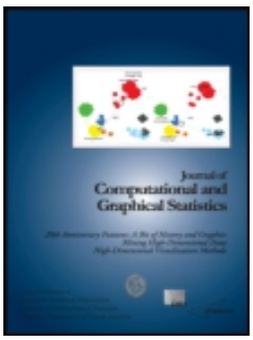

Taylor & Francis
Taylor & Francis Group

Journal of Computational and Graphical Statistics

ISSN: 1061-8600 (Print) 1537-2715 (Online) Journal homepage: http://www.tandfonline.com/loi/ucgs20

An Improved Iterative Proportional Scaling Procedure for Gaussian Graphical Models

Ping-Feng Xu, Jianhua Guo & Xuming He

To cite this article: Ping-Feng Xu, Jianhua Guo & Xuming He (2011) An Improved Iterative Proportional Scaling Procedure for Gaussian Graphical Models, Journal of Computational and Graphical Statistics, 20:2, 417-431, DOI: 10.1198/jcgs.2010.09044

To link to this article: https://doi.org/10.1198/jcgs.2010.09044

Supplementary materials for this article are available online.
Please click the JCGS link at http://pubs.amstat.org.

An Improved Iterative Proportional Scaling Procedure for Gaussian Graphical Models

Ping-Feng XU, Jianhua GUO, and Xuming HE

The maximum likelihood estimation of Gaussian graphical models is often carried out by the iterative proportional scaling (IPS) procedure. In this article, we propose an improvement to the IPS procedure by using local computation and by sharing computations on a junction tree T. The proposed procedure, called IIPS for short, adjusts node by node the marginals of the cliques of the underlying graph contained in the nodes of T, and sends messages between two adjacent nodes of T by an exchange operation for the propositional scaling step. We show, through complexity calculations and empirical examples, that the proposed IIPS procedure works more efficiently than the conventional IPS procedure for large Gaussian graphical models. Computer codes used in this article are available as an online supplement.

Key Words: Junction tree; Probability propagation algorithm.

1. INTRODUCTION

The Gaussian graphical models introduced by Dempster (1972) play an important role in statistical inference (see Whittaker 1990; Lauritzen 1996; Anderson 2003). In recent years, they have been used in high-dimensional problems with thousands of variables; see, for example, the work of Li and Gui (2006), and Ma, Gong, and Bohnert (2007). In those problems, the algorithmic efficiency for computing the maximum likelihood estimate (MLE) of the covariance matrix of the Gaussian graphical model becomes important. In this article, we propose an improvement to the iterative proportional scaling (IPS) procedure (see Speed and Kiiveri 1986), which is commonly used for computing the MLE.

The IPS procedure first introduced into the statistics literature by Deming and Stephan (1940) has been generalized and developed by many authors. From the theoretical side, Kullback (1968), Fienberg (1970), Darroch and Ratcliff (1972), Csiszár (1975, 1989),

Ping-Feng Xu is Ph.D. Student and Jianhua Guo is Professor (E-mail: *jhguo@nenu.edu.cn*), Key Laboratory for Applied Statistics of MOE and School of Mathematics and Statistics, Northeast Normal University, Changchun 130024, Jilin Province, China. Xuming He is Professor, Department of Statistics, University of Illinois at Urbana–Champaign, 725 South Wright Street, Champaign, IL 61820.

Journal of Computational and Graphical Statistics, Volume 20, Number 2, Pages 417–431
DOI: 10.1198/jcgs.2010.09044

Rüschendorf (1995), and Gao et al. (2010) provided information-theoretic interpretations of the IPS procedure together with its convergence property. Fienberg (1970), Darroch and Ratcliff (1972), and Speed and Kiiveri (1986) applied the IPS procedure to solve likelihood equations for contingency tables, log-linear models, and Gaussian graphical models, respectively. Jirousek and Preucil (1995) proposed a space saving modification of the IPS procedure for graphical models with contingency tables. Badsberg and Malvestuto (2001) implemented Jirousek and Preucil's method on Markov propagation trees, that is, junction trees.

Earlier improvements to the IPS procedure used the probability propagation algorithm, which makes exact inference efficiently on junction trees. In the conventional IPS procedure for Gaussian graphical models, Whittaker (1990) updated the covariance matrix by "adjusting a c-marginal," while Lauritzen (1996) used a similar operation to update the concentration matrix, that is, the inverse of covariance matrix. From the complexity viewpoint, both methods include matrix inverse operations. If the model involves a large number of variables, it becomes difficult to update the concentration matrices directly by the conventional IPS procedure. Hara and Takemura (2010) proposed a localized method for improving the computational efficiency of the conventional IPS procedure by taking advantage of the structure of graphical models.

