y - \bar{Y} \otimes 1_N)(Y - \bar{Y} \otimes 1_N)' \in \mathbb{R}^{V \times V}, \tag{5}$$

where $1_N = (1, \ldots, 1) \in \mathbb{R}^N$ and $\otimes$ is the Kronecker product. Estimation of Σ would proceed in the same way as in the centered case by using $\tilde{S}$ instead of S; the only change being that $n \geq p + 1$ ensures almost sure positive definiteness of $\tilde{S}$.

The density function of Y with respect to the Lebesgue measure is the function $f_\Sigma : \mathbb{R}^{V \times N} \to \mathbb{R}$, which can be expressed as

$$\begin{aligned} f_\Sigma(y) &= (2\pi)^{-np/2} |\Sigma|^{-n/2} \exp\{-\frac{1}{2}\mathrm{tr}(\Sigma^{-1} yy')\} \\ &= (2\pi)^{-np/2} |\Sigma|^{-n/2} \exp\{-\frac{n}{2}\mathrm{tr}(\Sigma^{-1} S)\}, \end{aligned} \tag{6}$$

see e.g. Edwards (2000, §3.1). Considered as a function of the unknown parameters for fixed data y this gives the likelihood function of the Gaussian ancestral graph model $\mathbf{N}(G)$ as the mapping $L : \mathbf{P}(G) \to \mathbb{R}$ where $L(\Sigma) = f_\Sigma(y)$. In ML estimation, the parameter Σ is estimated by the maximizer $\hat{\Sigma}$ of the likelihood L. Usually, one considers more conveniently the maximization of the log-likelihood $\ell = \log L$, which, ignoring an additive constant, takes the form

$$\ell(\Sigma) = -\frac{n}{2} \log |\Sigma| - \frac{n}{2} \mathrm{tr}\{\Sigma^{-1} S\}. \tag{7}$$

Positive definiteness of S guarantees the existence of the global maximum of $\ell(\Sigma)$ over $\mathbf{P}(G)$ but there may be multiple local maxima (Drton and Richardson 2004).

4.3 EMPLOYING THE DECOMPOSITION OF AN ANCESTRAL GRAPH

As described in RS (2002, §8.5), the decomposition of an ancestral graph G into an undirected and a directed-bidirected part is accompanied by a factorization of the density function of a distribution in the associated model $\mathbf{N}(G)$. More precisely, if $\Sigma \in \mathbf{P}(G)$, then

$$f_\Sigma(y) = f_\Lambda(y_{\mathrm{un}_G}) f_{B,\Omega}(y_{V \setminus \mathrm{un}_G} \mid y_{\mathrm{un}_G}). \tag{8}$$

Here $f_\Lambda(y_{\mathrm{un}_G})$ is the marginal density of Y_{un_G}, and $f_{B,\Omega}(y_{V \setminus \mathrm{un}_G} \mid y_{\mathrm{un}_G})$ is the conditional density of $(Y_{V \setminus \mathrm{un}_G} \mid Y_{\mathrm{un}_G})$. From (8) it follows that we can find the MLE of Λ by maximizing the marginal likelihood function $L(\Lambda) = f_\Lambda(y_{\mathrm{un}_G})$. Since $Y_{\mathrm{un}_G} \sim \mathcal{N}(0, \Lambda)$, this is precisely fitting an undirected graph model based on the graph G_{un_G} using only the observations for variables in un_G, that is $Y_{\mathrm{un}_G, N}$. Thus, the MLE $\hat{\Lambda}$ of Λ can be obtained by iterative proportional fitting, as described for example in Whittaker (1990, pp. 182–185).

In order to find the MLE $(\hat{B}, \hat{\Omega})$ of (B, Ω) we can maximize the conditional likelihood function

$$L(B, \Omega) = f_{B,\Omega}(y_{V \setminus \mathrm{un}_G} \mid y_{\mathrm{un}_G}). \tag{9}$$

It is easy to see that the global Markov property for the graph G implies that

$$L(B, \Omega) = f_{B,\Omega}(y_{V \setminus \mathrm{un}_G} \mid y_{\mathrm{pa}(V \setminus \mathrm{un}_G) \cap \mathrm{un}_G}). \tag{10}$$

Thus, only a subset $\mathrm{db}_G = [V \setminus \mathrm{un}_G] \cup \mathrm{pa}(V \setminus \mathrm{un}_G)$ of the variables is needed for estimating (B, Ω); i.e. we are using the observations $Y_{\mathrm{db}_G, N}$. The set db_G is the set of all vertices i in G that are the endpoint of at least one directed or bidirected edge, i.e. there is an edge $i \to j$, $i \leftarrow j$ or $i \leftrightarrow j$. For the graph G from Figure 1, the induced subgraph G_{db_G} is shown in Figure 4.

