Journal of Machine Learning Research 9 (2008) 893-914 Submitted 10/07; Revised 2/08; Published 5/08

Graphical Methods for Efficient Likelihood Inference in Gaussian Covariance Models

Mathias Drton DRTON@UCHICAGO.EDU
Department of Statistics
University of Chicago
5734 S. University Ave
Chicago, IL 60637, USA

Thomas S. Richardson TSR@STAT.WASHINGTON.EDU
Department of Statistics
University of Washington, Box 354322
Seattle, WA 98195-4322, USA

Editor: Max Chickering

Abstract

In graphical modelling, a bi-directed graph encodes marginal independences among random variables that are identified with the vertices of the graph. We show how to transform a bi-directed graph into a maximal ancestral graph that (i) represents the same independence structure as the original bi-directed graph, and (ii) minimizes the number of arrowheads among all ancestral graphs satisfying (i). Here the number of arrowheads of an ancestral graph is the number of directed edges plus twice the number of bi-directed edges. In Gaussian models, this construction can be used for more efficient iterative maximization of the likelihood function and to determine when maximum likelihood estimates are equal to empirical counterparts.

Keywords: ancestral graph, covariance graph, graphical model, marginal independence, maximum likelihood estimation, multivariate normal distribution

1. Introduction

In graphical modelling, bi-directed graphs encode marginal independences among random variables that are identified with the vertices of the graph (Pearl and Wermuth, 1994; Kauermann, 1996; Richardson, 2003). In particular, if two vertices are not joined by an edge, then the two associated random variables are assumed to be marginally independent. For example, the graph G in Figure 1, whose vertices are to be identified with a random vector (X_1, X_2, X_3, X_4), represents the pairwise marginal independences $X_1 \perp\!\!\!\perp X_3$, $X_1 \perp\!\!\!\perp X_4$, and $X_2 \perp\!\!\!\perp X_4$. While other authors (Cox and Wermuth, 1993, 1996; Edwards, 2000) have used dashed edges to represent marginal independences, the bi-directed graphs we employ here make explicit the connection to path diagrams (Wright, 1934; Koster, 1999).

Gaussian graphical models for marginal independence, also known as covariance graph models, impose zero patterns in the covariance matrix, which are linear hypotheses on the covariance matrix (Anderson, 1973). The graph in Figure 1, for example, imposes $\sigma_{13} = \sigma_{14} = \sigma_{24} = 0$. An estimation procedure designed for covariance graph models is described in Drton and Richardson (2003); see

Figure 1: A bi-directed graph G with (unique) minimally oriented graph $G^{\min}$.

also Chaudhuri et al. (2007). Other recent work involving these models includes Mao et al. (2004) and Wermuth et al. (2006).

In this paper we employ the connection between bi-directed graphs and the more general ancestral graphs with undirected, directed, and bi-directed edges (Section 2). For the statistical motivation of ancestral graphs see Richardson and Spirtes (2002); for causal interpretation see Richardson and Spirtes (2003). We show how to construct a maximal ancestral graph $G^{\min}$, which we call a minimally oriented graph, that is Markov equivalent to a given bi-directed graph G and such that the number of arrowheads is minimal (Sections 3–4). Two ancestral graphs are Markov equivalent if the independence models associated with the two graphs coincide; see for example Roverato (2005) for some recent results on Markov equivalence of different types of graphs. The number of arrowheads is the number of directed edges plus twice the number of bi-directed edges. Minimally oriented graphs provide useful nonparametric information about Markov equivalence of bi-directed, undirected and directed acyclic graphs. For example, the graph G in Figure 1 is not Markov equivalent to an undirected graph because $G^{\min}$ is not an undirected graph, and G is not Markov equivalent to a DAG because $G^{\min}$ contains a bi-directed edge. The graph G in Figure 1 has a unique minimally oriented graph but in general, minimally oriented graphs are not unique. Our construction procedure (Algorithm 14) involves a choice of a total order among the vertices. Varying the order one may obtain all minimally oriented graphs (Theorem 17).

For covariance graph models, minimally oriented graphs allow one to determine when the maximum likelihood estimate of a variance or covariance is available explicitly as its empirical counterpart (Section 5). For example, since no arrowheads appear at the vertices 1 and 4 in the graph $G^{\min}$ in Figure 1, the maximum likelihood estimates of σ_{11} and σ_{44} must be equal to the empirical variances of X_1 and X_4, respectively. The likelihood function for covariance graph models may be multi-modal, though simulations suggest this only occurs at small sample sizes, or under misspecification (Drton and Richardson, 2004a). However, when a minimally oriented graph reveals that a parameter estimate is equal to an empirical quantity (such as σ_{11} and σ_{44} in the above example) then even if the likelihood function is multi-modal this parameter will take the same value at every mode. Perhaps most importantly, minimally oriented graphs allow for computationally more efficient maximum likelihood fitting; see Remark 24 and the example in Section 5.3.

2. Ancestral Graphs and Their Global Markov Property

This paper deals with *simple mixed graphs*, which feature undirected ($v - w$), directed ($v \to w$) and bi-directed edges ($v \leftrightarrow w$) under the constraint that there is at most one edge between two vertices. In this section we give a formal definition of these graphs and discuss their Markov interpretation.

2.1 Simple Mixed Graphs

Let $\mathcal{E} = \{\emptyset, -, \leftarrow, \rightarrow, \leftrightarrow\}$ be the set of possible edges between an ordered pair of vertices; $\emptyset$ denoting that there is no edge. A *simple mixed graph* $G = (V, E)$ is a pair of a finite *vertex set* V and an *edge map* $E : V \times V \to \mathcal{E}$. The edge map E has to satisfy that for all $v, w \in V$,

(i) $E(v,v) = \emptyset$, that is, there is no edge between a vertex and itself,

(ii) $E(v,w) = E(w,v)$ if $E(v,w) \in \{-, \leftrightarrow\}$,

(iii) $E(v,w) = \rightarrow \iff E(w,v) = \leftarrow$.

In the sequel, we write $v - w \in G$, $v \to w \in G$, $v \leftarrow w \in G$ or $v \leftrightarrow w \in G$ if $E(v,w)$ equals $-$, $\rightarrow$, $\leftarrow$ or $\leftrightarrow$, respectively. If $E(v,w) \neq \emptyset$, then v and w are *adjacent*. If there is an edge $v \leftarrow w \in G$ or $v \leftrightarrow w \in G$ then there is an *arrowhead at* v on this edge. If there is an edge $v \to w \in G$ or $v - w \in G$ then there is a *tail at* v on this edge. A vertex w is in the *boundary* of v, denoted by $\mathrm{bd}(v)$, if v and w are adjacent. The boundary of vertex set $A \subseteq V$ is the set $\mathrm{bd}(A) = [\cup_{v \in A} \mathrm{bd}(v)] \setminus A$. We write $\mathrm{Bd}(v) = \mathrm{bd}(v) \cup \{v\}$ and $\mathrm{Bd}(A) = \mathrm{bd}(A) \cup A$. An induced subgraph of G over a vertex set A is the mixed graph $G_A = (A, E_A)$ where E_A is the restriction of the edge map E on $A \times A$. The *skeleton* of a simple mixed graph is obtained by making all edges undirected.

In a simple mixed graph a sequence of adjacent vertices $(v_1, \ldots, v_k)$ uniquely determines the sequence of edges joining consecutive vertices v_i and v_{i+1}, $1 \leq i \leq k-1$. Hence, we can define a *path* π between two vertices v and w as a sequence of distinct vertices $\pi = (v, v_1, \ldots, v_k, w)$ such that each vertex in the sequence is adjacent to its predecessor and its successor. A path $v \to \cdots \to w$ with all edges of the form $\rightarrow$ and pointing toward w is a directed path from v to w. If there is such a directed path from v to $w \neq v$, or if $v = w$, then v is an *ancestor* of w. We denote the set of all ancestors of a vertex v by $\mathrm{An}(v)$ and for a vertex set $A \subseteq V$ we define $\mathrm{An}(A) = \cup_{v \in A} \mathrm{An}(v)$. Finally, a directed path from v to w together with an edge $w \to v \in G$ is called a *directed cycle*.

Important subclasses of simple mixed graphs are illustrated in Figure 2. *Bi-directed*, *undirected* and *directed graphs* contain only one type of edge. *Directed acyclic graphs* (DAGs) are directed graphs without directed cycles. These three types of graph are special cases of *ancestral graphs* (Richardson and Spirtes, 2002).

Definition 1 *A simple mixed graph G is an* ancestral graph *if it holds that*

(i) G does not contain any directed cycles;

(ii) if $v - w \in G$, then there does not exist u such that $u \to v \in G$ or $u \leftrightarrow v \in G$;

(iii) if $v \leftrightarrow w \in G$, then v is not an ancestor of w.

2.2 Global Markov Property for Ancestral Graphs

Ancestral graphs can be given an independence interpretation, known as the global Markov property, by a graphical separation criterion called *m*-separation (Richardson and Spirtes, 2002, §3.4). An extension of Pearl's (1988) *d*-separation for DAGs, *m*-separation uses the notion of *colliders*. A non-endpoint vertex v_i on a path is a *collider on the path* if the edges preceding and succeeding v_i on the path both have an arrowhead at v_i, that is, $v_{i-1} \to v_i \leftarrow v_{i+1}$, $v_{i-1} \to v_i \leftrightarrow v_{i+1}$, $v_{i-1} \leftrightarrow v_i \leftarrow v_{i+1}$ or $v_{i-1} \leftrightarrow v_i \leftrightarrow v_{i+1}$ is part of the path. A non-endpoint vertex that is not a collider is a *non-collider on the path*.

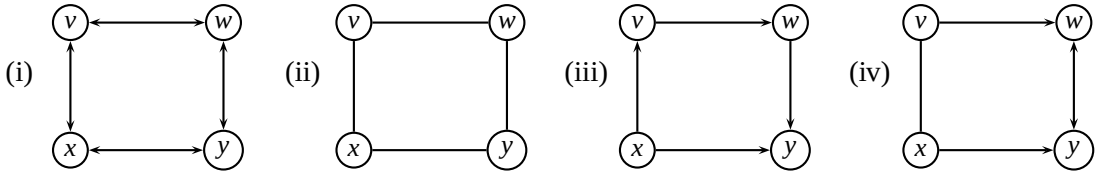

Figure 2: Simple mixed graphs. (i) A bi-directed graph, (ii) an undirected graph, (iii) a DAG, (iv) an ancestral graph.