In this article, we show that the operation of "adjusting a c-marginal" can operate locally on sub-covariance matrices by utilizing the conditional independence in the model. We propose an improved IPS procedure, which consists of a compilation stage and an iterative proportional scaling and propagation stage. In the first stage we find a triangulation of the underlying graph and construct a junction tree T of the triangulation as the computation structure, and then the junction tree is loaded with potentials. In the second stage we adjust node by node the marginals contained in the nodes of T, and send messages between two adjacent nodes of T by an exchange operation for the next propositional scaling step. A significant difference between our IIPS procedure and the method of Hara and Takemura (2010), to be called the HT method hereafter, is that IIPS shares computations in adjusting different c-marginals. This property of sharing computations reduces greatly the complexity of the conventional IPS procedure. Our studies show that the proposed IIPS procedure works more efficiently than the conventional IPS procedure, and routinely outperforms the HT method, for large Gaussian graphical models with hundreds or more variables.

The rest of this article is organized as follows. In Section 2, we provide some preliminaries about undirected graphs, Gaussian graphical models, the conventional IPS procedure, and potentials representing conditional density functions. Those readers who are familiar with the literature in this area may skip this section. Section 3 presents a theoretical result concerning the implementation of the IPS procedure by local computation on a junction tree. This is the basis of the proposed IIPS procedure. The detailed IIPS procedure and some complexity calculations are also given in this section. In Section 4, we compare the IIPS procedure with the conventional IPS procedure through numerical experiments. We make some concluding comments in Section 5, and defer the proof for our main results to the Appendix.

2. PRELIMINARIES

2.1 UNDIRECTED GRAPH

In this article we consider any graph $G = (V, E)$ that is simple and undirected, where V and E are a finite set of vertices and a set of edges, respectively. We say that two vertices $\alpha, \beta \in V$ are adjacent and write $\alpha \sim \beta$, if there is an edge between them, that is, if $(\alpha, \beta) \in E$. We call a subset of V complete if every pair of vertices are adjacent. A complete subset that is maximal with respect to inclusion is called a clique. The set of cliques of a graph G is denoted by $\mathcal{K}(G)$.

A graph G is called a triangulated graph if it contains no cycles of length ≥ 4 without a chord. We call $G^t = (V, E \cup F)$ a triangulation of $G = (V, E)$ if G^t is a triangulated graph, where F is the set of new edges filled into G^t. A graph G is a triangulated graph if and only if there exists a D-ordered sequence $K_1, K_2, \ldots, K_M$ of all cliques of G, that is, there is a $p(m) < m$ such that $S_m = K_m \cap (\bigcup_{i=1}^{m-1} K_i) \subseteq K_{p(m)}$, for all $m = 2, \ldots, M$. We call $K_{p(m)}$ a possible parent node of K_m and K_m a child node of $K_{p(m)}$. According to a D-ordered clique sequence $K_1, K_2, \ldots, K_M$ of G, a junction tree $T = (\mathcal{K}(G), E_T)$ can be constructed by adding an edge from a node to its parent node. Meanwhile, K_1 is called the root of T. For each edge $(K_i, K_j) \in E_T$, it is attached with a separator $S = K_i \cap K_j$ that connects nodes K_i and K_j in T. For more details about junction trees, we refer to the works of Lauritzen (1996), Cowell et al. (1999), and Wang and Guo (2008).

2.2 GAUSSIAN GRAPHICAL MODEL

The Gaussian graphical model $N(G)$ associated with an undirected graph $G = (V, E)$ is defined as the family of all normal distributions $N(\mu, \Sigma)$ that obey the pairwise Markov property, that is, for any random variable vector $Y \sim N(\mu, \Sigma)$ in this family, and $u, v \in V$,

$$\text{if } u \not\sim v, \text{ then } Y_u \perp\!\!\!\perp Y_v | Y_{V \setminus \{u,v\}}, \text{ that is, } \omega_{uv} = 0,$$

where ω_{uv} is an element of the concentration matrix $\Omega = \Sigma^{-1}$.

Now, we consider the MLE of Gaussian graphical model $N(G)$. Let $y_1, y_2, \ldots, y_n$ be an iid sample from $N(G)$, and

$$\overline{y} = \frac{1}{n} \sum_{i=1}^{n} y_i, \qquad S = \frac{1}{n} \sum_{i=1}^{n} (y_i - \overline{y})(y_i - \overline{y})^T.$$

The likelihood function is expressed as

$$L(\mu, \Sigma) \propto (\det \Sigma)^{-n/2} \exp\left\{ -\frac{n}{2} \operatorname{tr}(\Sigma^{-1} S) - \frac{n}{2} \operatorname{tr}(\Sigma^{-1} (\overline{y} - \mu)(\overline{y} - \mu)^T) \right\},$$

and the MLE of μ is $\overline{y}$. However, the MLE of the covariance matrix Σ and the MLE of the concentration matrix Ω satisfy the following likelihood equations:

$$\Sigma_{cc} = (\Omega^{-1})_{cc} = S_{cc} \quad \text{for all } c \in \mathcal{K}(G).$$

Note that we use lowercase c or c_i to represent a clique in $\mathcal{K}(G)$ in this article, and Σ_{cc} is a submatrix of Σ corresponding to $c \in \mathcal{K}(G)$.