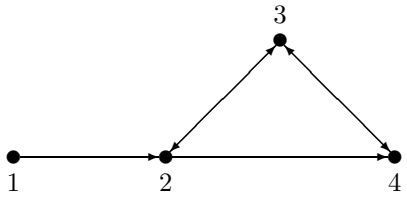


Figure 4: The graph G_{db_G}.

In the next section, we present an algorithm for estimating (B, Ω). This algorithm extends an idea developed in Drton and Richardson (2003), who in fact consider ancestral graphs with bidirected edges only. The idea is to iteratively fit a conditional distribution for a fixed marginal distribution. Thus we call this algorithm *iterative conditional fitting*. Note the duality with the iterative proportional fitting algorithm, in which, cycling through a list of cliques, the marginal distribution over a clique $C \subseteq V$ is fitted while fixing the conditional distribution $(Y_{V \setminus C} \mid Y_C)$.

5 ITERATIVE CONDITIONAL FITTING

5.1 THE GENERAL ALGORITHM

Let G be a maximal ancestral graph. The idea of iterative conditional fitting (ICF) put forward by Drton and Richardson (2003) is to repeatedly iterate through all vertices $i \in V$, and

(i) Fix the marginal distribution for $Y_{-i} = Y_{V \setminus \{i\}}$.

(ii) Fit the conditional distribution $(Y_i \mid Y_{-i})$ under the constraints implied by the Gaussian ancestral graph model $\mathbf{N}(G)$.

(iii) Find a new estimate of Σ from the estimated parameters of the conditional distribution $(Y_i \mid Y_{-i})$ and the fixed parameters of the marginal distribution of Y_{-i}.

In Drton and Richardson (2003), where the graph G only contained bidirected edges, the problem of fitting $(Y_i \mid Y_{-i})$ under constraints could be rephrased as a least squares regression problem. Here, however, where G also contains directed edges, the consideration of $(Y_i \mid Y_{-i})$ is complicated. Fortunately, we can "remove" directed edges by forming residuals as in (3), and considering the conditional distribution $(Y_i \mid \varepsilon_{-i})$ as presented in the following. Such a residual trick can already be found in Telser (1964).

Before formulating ICF for ancestral graphs, we note two facts. First, the maximization of $L(B, \Omega)$ from (9) is by (8) equivalent to maximizing $L(\Sigma) = L(\Lambda, B, \Omega)$ with holding Λ fixed to some feasible value Λ_0, which could be for example the identity matrix or the matrix found by iterative proportional fitting as described in 4.3. Second, fixing $\Lambda = \Lambda_0$, the matrix $\varepsilon = (I_V - B)Y$ has i-th row equal to the residual ε_i defined in (3) and each column of ε has covariance matrix $\Psi = (\psi_{ij})$ equal to

$$\Psi = \begin{pmatrix} \Lambda_0^{-1} & 0 \\ 0 & \Omega \end{pmatrix}. \tag{11}$$

From this and the fact $B_{\mathrm{un}_G, V} = 0$ we see that, in order to estimate (B, Ω) we need only cycle through the vertices $i \notin \mathrm{un}_G$.