Definition 2 *A path π between vertices v and w in a simple mixed graph G is m-*connecting *given a possibly empty set $C \subseteq V \setminus \{v,w\}$ if (i) every non-collider on π is not in C, and (ii) every collider on π is in* $\mathrm{An}(C)$. *If no path m-connects v and w given C, then v and w are m-*separated *given C. Two non-empty and disjoint sets A and B are m-separated given $C \subseteq V \setminus (A \cup B)$, if any two vertices $v \in A$ and $w \in B$ are m-separated given C.*

Let $G = (V,E)$ be an ancestral graph whose vertices index a random vector $(X_v \mid v \in V)$. For $A \subseteq V$, let X_A be the subvector $(X_v \mid v \in A)$. The *global Markov property* for G states that X_A is conditionally independent of X_B given X_C whenever A, B and C are pairwise disjoint subsets such that A and B are m-separated given C in G. Subsequently, we denote such conditional independence using the shorthand $A \perp\!\!\!\perp B \mid C$ that avoids making the probabilistic context explicit. The global Markov property, when applied to each of the graphs in Figure 2 in turn, implies (among other independences) that:

(i) $v \perp\!\!\!\perp y$ and $w \perp\!\!\!\perp x$;

(ii) $v \perp\!\!\!\perp y \mid \{w,x\}$ and $w \perp\!\!\!\perp x \mid \{v,y\}$;

(iii) $v \perp\!\!\!\perp y \mid \{w,x\}$ and $w \perp\!\!\!\perp x \mid v$;

(iv) $v \perp\!\!\!\perp y \mid x$ and $w \perp\!\!\!\perp x \mid v$.

If G is a bi-directed graph, then the global Markov property states the marginal independence $v \perp\!\!\!\perp w$ if v and w are not adjacent. In a multivariate normal distribution such pairwise marginal independences hold iff all independences stated by the global Markov property for G hold (Kauermann, 1996). Without any distributional assumption, Richardson (2003, §4) shows that the independences stated by the global Markov property of a bi-directed graph hold iff certain (not only pairwise) marginal independences hold; see also Matúš (1994). Two ancestral graphs G_1 and G_2 are *Markov equivalent* if they have the same vertex set and the global Markov property states the same independences for G_1 as for G_2.

The graphs in Figure 2 have the property that for every pair of non-adjacent vertices v and w there exists some subset C such that the global Markov property states that $v \perp\!\!\!\perp w \mid C$. Ancestral graphs with this property are called *maximal*. If an ancestral graph G is not maximal, then there exists a unique Markov equivalent maximal ancestral graph $\overline{G}$ that contains all the edges present in G. Moreover, any edge in $\overline{G}$ that is not present in G is bi-directed (Richardson and Spirtes, 2002, §3.7).

The following facts are easily established; see also Richardson and Spirtes (2002).

Lemma 3 *(i) Markov equivalent maximal ancestral graphs have the same skeleton.*

(ii) If $\bar{G}$ is an ancestral graph that is Markov equivalent to a maximal ancestral graph G and has the same skeleton as G, then $\bar{G}$ is also a maximal ancestral graph.

(iii) Bi-directed, undirected and directed acyclic graphs are maximal ancestral graphs.

2.3 Boundary Containment

In the subsequent Sections 3 and 4 we will construct maximal ancestral graphs that are Markov equivalent to a given bi-directed graph. Via Theorem 5 below, the following property plays a crucial role in these constructions.

Definition 4 *A simple mixed graph G has the* boundary containment property *if for all distinct vertices $v, w \in V$ the presence of an edge $v - w$ implies that* $\mathrm{Bd}(v) = \mathrm{Bd}(w)$ *and the presence of an edge $v \to w$ in G implies that* $\mathrm{Bd}(v) \subseteq \mathrm{Bd}(w)$.

In the Appendix we present lemmas on the structure of m-connecting paths in graphs with the boundary containment property. These lemmas yield the following key result.

Theorem 5 *If $\bar{G}$ is an ancestral graph that has the same skeleton as a bi-directed graph G, then G and $\bar{G}$ are Markov equivalent iff $\bar{G}$ has the boundary containment property.*

Proof Two vertices are adjacent in G iff they are adjacent in $\bar{G}$. Therefore, G and $\bar{G}$ are Markov equivalent iff it holds that two non-adjacent vertices v and w are m-connected given $C \subseteq V$ in G iff they are m-connected given C in $\bar{G}$.

($\Longrightarrow$:) Suppose $\bar{G}$ does not have the boundary containment property, that is, there exists an edge $v - w \in \bar{G}$ or an edge $v \to w \in \bar{G}$ such that $\mathrm{Bd}(v) \not\subseteq \mathrm{Bd}(w)$. Choose $u \in \mathrm{Bd}(v) \setminus \mathrm{Bd}(w)$. Since u and w are not adjacent, they are m-separated given $C = \emptyset$ in G. In $\bar{G}$, however, the path (u, v, w) m-connects u and w given $C = \emptyset$. Hence, G and $\bar{G}$ are not Markov equivalent.

($\Longleftarrow$:) First, let v and w be non-adjacent vertices that are m-connected given $C \subseteq V$ in $\bar{G}$. By Lemma 29, there is a path $\bar{\pi} = (v, v_1, \ldots, v_k, w)$ that m-connects v and w given C in $\bar{G}$ and is such that $v_1, \ldots, v_k$ are colliders with $\{v_1 \ldots, v_k\} \subseteq C$. Since G is a bi-directed graph, the corresponding path $\pi = (v, v_1, \ldots, v_k, w)$ in G also m-connects v and w given C.

Conversely, let v and w be non-adjacent vertices that are m-connected given $C \subseteq V$ in G. Let $\pi = (v_0, v_1, \ldots, v_k, v_{k+1})$ m-connect $v = v_0$ and $w = v_{k+1}$ given C in G such that no shorter path m-connects v and w given C. Then $v_1, \ldots, v_k$ are colliders, $\{v_1 \ldots, v_k\} \subseteq C$, and v_{i-1} and v_{i+1}, $i = 1, \ldots, k$, are not adjacent in G. (This is a special case of Lemmas 27 and 29 because a bi-directed graph trivially satisfies the boundary containment property.) It follows that, for all $i = 1, \ldots, k-1$, $v_{i-1} \in \mathrm{Bd}(v_i)$ but $v_{i-1} \notin \mathrm{Bd}(v_{i+1})$, and similarly $v_{i+2} \notin \mathrm{Bd}(v_i)$ but $v_{i+2} \in \mathrm{Bd}(v_{i+1})$. This implies that $\mathrm{Bd}(v_i) \not\subseteq \mathrm{Bd}(v_{i+1})$ and $\mathrm{Bd}(v_i) \not\supseteq \mathrm{Bd}(v_{i+1})$ for all $i = 1, \ldots, k-1$. Since $\bar{G}$ has the boundary containment property, it must hold that $v_i \leftrightarrow v_{i+1} \in \bar{G}$ for all $i = 1, \ldots, k-1$. Therefore, $v_2, \ldots, v_{k-1}$ are colliders on the path $\bar{\pi} = (v, v_1, \ldots, v_k, w)$ in $\bar{G}$. Similarly, it follows that $v_2 \in \mathrm{Bd}(v_1) \setminus \mathrm{Bd}(v)$, which entails $\mathrm{Bd}(v_1) \not\subseteq \mathrm{Bd}(v)$. Thus, v_1 is a collider on $\bar{\pi}$. Analogously, we can show that v_k is a collider on $\bar{\pi}$, which yields that $\bar{\pi}$ is a path in $\bar{G}$ that m-connects v and w given C. ■

3. Simplicial Graphs

In this section we show how simplicial vertex sets of a bi-directed graph can be used to construct a Markov equivalent maximal ancestral graph by removing arrowheads from certain bi-directed edges. Simplicial sets are also important in other contexts such as collapsibility (Madigan and Mosurski, 1990; Kauermann, 1996; Lauritzen, 1996, §2.1.3, p.121 and 219) and triangulation of graphs (Jensen, 2001, §5.3).

Definition 6 *A vertex $v \in V$ is* simplicial, *if* $\mathrm{Bd}(v)$ *is complete, that is, every pair of vertices in* $\mathrm{Bd}(v)$ *are adjacent. Similarly, a set $A \subseteq V$ is* simplicial, *if* $\mathrm{Bd}(A)$ *is complete.*

Simplicial vertices can be characterized in terms of boundary containment as follows.

Proposition 7 *A vertex $v \in V$ is simplicial iff* $\mathrm{Bd}(v) \subseteq \mathrm{Bd}(w)$ *for all* $w \in \mathrm{Bd}(v)$.

If an edge between v and w has an arrowhead at v, then we say that we *drop the arrowhead at v* when either $v \leftarrow w$ is replaced by $v - w$ or $v \leftrightarrow w$ is replaced by $v \rightarrow w$.

Definition 8 *Let G be a bi-directed graph. The* simplicial graph G^s *is the simple mixed graph obtained by dropping all the arrowheads at simplicial vertices of G.*

For the graph from Figure 1, G^s is equal to the depicted graph $G^{\min}$; additional examples are given in Figure 3. Parts (i) and (ii) of the next lemma show that simplicial graphs have the boundary containment property.