2.3 A REVIEW OF THE IPS PROCEDURE

The IPS procedure iteratively and successively adjusts the covariance matrix for clique marginals. There are two schemes for adjusting the covariance matrix Σ or the concentration matrix Ω described by Whittaker (1990) and Lauritzen (1996), respectively. The operation A_c of "adjusting a c-marginal" on Σ is defined by

$$A_c\Sigma = \begin{pmatrix} S_{cc} & S_{cc}\Sigma_{cc}^{-1}\Sigma_{ca} \\ \Sigma_{ac}\Sigma_{cc}^{-1}S_{cc} & \Sigma_{aa} - \Sigma_{ac}\Sigma_{cc}^{-1}\Sigma_{ca} + \Sigma_{ac}\Sigma_{cc}^{-1}S_{cc}\Sigma_{cc}^{-1}\Sigma_{ca} \end{pmatrix},$$

where $a = V \setminus c$. For any ordering $c_1, c_2, \ldots, c_k \in \mathcal{K}(G)$, we define recursively

$$\Sigma^{m+1} = \big(A_{c_1}A_{c_2}\cdots A_{c_k}\big)\Sigma^m$$

for $m = 0, 1, 2, \ldots$, where Σ^0 is a covariance matrix of a normal distribution in $N(G)$. Then the MLE of the covariance matrix is $\hat{\Sigma} = \lim_{m\to+\infty}\Sigma^m$ by the convergence of the IPS procedure proven by Speed and Kiiveri (1986).

Similarly, the operation $\tilde{A}_c$ of "adjusting a c-marginal"; on Ω is defined by

$$\tilde{A}_c\Omega = \begin{pmatrix} S_{cc}^{-1} + \Omega_{ca}\Omega_{aa}^{-1}\Omega_{ac} & \Omega_{ca} \\ \Omega_{ac} & \Omega_{aa} \end{pmatrix},$$

where $a = V \setminus c$. Exploiting the inverse of the partitioned matrix, we have $((\tilde{A}_c\Omega)^{-1})_{cc} = S_{cc}$, which implies that $\tilde{A}_c$ does indeed adjust the concentration matrix to satisfy the likelihood equations. For any ordering $c_1, c_2, \ldots, c_k \in \mathcal{K}(G)$, we define recursively

$$\Omega^{m+1} = \big(\tilde{A}_{c_1}\tilde{A}_{c_2}\cdots\tilde{A}_{c_k}\big)\Omega^m$$

for $m = 0, 1, 2, \ldots$, where Ω^0 is a concentration matrix of a normal distribution in $N(G)$. Then the MLE of the concentration matrix is $\hat{\Omega} = \lim_{m\to+\infty}\Omega^m$ by the convergence of the IPS procedure proven by Lauritzen (1996).

In fact, both A_c and $\tilde{A}_c$ scale proportionally in the sense that

$$f(y|\hat{\mu}, A_c\Sigma) = f(y|\hat{\mu}, \Sigma)\frac{f(y_c|\hat{\mu}_c, S_{cc})}{f(y_c|\hat{\mu}_c, \Sigma_{cc})}$$

and

$$f(y|\hat{\mu}, (\tilde{A}_c\Omega)^{-1}) = f(y|\hat{\mu}, \Omega^{-1})\frac{f(y_c|\hat{\mu}_c, S_{cc})}{f(y_c|\hat{\mu}_c, (\Omega^{-1})_{cc})},$$

where $f(y|\mu, \Sigma)$ is the normal density function (see Speed and Kiiveri 1986; Whittaker 1990; Lauritzen 1996).

Although the operations A_c and $\tilde{A}_c$ can adjust the marginals equipollently, they have different operational complexities. The most computationally intensive operations of A_c and $\tilde{A}_c$ are matrix inversion and matrix multiplication. Let $\mathrm{mul}(n)$, $\mathrm{div}(n)$, $\mathrm{add}(n)$ be the number of multiplications, the number of divisions, and the number of additions and subtractions involved in computing the inversion of an n-by-n positive definite matrix, respectively. They depend on the algorithm used, but even the most efficient matrix inverse algorithms require $\mathrm{mul}(n)$ to be in the order of $n^{2+\eta}$ for some $\eta > 0$. Table 1(a) and (b) gives the numbers of multiplications needed in A_c and $\tilde{A}_c$, respectively. We typically

Table 1. Parts (a) and (b) describe the number of multiplications of A_c and $\tilde{A}_c$, respectively.