Next we compute the conditional distribution $(Y_i \mid \varepsilon_{-i})$ for $i \notin \mathrm{un}_G$. This distribution is obviously Gaussian and its conditional variance equals

$$\mathrm{Var}[Y_i \mid \varepsilon_{-i}] = \omega_{ii.-i}, \tag{12}$$

which is defined as

$$\omega_{ii.-i} = \omega_{ii} - \Omega_{i,-i}(\Omega_{-i,-i})^{-1}\Omega_{-i,i}. \tag{13}$$

Equation (12) holds because, for $i \in \mathrm{un}_G$

$$\begin{aligned} \mathrm{Var}[Y_i \mid \varepsilon_{-i}] &= \mathrm{Var}[\varepsilon_i \mid \varepsilon_{-i}] \\ &= \psi_{ii} - \Psi_{i,-i}(\Psi_{-i,-i})^{-1}\Psi_{-i,i} \\ &= \omega_{ii} - \Omega_{i,-i}(\Omega_{-i,-i})^{-1}\Omega_{-i,i}. \end{aligned} \tag{14}$$

Note that when writing $\Omega_{-i,-i}$ we mean the $[V \setminus (\mathrm{un}_G \cup \{i\})] \times [V \setminus (\mathrm{un}_G \cup \{i\})]$ submatrix of Ω. The normal distribution $(Y_i \mid \varepsilon_{-i})$ is now specified by (12) and the conditional expectation, which equals

$$\begin{aligned} \mathrm{E}[Y_i \mid \varepsilon_{-i}] &= \sum_{j \in \mathrm{pa}(i)} \beta_{ij} \mathrm{E}[Y_j \mid \varepsilon_{-i}] + \mathrm{E}[\varepsilon_i \mid \varepsilon_{-i}] \\ &= \sum_{j \in \mathrm{pa}(i)} \beta_{ij} Y_j + \sum_{k \in \mathrm{sp}(i)} \omega_{ik} Z_k, \end{aligned} \tag{15}$$

where the *pseudo-variable* Z_k is equal to the k-th row in

$$Z_{\mathrm{sp}(i)} = [(\Omega_{-i,-i})^{-1}]_{\mathrm{sp}(i),-i}\varepsilon_{-i}. \tag{16}$$

The derivation of (15) relies on two facts. First, for any $j \in \mathrm{pa}(i)$, Y_j is a function of $\varepsilon_{\mathrm{an}(i) \setminus \{i\}}$ and thus, $\mathrm{E}[Y_j \mid \varepsilon_{-i}] = Y_j$. Second, for $i \notin \mathrm{un}_G$ the residual covariance $\psi_{ij} = 0$ if $i \not\leftrightarrow j$. Thus, $\Psi_{i,\mathrm{nsp}(i)} = 0$, which implies that for $i \notin \mathrm{un}_G$:

$$\begin{aligned} \mathrm{E}[\varepsilon_i \mid \varepsilon_{-i}] &= \Psi_{i,-i}(\Psi_{-i,-i})^{-1}\varepsilon_{-i} \\ &= \Psi_{i,\mathrm{sp}(i)}[(\Psi_{-i,-i})^{-1}]_{\mathrm{sp}(i),-i}\varepsilon_{-i} \\ &= \Omega_{i,\mathrm{sp}(i)}[(\Omega_{-i,-i})^{-1}]_{\mathrm{sp}(i),-i}\varepsilon_{-i} \\ &= \sum_{k \in \mathrm{sp}(i)} \omega_{ik} Z_k. \end{aligned} \tag{17}$$

After this preparation we are now ready to formulate ICF for ancestral graphs to find the MLE $(\hat{B}, \hat{\Omega})$. Until convergence, for each $i \notin \mathrm{un}_G$:

1. Fix $\Omega_{-i,-i}$ and all $B_{j,\mathrm{pa}(j)} = (\beta_{j\ell} \mid \ell \in \mathrm{pa}(j))$ for $j \neq i$;
2. Use the fixed $\beta_{j\ell}$ to compute the residuals ε_j for $j \neq i$ from (3);
3. Use the fixed $\Omega_{-i,-i}$ to compute the pseudo-variables Z_k for $k \in \mathrm{sp}(i)$;
4. Carry out a least squares regression with response variable Y_i and covariates Y_j, $j \in \mathrm{pa}(i)$, and Z_k, $k \in \mathrm{sp}(i)$ to obtain estimates of β_{ij}, $j \in \mathrm{pa}(i)$, ω_{ik}, $k \in \mathrm{sp}(i)$, and $\omega_{ii.-i}$;
5. Compute an estimate of ω_{ii} using the new estimates and the fixed parameters from the relation $\omega_{ii} = \omega_{ii.-i} + \Omega_{i,-i}(\Omega_{-i,-i})^{-1}\Omega_{-i,i}$, compare (13).