Lemma 9 *Let v and w be adjacent vertices in a simplicial graph G^s. Then*

(i) if $v - w \in G^s$, then $\mathrm{Bd}(v) = \mathrm{Bd}(w)$*;*

(ii) if $v \rightarrow w \in G^s$, then $\mathrm{Bd}(v) \subsetneq \mathrm{Bd}(w)$*;*

(iii) if $v \leftrightarrow w \in G^s$, then each of $\mathrm{Bd}(v) = \mathrm{Bd}(w)$, $\mathrm{Bd}(v) \subsetneq \mathrm{Bd}(w)$, *and* $\mathrm{Bd}(v) \not\subseteq \mathrm{Bd}(w) \not\subseteq \mathrm{Bd}(v)$ *might be the case.*

Proof (i) and (ii) follow from Proposition 7. For (iii) see, respectively, the graphs G^s_1, G^s_2 in Figure 3, and $G^s = G^{\min}$ in Figure 1. ■

Theorem 10 *The simplicial graph G^s of a bi-directed graph G is a maximal ancestral graph that is Markov equivalent to G.*

Proof By Lemma 3, Theorem 5 and Lemma 9, it suffices to show that G^s is an ancestral graph. This, however, follows from Lemma 11 below. ■

Lemma 11 *If G is an ancestral graph that has the boundary containment property, then dropping all arrowheads at simplicial vertices of G yields an ancestral graph.*

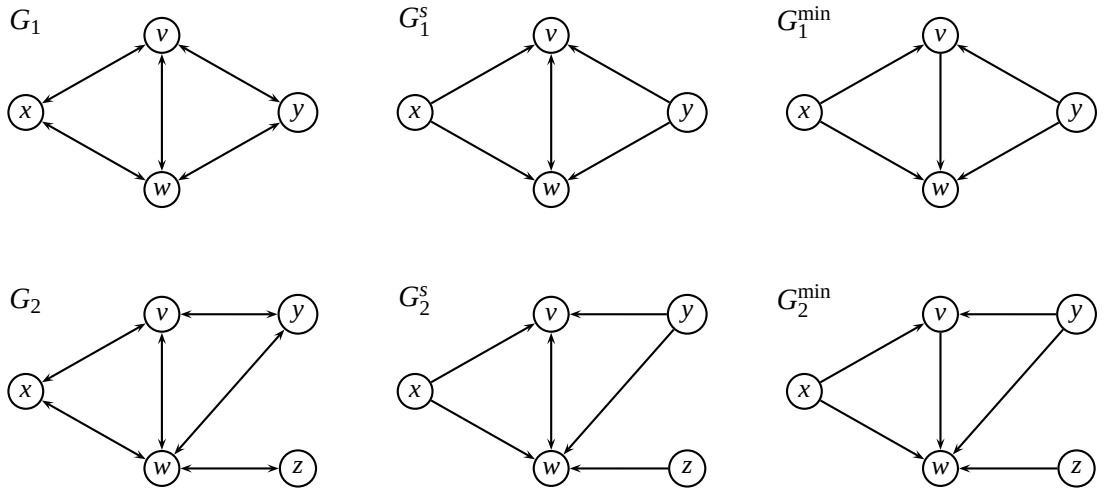

Figure 3: Bi-directed graphs with simplicial and minimally oriented graphs.

Proof Let $\bar{G}$ be the graph obtained by dropping the arrowheads at simplicial vertices. First, suppose $v \to w \in \bar{G}$ or $v \leftrightarrow w \in \bar{G}$ but that there is a path π from w to v that is a directed path in $\bar{G}$. Since there are no arrowheads at simplicial vertices in $\bar{G}$, no vertex on π including the endpoints v and w can be simplicial. This implies that π is a directed path from w to v in G. However, since $v \to w \in G$ or $v \leftrightarrow w \in G$, this is a contradiction to G being ancestral. We conclude that $\bar{G}$ satisfies conditions (i) and (iii) of Definition 1.

Next, suppose $v - w \in \bar{G}$ but that there exists another vertex u such that $u \to v \in \bar{G}$ or $u \leftrightarrow v \in \bar{G}$. It follows that v is not simplicial. Since G is ancestral, this implies that $v \to w \in G$ which in turn implies that $\mathrm{Bd}(v) \subseteq \mathrm{Bd}(w)$ because G has the boundary containment property. The set $\mathrm{Bd}(v)$ is not complete because v is not simplicial. Thus $\mathrm{Bd}(w)$ is not complete, that is, w is not a simplicial vertex. However, this is a contradiction to the fact that $v \to w \in G$ but $v - w \in \bar{G}$. Thus, $\bar{G}$ is indeed an ancestral graph. ■

Proposition 12 *A bi-directed graph G is Markov equivalent to an undirected graph iff the simplicial graph G^s induced by G is an undirected graph iff G is a disjoint union of complete (bi-directed) graphs.*

Proof If G^s is an undirected graph, then by Theorem 10, G is Markov equivalent to an undirected graph, namely G^s. Conversely, assume that there exists an undirected graph U that is Markov equivalent to G. Necessarily, G and U have the same skeleton (recall Lemma 3). By Theorem 5, U has the boundary containment property, which implies that every vertex is simplicial and thus that G^s is an undirected graph (and equal to U).

The simplicial graph G^s is an undirected graph iff the vertex set of the inducing bi-directed graph G can be partitioned into pairwise disjoint sets $A_1, \ldots, A_q$ such that (a) if $v \in A_i$, $1 \leq i \leq q$, and $w \in A_j$, $1 \leq j \leq q$, are adjacent, then $i = j$, and (b) all the induced subgraphs G_{A_i}, $i = 1, \ldots, q$ are complete graphs (Kauermann, 1996). ■

Under multivariate normality, a bi-directed graph that is Markov equivalent to an undirected graph represents a hypothesis that is linear in the covariance matrix as well as in its inverse. The general structure of such models is studied in Jensen (1988).

4. Minimally Oriented Graphs

The simplicial graph G^s sometimes may be a DAG. For example, the graph $u \leftrightarrow v \leftrightarrow w$ has the simplicial graph $u \rightarrow v \leftarrow w$. However, there exist bi-directed graphs that are Markov equivalent to a DAG and yet the simplicial graph contains bi-directed edges. For example, the graph G_1 in Figure 3 is Markov equivalent to the DAG $G_1^{\min}$ in the same Figure. Hence, some arrowheads may be dropped from bi-directed edges in a simplicial graph while preserving Markov equivalence. In this section we construct maximal ancestral graphs from which no arrowheads may be dropped without destroying Markov equivalence.

4.1 Definition and Construction

The following definition introduces the key object of this section.

Definition 13 *Let G be a bi-directed graph. A* minimally oriented graph *of G is a graph $G^{\min}$ that satisfies the following three properties:*

(i) *$G^{\min}$ is a maximal ancestral graph;*

(ii) *G and $G^{\min}$ are Markov equivalent;*

(iii) *$G^{\min}$ has the minimum number of arrowheads of all maximal ancestral graphs that are Markov equivalent to G. Here the number of arrowheads of an ancestral graph G with d directed and b bi-directed edges is defined as* $\mathrm{arr}(G) = d + 2b$.

By Lemma 3, a minimally oriented graph $G^{\min}$ has the same skeleton as the underlying bi-directed graph G. According to Theorem 5, $G^{\min}$ has the boundary containment property. Examples of minimally oriented graphs are shown in Figure 3. Given the small number of vertices of these graphs the claim that these graphs are indeed minimally oriented graphs can be verified directly. The example of graph G_1 in Figure 3 also illustrates that minimally oriented graphs are not unique. By symmetry, reversing the direction of the edge $v \rightarrow w$ in the depicted $G_1^{\min}$ yields a second minimally oriented graph for G_1.

We now turn to the problem of how to construct a minimally oriented graph. Define a relation on the vertex set V of the given bi-directed graph G by letting $v \preccurlyeq_B w$ if $v = w$ or if $\mathrm{Bd}(v) \subsetneq \mathrm{Bd}(w)$ in G. The relation $\preccurlyeq_B$ is a partial order and can thus be extended to a total order $\leq$ on V such that the strict boundary containment $\mathrm{Bd}(v) \subsetneq \mathrm{Bd}(w)$ implies that $v < w$. In general, the choice of such an extension to a total order is not unique.

Algorithm 14 *Let G be a bi-directed graph, and $\leq$ a total order on V that extends the partial order $\preccurlyeq_B$ obtained from strict boundary containment. Create a new graph $G_<^{\min}$ as follows:*

(a) *find the simplicial graph G^s of G;*

(b) set $G^{\min}_{<} = G^s$*;*

(c) replace every bi-directed edge $v \leftrightarrow w \in G^{\min}_{<}$ *with* $\mathrm{Bd}(v) \subseteq \mathrm{Bd}(w)$ *and* $v < w$ *by the directed edge* $v \rightarrow w$*.*

The notation $G^{\min}_{<}$ indicates the dependence of this graph on *both* the bi-directed graph G and the total order $\leq$. Clearly, by Theorem 5, in order for $G^{\min}_{<}$ to be a minimally oriented graph it is necessary that it satisfies the boundary containment property. The next lemma shows that this is true.

Lemma 15 *Let G be a bi-directed graph and $G^{\min}_{<}$ the graph constructed in Algorithm 14. It then holds that*

(i) if $v - w$ *is an undirected edge in* $G^{\min}_{<}$*, then* $\mathrm{Bd}(v) = \mathrm{Bd}(w)$*;*

(ii) if $v \rightarrow w$ *is a directed edge in* $G^{\min}_{<}$*, then* $\mathrm{Bd}(v) \subseteq \mathrm{Bd}(w)$*;*

(iii) $v \leftrightarrow w$ *is a bi-directed edge in* $G^{\min}_{<}$ *iff* $\mathrm{Bd}(v) \not\subseteq \mathrm{Bd}(w) \not\subseteq \mathrm{Bd}(v)$*.*

Proof (i) follows directly from Lemma 9(i) because it follows from Algorithm 14 that $G^{\min}_{<}$ and G^s contain the same undirected edges.

(ii) If the edge $v \rightarrow w$ is already present in G^s, then $\mathrm{Bd}(v) \subsetneq \mathrm{Bd}(w)$ according to Lemma 9(ii). If $v \rightarrow w$ is not already present in G^s, then $v < w$ and $\mathrm{Bd}(v) \subseteq \mathrm{Bd}(w)$.