Operations	Number of multiplications
(a) A_c	
Σ_{cc}^{-1}	$\text{mul}(\lvert c\rvert)$
$\Sigma_{ac}\Sigma_{cc}^{-1}$	$\lvert a\rvert\lvert c\rvert^2$
$(\Sigma_{ac}\Sigma_{cc}^{-1})\Sigma_{ca}$	$\lvert a\rvert^2\lvert c\rvert$
$(\Sigma_{ac}\Sigma_{cc}^{-1})S_{cc}$	$\lvert a\rvert\lvert c\rvert^2$
$(\Sigma_{ac}\Sigma_{cc}^{-1}S_{cc})(\Sigma_{cc}^{-1}\Sigma_{ca})$	$\lvert a\rvert^2\lvert c\rvert$
Total	$\text{mul}(\lvert c\rvert) + 2\lvert a\rvert\lvert c\rvert^2 + 2\lvert a\rvert^2\lvert c\rvert$
(b) $\tilde{A}_c$	
S_{cc}^{-1}	$\text{mul}(\lvert c\rvert)$
Ω_{aa}^{-1}	$\text{mul}(\lvert a\rvert)$
$\Omega_{ca}\Omega_{aa}^{-1}$	$\lvert a\rvert^2\lvert c\rvert$
$(\Omega_{ca}\Omega_{aa}^{-1})\Omega_{ac}$	$\lvert a\rvert\lvert c\rvert^2$
Total	$\text{mul}(\lvert c\rvert) + \text{mul}(\lvert a\rvert) + \lvert a\rvert^2\lvert c\rvert + \lvert a\rvert\lvert c\rvert^2$

have $|c| \ll |a|^{\min\{\eta,0.5\}}$ for large sparse models, so the dominating term of the multiplication complexity is $\text{mul}(|c|)$ or $2|a|^2|c|$ for A_c, which is of lower order than that for $\tilde{A}_c$. The number of additions and subtractions is $\text{add}(|c|) + 2|a|^2$ in A_c, as compared with $\text{add}(|c|) + \text{add}(|a|) + |c|^2$ in $\tilde{A}_c$. These calculations favor the use of A_c, as suggested by Whittaker (1990).

Hara and Takemura (2010) were among the first to suggest a localized method for improving the computational efficiency of the conventional IPS procedure. They gave another expression of $\tilde{A}_c\Omega$ as

$$\tilde{A}_c\Omega = \begin{pmatrix} S_{cc}^{-1} + \Omega_{cc} - ((\Omega^{-1})_{cc})^{-1} & \Omega_{ca} \\ \Omega_{ac} & \Omega_{aa} \end{pmatrix},$$

and provided an algorithm to compute $((\Omega^{-1})_{cc})^{-1}$ by using the structure of a triangulation of the underlying graph. First, a triangulation G^t of G is obtained with M cliques. For $c \in \mathcal{K}(G)$, find a D-ordered clique sequence $K_1, K_2, \ldots, K_M$ of G^t such that $c \subseteq K_1$. Let $R_1 = K_1 \setminus c$, and $R_m = K_m \setminus (\bigcup_{i=1}^{m-1} K_i)$, for $m = 2, \ldots, M$. Then, for adjusting c-marginal, the HT method requires $\sum_{m=1}^{M}\{|R_m||K_m|^2 - |R_m||K_m| - |R_m|^2|K_m| + \frac{|R_m|(|R_m|+1)(2|R_m|+1)}{6}\}$ multiplications, $\sum_{m=1}^{M}\{|R_m||K_m| - \frac{|R_m|(|R_m|+1)}{2}\}$ divisions, $\sum_{m=1}^{M}\{|R_m||K_m|^2 - |R_m||K_m| - |R_m|^2|K_m| + \frac{|R_m|(|R_m|+1)(2|R_m|+1)}{6}\} + |c|^2$ subtractions, and $|c|^2$ additions. The computational complexity of the HT method improves on that of $\tilde{A}_c$ for large sparse models.

2.4 PROPAGATION ALGORITHM

In this article, we use a propagation scheme similar to that described by Lauritzen and Jensen (2001) to send message between the nodes of a junction tree. First, we describe potentials and their operations, and then we load a rooted junction tree with potentials.

The IPS procedure adjusts the covariance matrix, but not the mean. Therefore, we consider potentials $\phi = [\Psi, \Sigma](X|Y)$ that specify the distribution of X given $Y = y$, where Ψ is the matrix of the regression coefficients, and Σ is the conditional covariance matrix of X given $Y = y$. Also, $\phi = [-, \Sigma](X|-)$ specifies the marginal distribution of X.