After steps (1)-(5), we move on to the next vertex in $V \setminus \mathrm{un}_G$. The procedure is continued until convergence.

5.2 CONVERGENCE

It is easy to see that this ICF algorithm is an iterative partial maximization algorithm (Lauritzen 1996, App. A.4) since in the i-th step we maximize the conditional likelihood $L(B, \Omega)$ from (9) over the section in the parameter space defined by fixing the parameters $\Omega_{-i,-i}$, and $B_{j,\mathrm{pa}(j)}$, $j \neq i$. The same reasoning as in Drton and Richardson (2003, §4.5) applies and we can deduce that for any feasible starting value the algorithm produces a sequence of estimates, for which each accumulation point is a local maximum or a saddle point of the likelihood. Furthermore, evaluating the likelihood at each accumulation point must give the same value.

5.3 APPLYING ICF TO DAGS

It is well known that the MLE of the parameters of a Gaussian DAG model can be found by carrying out a finite number of regressions (see e.g. Goldberger 1964, or Andersson and Perlman 1998). DAG models form a special case of ancestral graph models so we can also apply ICF to a Gaussian DAG model. If the graph G is a DAG then $\mathrm{sp}(i) = \emptyset$ for all $i \in V$. Therefore, the conditional distribution of $(Y_i \mid \varepsilon_{-i})$ is fitted by regressing solely on the parents Y_j, $j \in \mathrm{pa}(i)$; compare (15). Thus the least squares regression carried out in the i-th step of ICF is always the same $(Y_i \mid Y_{\mathrm{pa}(i)})$ since it involves no pseudo-variables, which could change from one iteration to the other. This means that ICF reduces to the standard approach of fitting Gaussian DAG models if the ancestral graph under consideration is in fact a DAG.

5.4 ICF IN AN EXAMPLE

We illustrate estimation of (B, Ω) by ICF in the example of the ancestral graph depicted in Figure 1. The set un_G equals $\{0, 1\}$ and in fact only the variables $\mathrm{db}_G = \{1, 2, 3, 4\}$ are relevant for estimating (B, Ω), compare Figure 4. The iteration steps, described in items (1)-(5) in Section 5.1, have to be carried only for $i \in V \setminus \mathrm{un}_G = \{2, 3, 4\}$. In Figure 5, we show the response variable Y_i in the i-th ICF update step as a filled circle. The remaining variables are depicted as unfilled circles. A vertex labelled by a number j represents the random variable Y_j and a vertex labelled by Z_j represents the pseudo-variable defined in (16). Finally, we use directed edges pointing from a covariate Y_j or Z_j to the response Y_i to indicate the structure of the least squares regression that has to be performed in the i-th ICF step. Note that we suppress the irrelevant variable 0.

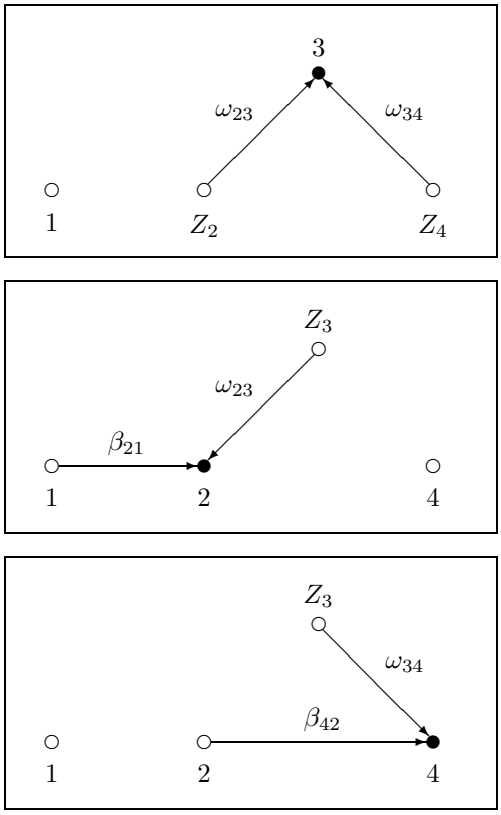


Figure 5: Illustration of the ICF update steps.