(iii) Suppose v and w are two adjacent vertices such that $\mathrm{Bd}(v) \not\subseteq \mathrm{Bd}(w) \not\subseteq \mathrm{Bd}(v)$. Then $v \leftrightarrow w$ in G^s and this edge cannot be replaced by a directed edge in step (c) of Algorithm 14. For the converse, consider two adjacent vertices v and w such that $\mathrm{Bd}(v) \subseteq \mathrm{Bd}(w)$. (The other case is symmetric.) If $v < w$, then according to the definition of the simplicial graph and step (c) of Algorithm 14 the edge between v and w in $G^{\min}_{<}$ cannot have an arrowhead at v and thus cannot be bi-directed. If $v > w$, then $\mathrm{Bd}(v) = \mathrm{Bd}(w)$ because $\mathrm{Bd}(v) \subsetneq \mathrm{Bd}(w)$ would imply $v < w$. It follows that the edge between v and w in $G^{\min}_{<}$ cannot be bi-directed as the arrowhead at w would be removed in step (c). ■

By Lemma 15(iii), $v \leftrightarrow w \in G^{\min}_{<}$ iff there exist vertices $x \in \mathrm{bd}(v) \setminus \{w\}$ and $y \in \mathrm{bd}(w) \setminus \{v\}$ such that the induced subgraph $G_{\{x,y,v,w\}}$ equals one of the two graphs shown in Figure 4. Graphs that do not contain the four-cycle from Figure 4(ii) as an induced subgraph are known as chordal or decomposable and play an important role in graphical modelling (Lauritzen, 1996). Graphs not containing the path from Figure 4(i) as an induced subgraph are called cographs and have favorable computational properties (Brandstädt et al., 1999). For instance, cographs can be recognized in linear time (Corneil et al., 1985).

Theorem 16 *The graph $G^{\min}_{<}$ constructed in Algorithm 14 is a minimally oriented graph for the bi-directed graph G.*

Proof We verify the conditions (i) and (iii) of Definition 13. This is sufficient because $G^{\min}_{<}$ has the boundary containment property (Lemma 15) and thus condition (i) implies condition (ii) by Theorem 5.

(i) $G^{\min}_{<}$ is a maximal ancestral graph:
By Lemma 3 it suffices to show that $G^{\min}_{<}$ is an ancestral graph. Let v and w be adjacent vertices

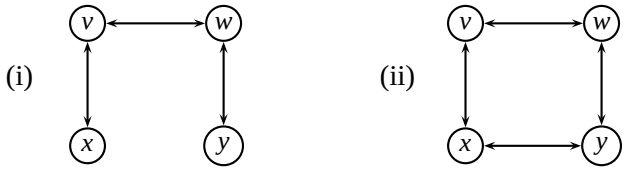

Figure 4: Induced subgraphs for which no arrowhead can be dropped from edge $v \leftrightarrow w$.

such that $v - w \in G^{\min}_{<}$. This is equivalent to $v - w \in G^s$, and it follows that there does not exist an arrowhead at v or w; compare the proof of Theorem 10. Furthermore, $G^{\min}_{<}$ does not contain any directed cycles because Algorithm 14 ensures that the presence of a directed edge $v \to w \in G^{\min}_{<}$ implies $v < w$ in the total order. Finally, assume that there exists $v \leftrightarrow w \in G^{\min}_{<}$. Then there cannot be a directed path from v to w, since by Lemma 15(ii) this would imply $\mathrm{Bd}(v) \subseteq \mathrm{Bd}(w)$, contradicting Lemma 15(iii).

(iii) $G^{\min}_{<}$ has the minimal number of arrowheads:
Let $\bar{G}$ be a maximal ancestral graph that is Markov equivalent to the (bi-directed) graph G, which requires that $\bar{G}$ and G, and thus also $G^{\min}_{<}$ have the same skeleton. Assume that $\mathrm{arr}(\bar{G}) < \mathrm{arr}(G^{\min}_{<})$. Then either (a) there exists $v \to w \in G^{\min}_{<}$ such that $v - w \in \bar{G}$ or (b) there exists $v \leftrightarrow w \in G^{\min}_{<}$ such that $v \to w \in \bar{G}$ or $v - w \in \bar{G}$.

Case (a): If $v \to w \in G^{\min}_{<}$, then w cannot be simplicial. Hence, there exist two vertices $x, y \in \mathrm{bd}(w)$ that are not adjacent in $G^{\min}_{<}$, and thus not adjacent in G; ($v = x$ is possible). The global Markov property of G states that $x \perp\!\!\!\perp y$. Since $\bar{G}$ is an ancestral graph and $v - w \in \bar{G}$, however, there may not be any arrowheads at w on the edges between x and w, and y and w in $\bar{G}$. Therefore, x and y are m-connected given $\emptyset$ in $\bar{G}$, which yields that the global Markov property of $\bar{G}$ does not imply $x \perp\!\!\!\perp y$; a contradiction.

Case (b): Suppose $v \leftrightarrow w \in G^{\min}_{<}$ but there is no arrowhead at v on the edge between v and w in $\bar{G}$. By Lemma 15(iii) there exists $x \in \mathrm{bd}(v) \setminus \mathrm{Bd}(w)$ such that x and w are not adjacent in $G^{\min}_{<}$. Thus x and w are not adjacent in G and $x \perp\!\!\!\perp w$ is stated by the global Markov property for G. In $\bar{G}$, however, v is a non-collider on the path (x, v, w) and thus this path m-connects x and w given $\emptyset$, which yields that the global Markov property of $\bar{G}$ does not imply $x \perp\!\!\!\perp w$; a contradiction. ■

The next result shows that our construction of minimally oriented graphs is complete in the sense that every minimally oriented graph can be obtained as the output of Algorithm 14 by appropriate choice of a total order on the vertex set.

Theorem 17 *If $G^{\min}$ is a minimally oriented graph for a bi-directed graph G, then there exists a total order $\leq$ on the vertex set such that $G^{\min} = G^{\min}_{<}$.*

Proof The graph $G^{\min}$ is an ancestral graph and thus contains no directed cycles. Hence, the directed edges in $G^{\min}$ yield a partial order $\preccurlyeq_D$ on the vertex set V in which $v \preccurlyeq_D w$ if $v = w$ or if there is a directed path from v to w. Define the relation $\preccurlyeq_{BD}$ by letting $v \preccurlyeq_{BD} w$ if $v \preccurlyeq_B w$ or $v \preccurlyeq_D w$. Clearly, $v \preccurlyeq_{BD} v$, that is, the relation is reflexive. We claim that the relation is in fact a partial order.

By Theorem 5, $G^{\min}$ has the boundary containment property such that $\mathrm{Bd}(v) \subseteq \mathrm{Bd}(w)$ if $v \preccurlyeq_D w$. Consequently, if $v \neq w$ then $v \preccurlyeq_D w$ implies $w \not\preccurlyeq_B v$ and $v \preccurlyeq_B w$ implies $w \not\preccurlyeq_D v$. This implies that

$\preccurlyeq_{BD}$ is anti-symmetric. In order to verify transitivity, it suffices to consider three distinct vertices satisfying $v \preccurlyeq_D w \preccurlyeq_B u$ or $v \preccurlyeq_B w \preccurlyeq_D u$. In the former case $\mathrm{Bd}(v) \subseteq \mathrm{Bd}(w) \subsetneq \mathrm{Bd}(u)$, and in the latter case $\mathrm{Bd}(v) \subsetneq \mathrm{Bd}(w) \subseteq \mathrm{Bd}(u)$. In both cases $\mathrm{Bd}(v) \subsetneq \mathrm{Bd}(u)$ such that $v \preccurlyeq_B u$, which implies the required conclusion $v \preccurlyeq_{BD} u$.

We can now choose a total order $\leq$ on V that extends the partial order $\preccurlyeq_{BD}$ and thus extends both $\preccurlyeq_B$ and $\preccurlyeq_D$. Let $G^{\min}_{<}$ be the output of Algorithm 14 when the bi-directed graph G and the chosen total order $\leq$ are given as the input. We claim that $G^{\min} = G^{\min}_{<}$.

First note that if v is a simplicial vertex of G, then there are no arrowheads at v in $G^{\min}$. Otherwise, we could drop all arrowheads at simplicial vertices in $G^{\min}$ to obtain an ancestral graph (Lemma 11) with fewer arrowheads. The new graph would have the boundary containment property and thus be Markov equivalent to G (by Theorem 5). This would contradict the assumed minimality of $G^{\min}$.

The observation about simplicial vertices implies that an undirected edge in the simplicial graph G^s is also an undirected edge in $G^{\min}$. Conversely, if $v - w \in G^{\min}$ then there may not be an arrowhead at v on any other edge, and likewise for w, because $G^{\min}$ is ancestral. Since $G^{\min}$ has the boundary containment property, it follows from Proposition 7 that both v and w are simplicial vertices. This implies that $v - w \in G^s$ and we conclude that $G^{\min}$ and G^s have the same undirected edges. By construction, the same holds for $G^{\min}_{<}$ and G^s. Hence, $G^{\min}$ and $G^{\min}_{<}$ have the same undirected edges.

Suppose $v \to w \in G^{\min}$. Then $\mathrm{Bd}(v) \subseteq \mathrm{Bd}(w)$ because $G^{\min}$ has the boundary containment property. Moreover, $v < w$ because the total order $\leq$ extends $\preccurlyeq_D$. It follows that $v \to w \in G^{\min}_{<}$. In other words, every directed edge in $G^{\min}$ is also in $G^{\min}_{<}$. This together with the fact that $G^{\min}$ and $G^{\min}_{<}$ have the same skeleton and the same number of arrowheads, $\mathrm{arr}(G^{\min}) = \mathrm{arr}(G^{\min}_{<})$, implies that $G^{\min} = G^{\min}_{<}$. ■

4.2 Markov Equivalence Results

The following corollary is an immediate consequence of Proposition 12 because a minimally oriented graph $G^{\min}$ is an undirected graph iff G^s is an undirected graph.

Corollary 18 *Let $G^{\min}$ be a minimally oriented graph for a bi-directed graph G. If G is Markov equivalent to an undirected graph U, then $G^{\min} = U$ is the unique minimally oriented graph of G.*

A minimally oriented graph also reveals whether the original bi-directed graph is Markov equivalent to a DAG.