For a potential $\phi = [-, \Sigma](X \cup Y|-)$ where $X \cap Y = \emptyset$, we define the marginal potential of Y by

$$\phi^{\downarrow Y} = [-, \Sigma_{YY}](Y|-).$$

Obviously, this operation respects the marginalization of a distribution. For two given potentials $\phi = [-, \Lambda](Z|-)$, and $\psi = [\Psi, \Sigma](X|Y)$ satisfying $Y \subseteq Z$, $X \cap Z = \emptyset$, and $X \perp\!\!\!\perp (Z \setminus Y)|Y$, the combination of ϕ and ψ is defined by

$$\phi \otimes \psi = [-, \Delta](X \cup Z|-),$$

where

$$\Delta = \begin{pmatrix} \Sigma + \Psi \Lambda_{YY} \Psi^T & \Psi \Lambda_{YZ} \\ \Lambda_{ZY} \Psi^T & \Lambda \end{pmatrix}.$$

This combination operation corresponds to the ordinary composition of marginal and conditional distributions. This combination operation needs $|X|^2$ additions, and $|X||Y|^2 + |X|^2|Y| + |X||Y||Z|$ multiplications, where $|X|$ denotes the number of variables in the set X.

For a potential $\phi = [-, \Sigma](X \cup Y|-)$, where $X \cap Y = \emptyset$, its complement $\phi^{|Y}$ is defined by

$$\phi^{|Y} = [\Sigma_{XY} \Sigma_{YY}^{-1}, \Sigma_{XX} - \Sigma_{XY} \Sigma_{YY}^{-1} \Sigma_{YX}](X|Y),$$

which requires $\text{mul}(|Y|) + |X||Y|^2 + |X|^2|Y|$ multiplications, $\text{div}(|Y|)$ divisions, and $\text{add}(|Y|) + |X|^2$ additions and subtractions. We also define $\phi^{\downarrow Y}$ through $\phi = \phi^{|Y} \otimes \phi^{\downarrow Y}$.

For two potentials $\phi = [-, \Sigma](X \cup Y|-)$ and $\psi = [\Psi, \Lambda](Z|Y)$ satisfying $X \cap Y = \emptyset$, $Y \cap Z = \emptyset$, and $X \cap Z = \emptyset$, we convert them to two potentials $\phi = \phi^{|Y} = [\Gamma, \Theta](X|Y)$ and $\psi = [-, \Delta](Z \cup Y|-)$, where

$$\Delta = \begin{pmatrix} \Lambda + \Psi \Sigma_{YY} \Psi^T & \Psi \Sigma_{YY} \\ \Sigma_{YY} \Psi^T & \Sigma_{YY} \end{pmatrix}.$$

The above process is called an exchange operation. When exchanging two potentials, we need $|Z||Y|^2 + |Y||Z|^2$ multiplications and $|Z|^2$ additions to compute $\psi = [-, \Delta](Z \cup Y|-)$. In fact, if X and Z are independent given Y, the exchange operation computes the distributions of $X|Y = y$ and $Z \cup Y$ by Bayes's theorem, given the distributions of $X \cup Y$ and $Z|Y = y$. Because the joint distribution of $X|Y = y$ and $Z \cup Y$ is the same as that of $X \cup Y$ and $Z|Y = y$, the exchange operation does not change the combination of potentials before and after the operation is performed.

Next, we consider loading a junction tree with potentials. Let $T = (\mathcal{K}(G^t), E_T)$ be a junction tree of some triangulation G^t with the corresponding D-ordered clique sequence

$K_1, K_2, \ldots, K_M$. For a normal distribution $N(\mu, \Sigma) \in N(G)$, we load T as follows:

$$\phi_{K_1} = \left[-, \Sigma_{K_1 K_1}\right](K_1|-),$$
$$\phi_{K_i} = \left[\Sigma_{R_i S_i} \Sigma_{S_i S_i}^{-1}, \Sigma_{R_i R_i \cdot S_i}\right](R_i|S_i),$$

where $\Sigma_{R_i R_i \cdot S_i} = \Sigma_{R_i R_i} - \Sigma_{R_i S_i} \Sigma_{S_i S_i}^{-1} \Sigma_{S_i R_i}$, $S_i = K_i \cap (\bigcup_{j=1}^{i-1} K_j)$ and $R_i = K_i \setminus (\bigcup_{j=1}^{i-1} K_j)$ for $i = 2, \ldots, M$. According to the D-ordered sequence, we can recursively combine all potentials $\phi_{K_1} \otimes \phi_{K_2} \otimes \cdots \otimes \phi_{K_M}$, defined as $\bigotimes_{K \in \mathcal{K}(G^t)} \phi_K$. Note that throughout this article we represent the node of a junction tree by capital letter K or K_i. Then we obtain the following propositions.