6 AN IMPLEMENTATION

The statistical programming language R (Ihaka and Gentleman 1996) provides a freeware environment for programming in interpreted code building on a large number of available routines. The team of devel-

opers of R provided a framework for writing extension libraries for R. As part of the "graphical models in R" initiative (Lauritzen 2002), Marchetti and Drton developed a function library called 'ggm', which implements functions for fitting Gaussian graphical models and, in particular, provides an implementation of ICF. The package can be downloaded from `http://cran.r-project.org/`.

In the following, we show an example session in R using 'ggm'. We begin by loading 'ggm'.

```
> library(ggm)
```

The library also contains one toy data set over 5 variables to which we can fit the ancestral graph from Figure 1. We load the data set containing the marks of $n = 88$ students in mathematics examinations in mechanics, vectors, algebra, analysis, statistics.

```
> data(marks)
> n <- 88
```

We compute the sample covariance matrix S and, for a more convenient display, we shorten the variable names to 'mec', 'vec', etc.

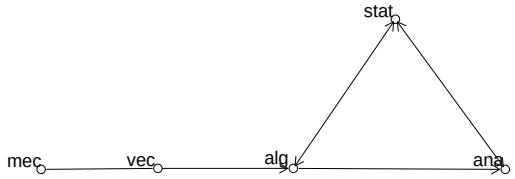
```
> S <- var(marks)
> dimnames(S) <-
+   list(c("mec","vec","alg","ana","stat"),
+        c("mec","vec","alg","ana","stat"))
> round(S, 2)
        mec    vec    alg    ana   stat
mec  305.69 127.04 101.47 106.32 117.49
vec  127.04 172.84  85.16  94.67  99.01
alg  101.47  85.16 112.89 112.11 121.87
ana  106.32  94.67 112.11 220.38 155.54
stat 117.49  99.01 121.87 155.54 297.76
```

Next we define the (maximal) ancestral graph from Figure 1, which is done by defining complete subsets for the undirected and bidirected subgraphs, and regression structures for the directed subgraph. The resulting graph is represented by an adjacency matrix $A = (a_{ij})$ where $a_{ij} = a_{ji} = 1$ if $i - j$, $a_{ij} = a_{ji} = 2$ if $i \leftrightarrow j$, and $a_{ij} = 1$, $a_{ji} = 0$ if $i \to j$.

```
> mag <- makeAG(ug=UG(~mec*vec),
+               dag=DAG(alg~vec, ana~alg),
+               bg=UG(~alg*stat+ana*stat))
> mag
     alg stat ana mec vec
alg    0    2   1   0   0
stat   2    0   2   0   0
ana    0    2   0   0   0
mec    0    0   0   0   1
vec    1    0   0   1   0
```

The package 'ggm' also provides a rudimentary tool for plotting graphs.

```
> drawGraph(mag)
```



Now we are able to fit the Gaussian ancestral graph model to the data. The function 'fitAncestralGraph' returns $\hat{\Sigma}$ as `Shat`, $\hat{\Lambda}$ as `Lhat`, $I_V - \hat{B}$ as `Bhat` and $\hat{\Omega}$ as `Ohat`. The output also includes the deviance statistic `dev`, the degrees of freedom `df` and the number of iterations `it`, that is, the number of full cycles through all $i \notin \mathrm{un}_G$ during ICF. The deviance statistic is computed as

$$\texttt{dev} = 2\ell(S) - 2\ell(\hat{\Sigma}). \tag{18}$$

Note, however, that the ICF estimate $\hat{\Sigma}$ may only be a local maximum.

```
> icf <- fitAncestralGraph(mag, S, n)
> lapply( icf , round, 2 )
$Shat
        mec    vec   alg    ana   stat
mec  305.69 127.04 40.55  38.78   0.00
vec  127.04 172.84 55.16  52.76   0.00
alg   40.55  55.16 93.74  89.66  90.30
ana   38.78  52.76 89.66 194.94 125.43
stat   0.00   0.00 90.30 125.43 297.82

$Lhat
        mec    vec alg ana stat
mec  305.69 127.04   0   0    0
vec  127.04 172.84   0   0    0
alg    0.00   0.00   0   0    0
ana    0.00   0.00   0   0    0
stat   0.00   0.00   0   0    0

$Bhat
     mec   vec   alg ana stat
mec    1  0.00  0.00   0    0
vec    0  1.00  0.00   0    0
alg    0 -0.32  1.00   0    0
ana    0  0.00 -0.96   1    0
stat   0  0.00  0.00   0    1

$Ohat
     mec vec   alg    ana   stat
mec    0   0  0.00   0.00   0.00
vec    0   0  0.00   0.00   0.00
alg    0   0 76.13   0.00  90.30
ana    0   0  0.00 109.19  39.06
stat   0   0 90.30  39.06 297.82

$dev
[1] 27.96
$df
[1] 5
$it
[1] 6
```