Theorem 19 *Let $G^{\min}$ be a minimally oriented graph for a bi-directed graph G. Then G is Markov equivalent to a DAG iff $G^{\min}$ contains no bi-directed edges.*

Proof Let G be a bi-directed graph such that $G^{\min}$ contains no bi-directed edges. If $A \subseteq V$ is a simplicial set, then the induced subgraph $(G^{\min})_A$ is undirected and complete (this follows directly from Theorem 17 and Algorithm 14). Let $A_1, \ldots, A_q$ be the inclusion-maximal simplicial sets of G. Let D be a directed graph obtained by replacing each induced subgraph $(G^{\min})_{A_i}$, $i = 1, \ldots, q$, by a complete DAG. Then D itself is acyclic, which can be seen as follows: First, since $G^{\min}$ does not contain any directed cycles, a directed cycle π in D must involve a vertex $v \in \cup_{i=1}^{q} A_i$. Let $v \in A_j$.

Since the induced subgraphs D_{A_i}, $i = 1,\dots,q$, are all acyclic, π must also involve a vertex not in A_j. Therefore, there exists an edge $x \to w$ on π such that $w \in A_j$ and $x \notin A_j$. Since the sets A_i are inclusion-maximal simplicial sets, no vertex in A_i, $i \neq j$, is adjacent to any vertex in A_j. Hence, $x \notin \cup_{i=1}^{q} A_i$, which implies that the edge $x \to w$ is also present in $G^{\min}$. This is a contradiction to w being a simplicial vertex.

Two vertices are adjacent in $G^{\min}$ iff they are adjacent in D. Moreover, D has the boundary containment property because $G^{\min}$ has this property, and if $u \to \bar{u}$ in D then either $u \to \bar{u}$ in $G^{\min}$ or $u - \bar{u}$ in $G^{\min}$. It thus follows from Theorem 5 that D is Markov equivalent to $G^{\min}$ and G.

Conversely, suppose that $v \leftrightarrow w \in G^{\min}$ and for a contradiction, that G is Markov equivalent to a DAG D. Note that D must have the same skeleton as G (and $G^{\min}$). By Lemma 15(iii), there exist two different vertices $x \in \mathrm{bd}(v) \setminus \{w\}$ and $y \in \mathrm{bd}(w) \setminus \{v\}$ such that, by the Markov property of G, $x \perp\!\!\!\perp w$ and $v \perp\!\!\!\perp y$. Hence, v and w must be colliders on the paths (x,v,w) and (v,w,y) in D, respectively. This is impossible in the DAG D. ■

Theorem 19 can be shown to be equivalent to a Markov equivalence result stated without proof in Theorem 1 in Pearl and Wermuth (1994). This latter theorem requires 'no chordless four-chain', which must be read as excluding graphs with induced subgraphs that are either of the graphs in Figure 4. Under this condition, Pearl and Wermuth (1994) also state that a Markov equivalent DAG can be constructed from the (undirected) skeleton of G by introducing directed and bi-directed edges in an operation they term 'sink orientation', and turning remaining undirected edges into directed ones. The sink orientation of the graph G_1 in Figure 3 has the directed edges of G_1^s but an undirected edge $v - w$. Thus sink orientation need not yield an ancestral graph. The bi-directed graphical models considered in Theorem 19 also appear in the construction of generalized Wishart distributions (Letac and Massam, 2007, Theorem 2.2). In that context the models are called *homogeneous* and characterized in terms of Hasse diagrams.

As the next result reveals, bi-directed graphs that are Markov equivalent to DAGs exhibit a structure that corresponds to a multivariate regression model. The graphs can also be termed chordal cographs; compare the paragraph before Theorem 16.

Proposition 20 *Let $G^{\min}$ be a minimally oriented graph for a connected bi-directed graph G. If $G^{\min}$ contains no bi-directed edges, then the set A of all simplicial vertices is non-empty, the induced subgraph $(G^{\min})_A$ is a disjoint union of complete undirected graphs, the induced subgraph $(G^{\min})_{V\setminus A}$ is a complete DAG, and an edge $v \to w$ joins any two vertices $v \in A$ and $w \notin A$ in $G^{\min}$.*

Proof For two adjacent vertices v and w in $G^{\min}$, Lemma 15(i)-(ii) implies that $\mathrm{Bd}(v) \subseteq \mathrm{Bd}(w)$ or $\mathrm{Bd}(w) \subseteq \mathrm{Bd}(v)$. Hence, we can list the vertex set as $V = \{v_1,\dots,v_p\}$ such that $\mathrm{Bd}(v_i) \subseteq \mathrm{Bd}(v_j)$ if v_i and v_j are adjacent and $i \leq j$. It follows that $v_1 \in A$ and thus $A \neq \emptyset$. Let $A_1,\dots,A_q$ be the inclusion-maximal simplicial sets of G. Then $(G^{\min})_A$ equals the union of the disjoint complete undirected graphs $(G^{\min})_{A_1},\dots,(G^{\min})_{A_q}$. Since $G^{\min}$ is an ancestral graph, $(G^{\min})_{V\setminus A}$ is a DAG.

We prove the remaining claims by induction on $|V \setminus A|$. If $|V \setminus A| = 0$, then the connected graph $G^{\min}$ is a complete undirected graph and there is nothing to show. Let $|V \setminus A| \geq 1$. It follows that $v_p \in V \setminus A$. If the shortest path between some vertex v_{i_1} and v_p in G is of the form $v_{i_1} \leftrightarrow \dots \leftrightarrow v_{i_k} \leftrightarrow v_p$, then $i_1 < \dots < i_k < p$ and $\mathrm{Bd}(v_{i_1}) \subseteq \dots \subseteq \mathrm{Bd}(v_{i_k}) \subseteq \mathrm{Bd}(v_p)$, which is easily shown by induction on k. However, since $v_{i_1} \in \mathrm{Bd}(v_{i_1})$ it must in fact hold that v_{i_1} and v_p are adjacent. Hence, there is an edge between every vertex $v \in V \setminus \{v_p\}$ and v_p, which for $v \in A$ is of

the form $v \to v_p$ because clearly $v_p \notin A$. The proof is finished by combining what we learned about v_p with the induction assumption applied to the induced subgraph G_W with $W = \{v_1, \ldots, v_{p-1}\}$. Note that for $v, w \in W$, the inclusion $\mathrm{Bd}_G(v) \subseteq \mathrm{Bd}_G(w)$ implies that $\mathrm{Bd}_{G_W}(v) \subseteq \mathrm{Bd}_{G_W}(w)$. Thus by Lemma 15 and Theorem 16, $(G_W)^{\min}$ does not contain any bi-directed edges. ■

5. Maximum Likelihood Estimation in Gaussian Models

In this section we consider the Gaussian covariance models associated with bi-directed graphs and demonstrate that the graphical constructions from Sections 3 and 4 can be employed for more efficient computation of maximum likelihood estimates.

5.1 Covariance Graphs and Gaussian Ancestral Graph Models

Let G be a bi-directed graph, and

$$\mathbf{P}(G) = \left\{ \Sigma \in \mathbb{R}^{V \times V} \mid \Sigma = (\sigma_{vw}) \text{ sym. pos. def., } \sigma_{vw} = 0 \; \forall (v,w) : v \leftrightarrow w \notin G \right\}$$

be the cone of symmetric positive definite matrices with zero pattern induced by G. The *covariance graph model* associated with G is the family of multivariate normal distributions $\mathbf{N}(G) = \big(\mathcal{N}(0,\Sigma) \mid \Sigma \in \mathbf{P}(G)\big)$. It can be shown that every distribution in $\mathbf{N}(G)$ satisfies all conditional independences stated by the global Markov property for the bi-directed graph G (Kauermann, 1996, Prop. 2.2). Conversely, if a distribution $\mathcal{N}(0,\Sigma)$ satisfies the global Markov property for G, then $\Sigma \in \mathbf{P}(G)$.

Let $S \in \mathbb{R}^{V \times V}$ be the empirical covariance matrix computed from an i.i.d. sample drawn from some unknown distribution $\mathcal{N}(0,\Sigma) \in \mathbf{N}(G)$, that is, the (v,w)-th entry in S is the dot-product of the vectors of observations for the v-th and w-th variables divided by the sample size n. The log-likelihood function $\ell_{S,n} : \mathbf{P}(G) \to \mathbb{R}$ of $\mathbf{N}(G)$ can be written as

$$\ell_{S,n}(\Sigma) = -\frac{n|V|}{2}\log(2\pi) - \frac{n}{2}\log|\Sigma| - \frac{n}{2}\mathrm{tr}\left(\Sigma^{-1}S\right).$$

If S is positive definite then the global maximum of $\ell_{S,n}$ over $\mathbf{P}(G)$ exists. The likelihood equations obtained by setting to zero the partial derivatives of $\ell_{S,n}$ with respect to the non-restricted entries in Σ take on the form

$$(\Sigma^{-1})_{vw} = (\Sigma^{-1}S\Sigma^{-1})_{vw} \quad \forall v,w \in V : \; v = w \;\text{ or }\; v \leftrightarrow w \in G; \tag{1}$$

compare Anderson and Olkin (1985, §2.1.1). A matrix $\hat{\Sigma}(S) \in \mathbf{P}(G)$ that solves (1) is *a solution to the likelihood equations* of $\mathbf{N}(G)$. Since subsequent theorems on the structure of the likelihood equations are obtained via Gaussian ancestral graph models, we briefly review the parametrization of these models.

Let G be an ancestral graph and $\mathrm{un}_G \subseteq V$ the set of vertices v that are such that any edge with endpoint v has a tail at v. By Definition 1(i), $v - w \in G$ implies $v, w \in \mathrm{un}_G$, and $v \leftrightarrow w \in G$ implies that $v, w \notin \mathrm{un}_G$. Let Λ be a symmetric positive definite $\mathrm{un}_G \times \mathrm{un}_G$ matrix such that $\Lambda_{vw} \neq 0$ only if $v = w$ or $v - w \in G$. Let Ω be a symmetric positive definite $(V \setminus \mathrm{un}_G) \times (V \setminus \mathrm{un}_G)$ matrix such that $\Omega_{vw} \neq 0$ only if $v = w$ or $v \leftrightarrow w \in G$. Finally, let B be a $V \times V$ matrix such that $B_{vw} \neq 0$ only if

$w \to v \in G$. Define the symmetric positive definite matrix

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

where I is the identity matrix.