Proposition 1. *For a normal distribution $N(\mu, \Sigma) \in N(G)$ and a junction tree $T = (\mathcal{K}(G^t), E_T)$ of some triangulation G^t loaded as above, $\bigotimes_{K \in \mathcal{K}(G^t)} \phi_K$ is equal to the potential $[-, \Sigma](V|-)$ representing $N(\mu, \Sigma)$.*

Proposition 2. *For a normal distribution $N(\mu, \Sigma) \in N(G)$ and a junction tree $T = (\mathcal{K}(G^t), E_T)$ with a root K_1 of some triangulation G^t loaded as above, we exchange the potentials ϕ_{K_1} and ϕ_{K_j} where K_j is a child of K_1; then*

$$\phi_{K_j} = \left(\bigotimes_{K \in \mathcal{K}(G^t)} \phi_K \right)^{\downarrow Y_{K_j}} = \left[-, \Sigma_{K_j K_j}\right](K_j|-),$$

$$\phi_{K_1} = \left[\Sigma_{K_1 \setminus S_j S_j} \Sigma_{S_j S_j}^{-1}, \Sigma_{K_1 \setminus S_j K_1 \setminus S_j} - \Sigma_{K_1 \setminus S_j S_j} \Sigma_{S_j S_j}^{-1} \Sigma_{S_j K_1 \setminus S_j}\right](K_1 \setminus S_j | S_j),$$

and the combinations of all potentials are equal before and after we exchange ϕ_{K_1} and ϕ_{K_j}.

In fact, after we exchange potentials ϕ_{K_1} and ϕ_{K_j}, K_j can be viewed as a "new" root of the junction tree T with the corresponding D-ordered clique sequence $K_j, K_1, \ldots, K_{j-1}, K_{j+1}, \ldots, K_M$, and T is also loaded as above. We call the new root K_j loaded with marginal potential a temporary root. Note that every node of T can be as a temporary root by exchanging potentials of two adjacent nodes along the junction tree T. Proposition 2 helps understand computation sharing in adjusting different clique marginals in the IIPS procedure to be described in Section 3.

3. IMPROVED ITERATIVE PROPORTIONAL SCALING PROCEDURE

First, we show some theoretical results about local computation of A_c on junction trees, and discuss the "depth first adjust" (DFA) algorithm, that is, Algorithm 1, in preparation for our main proposal.

Theorem 1. *Let $T = (\mathcal{K}(G^t), E_T)$ be a junction tree with a root K_1 of some triangulation G^t, and $K_1, K_2, \ldots, K_M$ be the corresponding D-ordered clique sequence. For a normal distribution $N(\mu, \Sigma) \in N(G)$, T is loaded as in Section 2.4. After we update ϕ_{K_1} by $[-, A_c \Sigma_{K_1 K_1}](K_1|-)$ where c is a clique of G contained in the root K_1, then we get that*

Algorithm 1 DFA(K_i)

```
1: for each clique c ∈ 𝒦(G) contained in the node K_i do
2:     Σ_{K_i K_i} = A_c Σ_{K_i K_i};
3: set K_i to be visited;
4: if there exists an unvisited child node K_j of K_i then
5:     exchange potentials φ_{K_i} and φ_{K_j};
6:     DFA(K_j);
7: if K_i has a parent node K_p in T then
8:     exchange potentials φ_{K_i} and φ_{K_p};
```

(i) $[-, A_c\Sigma](V|-) = \bigotimes_{K\in\mathcal{K}(G^t)} \phi_K$ *and that*

(ii) *if we load* $T = (\mathcal{K}(G^t), E_T)$ *as follows*:

$$\phi'_{K_1} = \left[-, (A_c\Sigma)_{K_1K_1}\right](K_1|-),$$
$$\phi'_{K_i} = \left[(A_c\Sigma)_{R_iS_i}(A_c\Sigma)^{-1}_{S_iS_i}, (A_c\Sigma)_{R_iR_i\cdot S_i}\right](R_i|S_i),$$

where $(A_c\Sigma)_{R_iR_i\cdot S_i} = (A_c\Sigma)_{R_iR_i} - (A_c\Sigma)_{R_iS_i}(A_c\Sigma)^{-1}_{S_iS_i}(A_c\Sigma)_{S_iR_i}$, *for* $i = 2, \ldots, M$, *then we have that* $\phi_{K_i} = \phi'_{K_i}$ *for* $i = 1, \ldots, M$.

It follows from (i) of Theorem 1 that the operation A_c of "adjusting a c-marginal" can operate locally on sub-covariance; thus we can avoid the inverse operation on high-order matrices to decrease the complexity of the IPS procedure. Furthermore, (ii) shows that, after we update $\phi_{K_1} = [-, A_c\Sigma_{K_1K_1}](K_1|-)$, all potentials ϕ_K are equal to the potentials of T, which is loaded by distribution $N(\mu, A_c\Sigma)$ as in Section 2.4. Therefore, we can adjust another clique marginal without updating other potentials, and as a consequence, we propose sharing computations on potentials in adjusting different clique marginals. This property accelerates the IPS procedure.