Comparing the deviance and the degrees of freedom using the asymptotic distribution of the deviance as $\chi^2_{\mathtt{df}}$ suggests a rather poor model fit. In fact, the mathematics marks data are a notorious example for undirected graph models (see e.g. Edwards 2000, Whittaker 1990), so we are not surprised that our toy data are not well described by the Gaussian ancestral graph model.

7 CONCLUSION

We have presented ICF = iterative conditional fitting, which is an iterative partial maximization algorithm for fitting Gaussian ancestral graph models. Fitting conditional distributions while fixing marginal distributions, ICF stands in duality with the iterative proportional fitting algorithm, in which marginal distributions are fitted while conditional distributions are fixed. ICF is particularly attractive since if the ancestral graph under consideration is in fact a DAG, then the likelihood is maximized in a finite number of step performing exactly the regressions commonly used for fitting Gaussian DAG models.

A topic of future work will be using Markov equivalence of ancestral graphs for improving efficiency. As it is true for DAGs, different ancestral graphs may induce the same statistical model, in which case the graphs are called Markov equivalent. Since the update steps of the ICF algorithm depend on the graph itself, it is important to work out which graph in a whole class of Markov equivalent graphs allows for the most efficient fitting of the associated model (see also Drton and Richardson 2003, §4.2.4).

Finally, ICF has the nice feature that its main idea of decomposing the complicated overall maximization problem into a sequence of simpler optimization problems seems also promising for the development of fitting methodology in the case of discrete variables. This extension will be the subject of future work.

Acknowledgements

We wish to thank Steffen Lauritzen for pointing out the duality between iterative conditional and iterative proportional fitting. We acknowledge support by NSF grants DMS 9972008 and DMS 0071818.

References

S. A. Andersson, D. Madigan, M. D. Perlman (2001). Alternative Markov properties for chain graphs. *Scandinavian Journal of Statistics* **28**:33–85.

S. A. Andersson, M. D. Perlman (1998). Normal linear regression models with recursive graphical Markov structure. *Journal of Multivariate Analysis* **66**:133–187.

M. Drton, T. S. Richardson (2003). A New Algorithm for Maximum Likelihood Estimation in Gaussian Graphical Models for Marginal Independence, in UAI (Uffe Kjærulff and Christopher Meek, eds.), San Francisco: Morgan Kaufmann, pp. 184–191.

M. Drton, T. S. Richardson (2004). Multimodality of the likelihood in the bivariate seemingly unrelated regressions model. *Biometrika,* to appear.

D. M. Edwards (2000). *Introduction to Graphical Modelling.* Second edition. New York: Springer-Verlag.

R. Ihaka, R. Gentleman (1996). R: A Language for Data Analysis and Graphics. *Journal of Computational and Graphical Statistics* **5**:299–314.

A. S. Goldberger (1964). *Econometric Theory.* New York: John Wiley & Sons.

S. L. Lauritzen (1996). *Graphical Models.* New York: The Clarendon Press [Oxford University Press].

S. L. Lauritzen (2002). gRaphical Models in R: A new initiative within the R project. *R News* **2(3)**, 39 (`http://cran.r-project.org/doc/Rnews/`).

J. Pearl (1988). *Probabilistic Reasoning in Intelligent Systems.* San Francisco: Morgan Kaufmann.

T. S. Richardson, P. Spirtes (2002). Ancestral graph Markov models. *Annals of Statistics* **30**:962-1030.

L. G. Telser (1964). Iterative estimation of a set of linear regression equations. *Journal of the American Statistical Association* **59**:845-862.

J. Whittaker (1990). *Graphical Models in Applied Multivariate Statistics.* Chichester: John Wiley & Sons.