Let $\mathbf{N}(G)$ be the Gaussian ancestral graph model associated with G, that is, the family of all centered normal distributions that are globally Markov with respect to G. As shown in Richardson and Spirtes (2002, §8), the normal distribution $\mathcal{N}(0, \Sigma)$ with $\Sigma = \Sigma(\Lambda, B, \Omega)$ defined in (2) is in $\mathbf{N}(G)$. Conversely, if G is *maximal*, then for any $\mathcal{N}(0, \Sigma) \in \mathbf{N}(G)$ there exist unique Λ, Ω, B of the above type such that $\Sigma = \Sigma(\Lambda, B, \Omega)$. (Note that Richardson and Spirtes, 2002, use B for what is here denoted by $I - B$).

Since a bi-directed graph G and a minimally oriented graph $G^{\min}$ are Markov equivalent and maximal, the parametrization map for $G^{\min}$, $(\Lambda, B, \Omega) \mapsto \Sigma(\Lambda, B, \Omega)$, has image equal to $\mathbf{P}(G)$. By Richardson and Spirtes (2002, Theorem 8.14, Lemma 8.22), we obtain the following Lemma.

Lemma 21 *Let G be a bi-directed graph. The covariance matrix $\Sigma(\Lambda, B, \Omega)$ solves the likelihood equations of $\mathbf{N}(G)$ iff (Λ, B, Ω) solves the likelihood equations of $\mathbf{N}(G^{\min})$.*

5.2 Empirical Maximum Likelihood Estimates

Using the graphical results established earlier, we can show that over simplicial sets a solution to the likelihood equations (1) agrees with its empirical counterpart in S.

Theorem 22 *Let G be a bi-directed graph with associated covariance graph model $\mathbf{N}(G)$. If $A \subseteq V$ is simplicial, S is a symmetric positive definite matrix, and $\hat{\Sigma}(S) \in \mathbf{P}(G)$ is a solution to the likelihood equations (1), then $\hat{\Sigma}(S)_{A \times A} = S_{A \times A}$.*

Proof By Theorem 10, the covariance graph model $\mathbf{N}(G)$ and the Gaussian ancestral graph model $\mathbf{N}(G^s)$ based on the simplicial graph G^s are equal. Let $\mathbf{N}(G^s)$ be parametrized by the precision matrix Λ, the matrix of regression coefficients B and the covariance matrix Ω as described in §5.1. In particular, it follows from Richardson and Spirtes (2002, Lemma 8.4) that if $\Sigma = \Sigma(\Lambda, B, \Omega)$, then $(\Lambda^{-1})_{A \times A} = \Sigma_{A \times A}$.

The inclusion-maximal simplicial sets $A_1, \ldots, A_q$ of G form a partition of un_{G^s}. The induced subgraphs $G^s_{A_i}$, $i = 1, \ldots, q$, are complete undirected graphs. It follows that Λ is a block-diagonal matrix such that $\Lambda_{vw} = 0$ if there does not exist an inclusion-maximal simplicial set A_i such that $v, w \in A_i$. Now the discussion in Richardson and Spirtes (2002, §8.5) and Lemma 21 imply that every solution to the likelihood equations for Λ, B, Ω in the Gaussian ancestral graph model $\mathbf{N}(G^s)$ satisfies that $(\hat{\Lambda}^{-1})_{A_i \times A_i} = S_{A_i \times A_i}$ for all $i = 1, \ldots, q$. Since $A \subseteq A_j$ for some j, it holds that $\hat{\Sigma}_{A \times A} = (\hat{\Lambda}^{-1})_{A \times A} = S_{A \times A}$. ■

Our graphical constructions also provide information on when maximum likelihood estimates of conditional parameters are equal to their empirical counterparts. The conditional parameters we consider are the regression coefficients and conditional variance for the conditional distribution of variable v given its *parents* $\mathrm{pa}(v) = \{w \in V \mid w \to v \in G^{\min}\}$ in a minimally oriented graph $G^{\min}$. If $\mathrm{pa}(v) = \emptyset$, then conditioning variable v on $\mathrm{pa}(v)$ is understood to yield the marginal distribution of v.

Theorem 23 *Let $G^{\min}$ be a minimally oriented graph for a bi-directed graph G, S a symmetric positive definite matrix, and $\hat{\Sigma}(S) \in \mathbf{P}(G)$ a solution to the likelihood equations (1). If v is a vertex such that there is no vertex w with $v \leftrightarrow w \in G^{\min}$, then the regression coefficients for v given $\mathrm{pa}(v)$ are*

$$\hat{\Sigma}(S)_{v\times \mathrm{pa}(v)} \left[\hat{\Sigma}(S)_{\mathrm{pa}(v)\times \mathrm{pa}(v)}\right]^{-1} = S_{v\times \mathrm{pa}(v)} \left(S_{\mathrm{pa}(v)\times \mathrm{pa}(v)}\right)^{-1}, \tag{3}$$

and that the conditional variance for v given $\mathrm{pa}(v)$ is

$$\hat{\Sigma}(S)_{vv} - \hat{\Sigma}(S)_{v\times \mathrm{pa}(v)} \left[\hat{\Sigma}(S)_{\mathrm{pa}(v)\times \mathrm{pa}(v)}\right]^{-1} \hat{\Sigma}(S)_{\mathrm{pa}(v)\times v} = S_{vv} - S_{v\times \mathrm{pa}(v)} \left(S_{\mathrm{pa}(v)\times \mathrm{pa}(v)}\right)^{-1} S_{\mathrm{pa}(v)\times v}. \tag{4}$$

Proof If $\mathrm{pa}(v) = \emptyset$, then v is a simplicial vertex, and the claim reduces to $\hat{\Sigma}(S)_{vv} = S_{vv}$, which follows from Theorem 22. Otherwise, using the parametrization of $\mathbf{N}(G^{\min})$, it follows from Richardson and Spirtes (2002, Theorem 8.7) that if $\Sigma = \Sigma(\Lambda, B, \Omega)$, then

$$\Sigma_{v\times \mathrm{pa}(v)} \left[\Sigma_{\mathrm{pa}(v)\times \mathrm{pa}(v)}\right]^{-1} = B_{v\times \mathrm{pa}(v)}$$

and

$$\Sigma_{vv} - \Sigma_{v\times \mathrm{pa}(v)} \left[\Sigma_{\mathrm{pa}(v)\times \mathrm{pa}(v)}\right]^{-1} \Sigma_{\mathrm{pa}(v)\times v} = \Omega_{vv}.$$

If $\hat{\Lambda}$, $\hat{B}$, $\hat{\Omega}$ solve the likelihood equations for $\mathbf{N}(G^{\min})$, then $\hat{B}_{v\times \mathrm{pa}(v)}$ and $\hat{\Omega}_{vv}$ solve the likelihood equations of the model in which all parameters in Λ, B, Ω except for $B_{v\times \mathrm{pa}(v)}$ and Ω_{vv} are held fixed. It follows from Drton and Richardson (2004b, §§5.1-2) that $\hat{B}_{v\times \mathrm{pa}(v)}$ and $\hat{\Omega}_{vv}$ are equal to the empirical expressions on the right hand side of (3) and (4), respectively. Applying Lemma 21 yields the claim. ■

Remark 24 *Iterative Conditional Fitting* is a special purpose algorithm for maximum likelihood estimation in covariance graph models (Drton and Richardson, 2003; Chaudhuri et al., 2007). However, it does not exploit the results of Theorems 22 and 23. On the other hand, if one runs the ancestral graph extension of iterative conditional fitting described in Drton and Richardson (2004b) on a minimally oriented graph, then unnecessary computations are avoided by implicitly exploiting Theorems 22 and 23. This is illustrated in the example in Section 5.3.

If a bi-directed graph G has a minimally oriented graph $G^{\min}$ without bi-directed edges then G is Markov equivalent to a DAG (Theorem 19) and the likelihood equations have a unique solution that is a rational function of the empirical covariance matrix S. However, this is no longer true if there is a bi-directed edge in $G^{\min}$. In this case, G contains one of the two graphs in Figure 4 as a subgraph; compare Lemma 15(iii). Solving the likelihood equations for the bi-directed four-chain in Figure 4(i) is equivalent to computing the roots of a quintic polynomial. There exist data for which this quintic has exactly three real roots (Drton and Richardson, 2004a). Galois theory (Stewart, 1989, Lemma 14.7) implies that for these data the quintic is unsolvable by radicals, that is, the roots of the quintic and thus the solutions to the likelihood equations cannot be computed from the data in finitely many steps involving addition, subtraction, multiplication, division, or taking r-th roots. (Geiger et al., 2006, obtain similar results in the context of undirected graphs). Similarly, solving the likelihood equations of the bi-directed four-cycle in Figure 4(ii) corresponds to solving

	GAL7	GAL10	GAL1	GAL3	GAL2	GAL80	GAL11	GAL4
GAL7	*1.000*	0.91	0.88	0.50	0.81	0.21	-0.07	-0.08
GAL10	0.910	*1.000*	0.92	0.46	0.87	0.26	-0.08	-0.07
GAL1	0.880	0.920	*1.000*	0.39	0.87	0.28	-0.10	-0.10
GAL3	0.489	0.447	0.374	*0.998*	0.44	0.20	-0.18	0.12
GAL2	0.807	0.865	0.865	0.422	*0.991*	0.26	-0.18	-0.03
GAL80	0.224	0.271	0.297	0.191	0.280	*1.001*	0.08	0.23
GAL11	0	0	0	-0.208	-0.103	0	*1.022*	0.24
GAL4	0	0	0	0	0.038	0.209	0.255	*0.987*

Table 1: Gene expression data. Empirical correlation matrix (above diagonal) and maximum likelihood estimate (below diagonal). The italicized diagonal entries are ratios between maximum likelihood and empirical variance estimates.

a polynomial equation system of degree 17. This can be verified in computer algebra systems such as `Singular` (Greuel et al., 2005); see also Drton and Sullivant (2007, §5). It is natural to conjecture that there exist data for which this system is also unsolvable by radicals.