After all the cliques $c_1, c_2, \ldots, c_m$ contained in the root K_1 are adjusted, we exchange potentials ϕ_{K_1} and ϕ_{K_j}, where K_j is a child node of K_1 in T. According to Proposition 2, T has a temporary root K_j, and it is loaded by distribution $N(\mu, (A_{c_1}A_{c_2}\cdots A_{c_m})\Sigma)$. Then by Theorem 1, we can adjust clique marginals contained in K_j. In fact, in preparing for adjusting clique marginals contained in K_j, we update just ϕ_{K_1} and ϕ_{K_j}, but not other potentials, which means that our strategy shares computations on potentials except ϕ_{K_1} and ϕ_{K_j}. After all the cliques contained in K_j are adjusted, we exchange ϕ_{K_j} and ϕ_{K_s}, where K_s is a child node of K_j in T, and adjust the clique marginals contained in K_s. If all the cliques contained in some K_i are adjusted, we say that K_i is visited; otherwise it is unvisited. We repeat the above process until the temporary root K_j has no children. At that time we backtrack, returning to the most recent node with an unvisited child node by exchanging the potentials. This way, we can adjust, node by node, in a depth first search (DFS) manner the marginals contained in nodes of T, and send messages between the parent node and child node by exchange operations.

We now introduce an improved IPS procedure in Algorithm 2. In this IIPS procedure, Lines 4, 5, 6 belong to the compilation stage and Lines 7, 8, 9, 10 to the iterative propor-

Algorithm 2 Improved IPS procedure

1: **Input**: A model $N(G)$ and an initial distribution $N(\overline{y}, \Sigma^0) \in N(G)$
2: **Output**: The MLE $\widehat{\Sigma}$
3: **begin**
4: Construct a triangulation G^t of G, and a junction tree $T = (\mathcal{K}(G^t), E_T)$ of G^t with a root K_1;
5: Find all cliques of G. For each clique c of G, assign c to a node K of T such that $c \subseteq K$;
6: Load T with potentials as in Section 2.4;
7: **while** *not convergent* **do**
8: set all nodes of T to be unvisited;
9: call DFA(K_1) to adjust clique marginals in each node of T and exchange potentials of the adjacent nodes of T;
10: Compute $[-, \widehat{\Sigma}](V|-) = \bigotimes_{K \in \mathcal{K}(G^t)} \phi_K$ by combination operation;
11: **end**

tional scaling and propagation stage. In Line 4, G may have several different triangulations and we wish to find an optimal triangulation that results in the least complexity. The optimal triangulation problem is NP-hard (Wen 1990), but there are several algorithms (Kjærulff 1992; Larranaga et al. 1997) for obtaining good triangulations. In this article, we use the MCS-M method proposed by Berry et al. (2004) to compute a minimal triangulation G^t. According to lemma 1 of Xu and Guo (2010), we know that any elimination ordering produced by MCS-M on G can be produced by MCS on G^t, so the expanded version of MCS proposed by Blair and Peyton (1993) can be used to construct a junction tree of G^t. To construct a junction tree in Line 4, we suggest using an optimal junction tree as in the work of Jensen and Jensen (1994). We will apply the Bron–Kerbosch algorithm introduced by Bron and Kerbosch (1973) to find all cliques in the IIPS procedure.

The validity of the IIPS procedure will be proven in the Appendix. Here, we analyze the complexity of the IIPS procedure. Because the MCS-M algorithm and the expanded MCS algorithm for constructing a triangulation and its junction tree are polynomially computable algorithms, the complexity of the IIPS procedure rests on adjusting all clique marginals and exchanging potentials of nodes. By the analysis of complexity of A_c in Section 2.3, we require $\sum_{K \in \mathcal{K}(G^t), c \subset K} \text{mul}(|c|) + 2|c||K \setminus c|^2 + 2|K \setminus c||c|^2$ multiplications, $\sum_{K \in \mathcal{K}(G^t), c \subset K} \text{div}(|c|)$ divisions, and $2\sum_{K \in \mathcal{K}(G^t), c \subset K} |K \setminus c|^2 + \text{add}(|c|)$ additions and subtractions to adjust clique marginals, where c is a clique of G assigned to K in Line 5 of Algorithm 2. By the analysis of complexity of the exchange operation and complement operation, we need $2\sum_{(K_1,K_2) \in E_T} (\text{mul}(s) + sr_1^2 + sr_2^2 + r_1 s^2 + r_2 s^2)$ multiplications, $2\sum_{(K_1,K_2) \in E_T} \text{div}(s)$ divisions, $2(\sum_{(K_1,K_2) \in E_T} r_1^2 + r_2^2 + \text{add}(s))$ additions and subtractions to exchange potentials along the junction tree T, where $s = |K_1 \cap K_2|$, $r_1 = |K_1 \setminus K_2|$, and $r_2 = |K_2 \setminus K_1|$. Although the IIPS procedure needs additional operations to exchange potentials of nodes of T, we adjust the clique marginals locally to avoid the inverse operation on large matrices, and we share computations in adjusting different clique marginals, so the IIPS procedure has a lower complexity than the conventional IPS