5.3 Example: Gene Expression Measurements

The application of covariance graph models to gene expression data has been promoted in Butte et al. (2000). For illustration, we select data from microarray experiments with yeast strands (Gasch et al., 2000). We focus on eight genes involved in galactose utilization. Expression measurements for all eight genes are available in $n = 134$ experiments, for which the empirical correlation matrix is shown in the upper-diagonal part of Table 1.

For these data, the covariance graph model induced by the graph G in Figure 5(i) has a deviance of 8.87 over 8 degrees of freedom, which indicates a good model fit; the p-value computed using a chi-square distribution is 0.35. Figure 5(ii) shows the unique minimally oriented graph $G^{\min}$. The maximum likelihood estimate obtained by fitting the model to the correlation matrix is shown in the lower-diagonal part of Table 1; note that this estimate is not a correlation matrix (not all the italicized diagonal entries are equal to one). As predicted by Theorem 22, the submatrix over GAL1, GAL7, and GAL10 equals the respective submatrix in the empirical correlation matrix. The regression coefficients for the regression of GAL2 on all remaining variables are identical when computed from the maximum likelihood versus the empirical estimate (Theorem 23).

The use of a minimally oriented graph $G^{\min}$ leads to a considerable gain in computational efficiency in the iterative calculation of the maximum likelihood estimate $\hat{\Sigma}$. With the identity matrix as starting value, iterative conditional fitting (Remark 24) on the original bi-directed graph G performs eight multiple regressions per iteration and converges after 103 iterations. Using the same starting value and termination criterion, iterative conditional fitting on $G^{\min}$ converges after only 5 iterations and requires only five multiple regressions per iteration (for the genes GAL2, GAL3, GAL4, GAL11, and GAL80), of which the one for GAL2 has to be executed only in the first iteration.

As in any application of covariance graph models, one might question the assumption of Gaussianity. Indeed there are 10 experiments in which the measurements for the genes GAL1, GAL7, GAL10 and GAL80 come out to be large negative values, and one in which GAL7 alone takes such

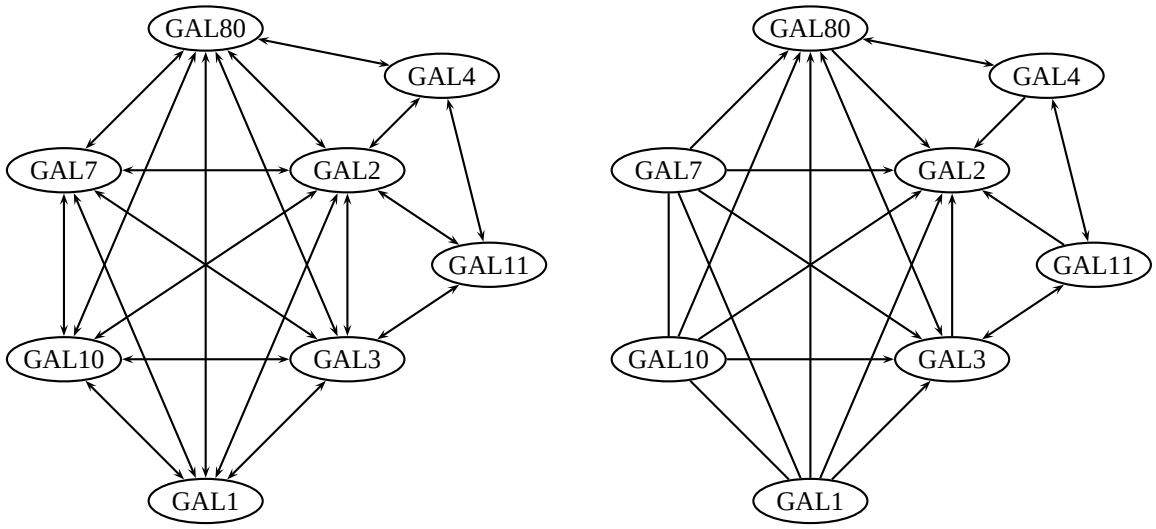

Figure 5: (i) Bi-directed graph G for gene expression measurements, (ii) the unique minimally oriented graph $G^{\min}$.

a value. These appear to be outliers (standardized values between -3 and -5), possibly produced by thresholding, as some values are identical. However, the measurements for the other genes are well within the range of the observations for the remaining 123 experiments. Thus it is unclear whether removing these 11 experiments from consideration is appropriate. If the 11 experiments are removed, then the correlations among GAL1, GAL7 and GAL10 decrease to values between 0.38 and 0.60, the latter value is the maximum of all correlations. Nevertheless the deviance for G only changes slightly to 10.09 (p-value 0.26). The iterative conditional fitting algorithm based on G now converges after only 20 iterations rather than 103. However, this is still four times as many iterations as required in iterative conditional fitting based on the minimally oriented graph $G^{\min}$; recall that in addition each iteration is also simpler.

The original correlation matrix in Table 1 exhibits an apparent similarity of the rows for GAL1, GAL7 and GAL10; this is also reflected in the graph G in which these variables form a complete set and have the same spouses. Such symmetry could be investigated further via a group symmetry model (Andersson and Madsen, 1998).

6. Conclusion

We showed how to remove a maximal number of arrowheads from the edges of a bi-directed graph G such that one obtains a maximal ancestral graph $G^{\min}$ that is Markov equivalent to G. The graph $G^{\min}$, called a minimally oriented graph, reveals whether G is Markov equivalent to an undirected graph, and also whether G is Markov equivalent to a DAG.

For the (Gaussian) covariance graph model associated with G, a minimally oriented graph $G^{\min}$ yields an alternative parametrization that provides insight into likelihood inference. The structure of the arrowheads in $G^{\min}$ allowed us to identify parts of the covariance matrix for which the maximum

likelihood estimates are equal to their empirical counterparts (this applies to all solutions to the likelihood equations if, as occasionally happens, there is more than one solution). This makes it possible to avoid or speed up iterative estimation of the full covariance matrix. We also saw that the maximum likelihood estimator of the covariance matrix in a covariance graph model is a rational function of empirical covariance matrix if $G^{\min}$ contains no bi-directed edge. This is similar to the results that identify decomposable models as the sub-class of all log-linear and all covariance selection models (Dempster, 1972) for which the maximum likelihood estimator is available in closed form.

Drton and Richardson (2008) formulate binary models based on the Markov property of bi-directed graphs. For these models, the maximum likelihood estimator is available in closed form if the model-inducing graph is Markov equivalent to a DAG. Moreover, we verified that in the example of the graph G in Figure 1, the maximum likelihood estimates of the marginal distributions of X_1 and X_4 are equal to the corresponding empirical proportions. We thus believe that analogs to the Gaussian results established here will hold in discrete models, but a general parametrization of discrete ancestral graph models is required to fully access the potential of the results obtained in this paper.

Acknowledgments

This paper is based upon work supported by the U.S. National Science Foundation (DMS-0505612 and 0505865) and the U.S. National Institutes for Health (R01-HG2362-3). We would also like to thank two anonymous referees for very helpful comments on the presentation of our work.

Appendix A. Connecting Paths and Boundary Containment

In this appendix we prove results about graphs that satisfy the *boundary containment property* from Definition 4. These results are used in the proof of Theorem 5.

Let v and w be two fixed distinct vertices that are m-connected given $C \subseteq V \setminus \{v,w\}$ in a simple mixed graph G. Define $\Pi_G(v,w|C)$ to be the set of paths that m-connect v and w given C in G, and let $\Pi_G^{\min}(v,w|C)$ be the set of paths that are of minimal length among the paths in $\Pi_G(v,w|C)$.

Lemma 25 *If a simple mixed graph G satisfies the boundary containment property, v_{i-1}, v_i and v_{i+1} are three consecutive vertices on a path π in G, and v_i is a non-collider on π, then v_{i-1} and v_{i+1} are adjacent.*

Proof If v_i is a non-collider, then the edge between v_i and v_{i-1} or the edge between v_i and v_{i+1} must have a tail at v_i. Suppose, without loss of generality, that the latter is the case. Then $\mathrm{Bd}(v_i) \subseteq \mathrm{Bd}(v_{i+1})$ and thus $v_{i-1} \in \mathrm{Bd}(v_{i+1})$, which is the claim. ■

Lemma 26 *Let G be a simple mixed graph, and $\pi = (v, v_1, \ldots, v_k, w) \in \Pi_G(v,w|C)$. Let $v_0 = v$ and $v_{k+1} = w$. If v_i is a non-endpoint vertex on π and there is an arrowhead at v_i on the edge between v_{i-1} and v_i, then either (i) $v_i \in \mathrm{An}(C)$ or (ii) the path $(v_i, v_{i+1}, \ldots, v_k, w)$ is a directed path from v_i to w.*

Proof Suppose the result is false. Let v_j be the vertex closest to w satisfying the antecedent of the Lemma, but not the conclusion. If v_j is a collider, then by definition of m-connection, $v_j \in \mathrm{An}(C)$, which is a contradiction. If v_j is a non-collider then $v_j \to v_{j+1}$ on π. If $v_{j+1} = w$, if $v_{j+1} \in \mathrm{An}(C)$, or if $(v_{j+1}, \ldots, v_k, w)$ is a directed path from v_{j+1} to w, then clearly v_j satisfies the conclusion of the Lemma, which is a contradiction. But if $v_{j+1} \notin \mathrm{An}(C)$ and $(v_{j+1}, \ldots, v_k, w)$ is not a directed path from v_{j+1} to w then v_{j+1} satisfies the conditions on v_j, but is closer to w, again a contradiction. ■

Lemma 27 *If G is an ancestral graph that satisfies the boundary containment property and $\pi = (v, v_1, \ldots, v_k, w) \in \Pi_G^{\min}(v, w|C)$ then no non-consecutive vertices on π are adjacent.*

Proof Let $v_0 = v$ and $v_{k+1} = w$, and suppose for a contradiction that there are non-consecutive vertices on the path π which are adjacent. Let (v_p, v_q), $p < q$, be a pair of adjacent vertices which are furthest apart on the path, that is, (p, q) maximizes the distance $|r - s|$ among pairs of indices of adjacent vertices v_r and v_s on the path. Since π is of minimal length, either $v \neq v_p$ or $w \neq v_q$.

Suppose that $v \neq v_p$. By definition of (p, q), v_{p-1} is not adjacent to v_q. Consequently, by Lemma 25, v_p is a collider on (v_{p-1}, v_p, v_q), and thus the edge between v_{p-1} and v_p has an arrowhead at v_p. It then follows by Lemma 26 that either $v_p \in \mathrm{An}(C)$ or $(v_p, v_{p+1}, \ldots, v_k, w)$ is a directed path from v_p to w. In the latter case $v_p \in \mathrm{An}(v_q)$, but there is an arrowhead at v_p on the edge between v_p and v_q, which contradicts that G is ancestral. Hence $v_p \in \mathrm{An}(C)$. If $v_q = w$ then the path $(v, v_1, \ldots, v_p, v_q = w)$ is m-connecting given C and shorter than π. Hence $v_q \neq w$. It then follows by the same argument that v_q is a collider on (v_p, v_q, v_{q+1}) and in $\mathrm{An}(C)$. However, this also leads to a contradiction since then the path $(v, v_1, \ldots, v_p, v_q, v_{q+1}, \ldots, v_k, w)$ is both m-connecting given C and shorter than π.

The case where $w \neq v_q$ may be argued symmetrically. ■

Corollary 28 *If G is an ancestral graph that satisfies the boundary containment property and $\pi = (v = v_0, v_1, \ldots, v_k, v_{k+1} = w) \in \Pi_G^{\min}(v, w|C)$, then all the non-endpoint vertices $v_1, \ldots, v_k$ are colliders on π.*

Proof This follows directly from Lemma 27 and Lemma 25. ■

Even though all non-endpoints on a path of the type described in Corollary 28 in $\Pi_G^{\min}(v, w|C)$ are colliders, not all non-endpoints must be in the set C. For example, in the graph $G_2^{\min}$ from Figure 3, the path (x, v, y) m-connects x and y given $\{w\}$ since the collider v is an ancestor of w. However, as the next Lemma shows, there will always exist a path in $\Pi_G^{\min}(v, w|C)$ such that all non-endpoints are colliders in C. In $G_2^{\min}$ from Figure 3, the path (x, w, y) m-connects x and y given $\{w\}$.

Lemma 29 *If G is an ancestral graph that satisfies the boundary containment property, and $\pi = (v = v_0, v_1, \ldots, v_k, v_{k+1} = w) \in \Pi_G^{\min}(v, w|C)$ is such that no other path in $\Pi_G^{\min}(v, w|C)$ has more non-endpoint vertices in C than π, then all non-endpoint vertices $v_1, \ldots, v_k$ on π are colliders that are in C.*

Proof By Corollary 28, all non-endpoints $v_1, \ldots, v_k$ are colliders. Assume that there exists $v_i \notin C$, $1 \leq i \leq k$. Since $\pi \in \Pi_G(v, w|C)$, and thus $v_i \in \mathrm{An}(C)$, there exists $c \in C$ such that $v_i \to \cdots \to c \in G$.

In particular, $c \neq v_{i-1}$ and $c \neq v_{i+1}$ because v_i is ancestral neither to v_{i-1} nor to v_{i+1}. The boundary containment property and the fact that G does not contain directed cycles imply that $v_i \to c \in G$. By Lemma 25, G contains edges between c and both v_{i-1} and v_{i+1}. Since the edge between v_{i-1} and v_i has an arrowhead at v_i and $v_i \to c \in G$, the edge between v_{i-1} and c must have an arrowhead at c because otherwise the fact that G is an ancestral graph would be contradicted. Similarly, the edge between v_{i+1} and c must have an arrowhead at c. If $v_{i-1} \to c$, then $v_i \neq v$, v_{i-2} is adjacent to c and by the same argument as above there must be an arrowhead at c on the edge between v_{i-2} and c. Repeating this argument yields that there exists a vertex v_ℓ, $\ell \leq i-1$, such that either $v_\ell \leftrightarrow c \in G$, or $v_\ell = v$ and $v \to c$. The same arguments also imply that there exists a vertex v_j, $j \geq i+1$, such that either $v_j \leftrightarrow c \in G$, or $v_j = w$ and $w \to c$. Therefore, the path $(v, v_1, \ldots, v_\ell, c, v_j, \ldots, v_k, w)$ is in $\Pi_G(v, w|C)$ and is either shorter than π or of equal length but with more non-endpoint vertices in C. This contradicts the choice of π and therefore the assumption of a non-endpoint on π that is not in C must be false. ■

References

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

T. W. Anderson and I. Olkin. Maximum-likelihood estimation of the parameters of a multivariate normal distribution. *Linear Algebra Appl.*, 70:147–171, 1985.

S. A. Andersson and J. Madsen. Symmetry and lattice conditional independence in a multivariate normal distribution. *Ann. Statist.*, 26:525–572, 1998.

A. Brandstädt, V. B. Le, and J. P. Spinrad. *Graph Classes: A Survey*. Society for Industrial and Applied Mathematics (SIAM), Philadelphia, PA, 1999.

A. J. Butte, P. Tamayo, D. Slonim, T. R. Golub, and I. S. Kohane. Discovering functional relationships between RNA expression and chemotherapeutic susceptibility using relevance networks. *Proc. Nat. Acad. Sci. USA*, 97:12182–12186, 2000.

S. Chaudhuri, M. Drton, and T. S. Richardson. Estimation of a covariance matrix with zeros. *Biometrika*, 94:199–216, 2007.

D. G. Corneil, Y. Perl, and L. K. Stewart. A linear recognition algorithm for cographs. *SIAM J. Comput.*, 14:926–934, 1985.

D. R. Cox and N. Wermuth. *Multivariate Dependencies: Models, Analysis and Interpretation*. Chapman and Hall, London, 1996.

D. R. Cox and N. Wermuth. Linear dependencies represented by chain graphs (with discussion). *Statist. Sci.*, 8:204–218,247–277, 1993.

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

M. Drton and T. S. Richardson. A new algorithm for maximum likelihood estimation in Gaussian graphical models for marginal independence. In U. Kjærulff and C. Meek, editors, *Proceedings of the 19th Conference on Uncertainty in Artificial Intelligence*, pages 184–191. Morgan Kaufmann, San Francisco, 2003.

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

M. Drton and T. S. Richardson. Iterative conditional fitting for Gaussian ancestral graph models. In M. Chickering and J. Halpern, editors, *Proceedings of the 20th Conference on Uncertainty in Artificial Intelligence*, pages 130–137. Morgan Kaufmann, San Francisco, 2004b.

M. Drton and T. S. Richardson. Binary models for marginal independence. *J. Roy. Statist. Soc. Ser. B*, 70:287–309, 2008.

M. Drton and S. Sullivant. Algebraic statistical models. *Statist. Sinica*, 17:1273–1297, 2007.

D. M. Edwards. *Introduction to Graphical Modelling*. Springer-Verlag, New York, second edition, 2000.

A. P. Gasch, P. T. Spellman, C. M. Kao, O. Carmel-Harel, M. B. Eisen, G. Storz, D. Botstein, and P. O. Brown. Genomic expression programs in the response of yeast cells to environmental changes. *Molecular Biology of the Cell*, 11:4241–4257, 2000.

D. Geiger, C. Meek, and B. Sturmfels. On the toric algebra of graphical models. *Ann. Statist.*, 34: 1463–1492, 2006.

G.-M. Greuel, G. Pfister, and H. Schönemann. SINGULAR 3.0. A Computer Algebra System for Polynomial Computations, Centre for Computer Algebra, University of Kaiserslautern, 2005. `http://www.singular.uni-kl.de`.

F. V. Jensen. *Bayesian Networks and Decision Graphs*. Springer-Verlag, New York, 2001.

S. T. Jensen. Covariance hypotheses which are linear in both the covariance and the inverse covariance. *Ann. Statist.*, 16:302–322, 1988.

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

J. T. A. Koster. On the validity of the Markov interpretation of path diagrams of Gaussian structural equation systems with correlated errors. *Scand. J. Statist.*, 26:413–431, 1999.

S. L. Lauritzen. *Graphical Models*. Clarendon Press, Oxford, UK, 1996.

G. Letac and H. Massam. Wishart distributions for decomposable graphs. *Ann. Statist.*, 35:1278–1323, 2007.

D. Madigan and K. Mosurski. An extension of the results of Asmussen and Edwards on collapsibility in contingency tables. *Biometrika*, 77:315–319, 1990.

Y. Mao, F. R. Kschischang, and B. J. Frey. Convolutional factor graphs as probabilistic models. In U. Kjærulff and C. Meek, editors, *Procceding of the 20th Conference on Uncertainty in Artificial Intelligence*, pages 374–381. Morgan Kaufmann, San Francisco, 2004.

F. Matúš. Stochastic independence, algebraic independence and abstract connectedness. *Theoretical Computer Science*, 134:455–471, 1994.

J. Pearl. *Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference*. Morgan Kaufmann, San Mateo, CA, 1988.

J. Pearl and N. Wermuth. 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*, pages 205–214. Springer, New York, 1994.

T. S. Richardson. Markov properties for acyclic directed mixed graphs. *Scand. J. Statist.*, 30: 145–157, 2003.

T. S. Richardson and P. Spirtes. Ancestral graph Markov models. *Ann. Statist.*, 30:962–1030, 2002.

T. S. Richardson and P. Spirtes. Causal inference via ancestral graph models (with discussion). In P. Green, N. Hjort, and S. Richardson, editors, *Highly Structured Stochastic Systems*, chapter 3, pages 83–105. Oxford University Press, Oxford, UK, 2003.

A. Roverato. A unified approach to the characterization of equivalence classes of DAGs, chain graphs with no flags and chain graphs. *Scand. J. Statist.*, 32:295–312, 2005.

I. Stewart. *Galois Theory*. Chapman and Hall, London, second edition, 1989.

N. Wermuth, D. R. Cox, and G. Marchetti. Covariance chains. *Bernoulli*, 12:841–862, 2006.

S. Wright. The method of path coefficients. *Ann. Math. Statist.*, 5:161–215, 1934.