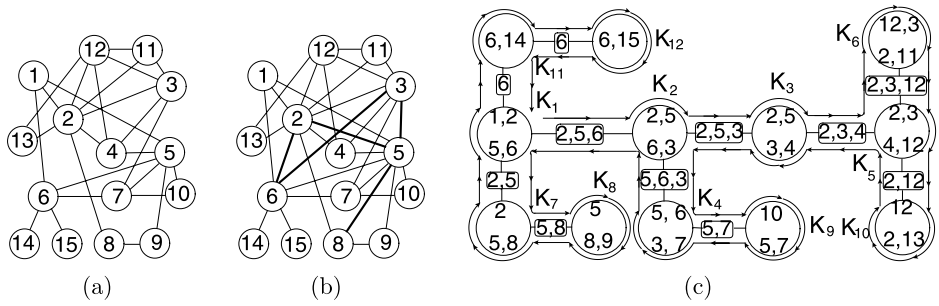

Figure 1. Part (a) presents the coherent subgraph for starch metabolism. Part (b) is a triangulation G^t of the coherent subgraph, and the thick lines are new edges filled into G^t. Part (c) is an optimal junction tree of G^t with root $K_1 = \{1, 2, 5, 6\}$. The ellipses and round rectangles respect the nodes and separators, respectively. The curve with arrows surrounding the junction tree respects the sequence of nodes to be adjusted in a depth first search manner by Algorithm 1.

procedure, especially when the underlying graph is sparse. Now, we present an example to illustrate the difference between the IIPS procedure, the HT method, and the conventional IPS procedure by the operation A_c.

Example 1: Shown in Figure 1(a) is the coherent subgraph for starch metabolism identified by Ma, Gong, and Bohnert (2007). The subgraph around starch catabolism includes 15 genes represented by vertices 1 to 15, respectively. We consider the Gaussian graphical model for a 15-dimensional random vector with the graph in Figure 1(a). The set of cliques of the graph is $\{\{2, 12, 3, 4\}, \{2, 12, 3, 11\}, \{2, 12, 13\}, \{2, 8\}, \{2, 1\}, \{5, 6, 1\}, \{5, 6, 7\}, \{5, 4\}, \{5, 9\}, \{5, 10, 7\}, \{6, 15\}, \{6, 14\}, \{7, 3\}, \{9, 8\}\}$. To be specific, we use the Gauss–Jordan elimination (cf. Stoer and Bulirsch 2002) for matrix inversion in this example. Then, we need 13,500 multiplications, 36 divisions, and 4476 additions and subtractions to adjust all clique marginals by A_c.

Consider the triangulation G^t of the coherent subgraph in Figure 1(b). If all cliques of G are adjusted once, the HT method needs 1022 multiplications, 402 divisions, 1222 subtractions and additions, respectively, making the HT method favorable to the IPS procedure.

For the IIPS procedure, we use the same triangulation, with the optimal junction tree T of the triangulation shown in Figure 1(c). Table 2 shows the corresponding D-ordered clique sequence of the junction tree. The IIPS procedure needs 282 multiplications, 18 divisions, 70 additions and subtractions, to adjust all clique marginals, and then 770 multiplications, 50 divisions, 214 additions and subtractions to exchange potentials. If we adjust all clique marginals in G once by calling the DFA algorithm, the total numbers of multiplications, divisions, additions and subtractions needed are 1052, 68, and 284, respectively, making the IIPS procedure preferable to the other methods. The advantage of the IIPS procedure over the HT method will become more evident as the number of variables and the number of cliques increase.

Next, we compare the IIPS procedure with the HT method through an example of an n-cycle.

Table 2. The D-ordered clique sequence of T and the cliques assigned to each node of T.

Nodes of T	Parent	Clique marginals
$K_1 = \{1, 2, 5, 6\}$	–	$\{1, 2\}, \{1, 5, 6\}$
$K_2 = \{2, 5, 6, 3\}$	K_1	–
$K_3 = \{2, 5, 3, 4\}$	K_2	$\{5, 4\}$
$K_4 = \{5, 6, 3, 7\}$	K_2	$\{5, 6, 7\}, \{3, 7\}$
$K_5 = \{2, 3, 4, 12\}$	K_3	$\{2, 3, 4, 12\}$
$K_6 = \{2, 3, 12, 11\}$	K_5	$\{2, 3, 12, 11\}$
$K_7 = \{2, 5, 8\}$	K_1	$\{2, 8\}$
$K_8 = \{5, 8, 9\}$	K_7	$\{5, 9\}, \{8, 9\}$
$K_9 = \{5, 7, 10\}$	K_4	$\{5, 7, 10\}$
$K_{10} = \{2, 12, 13\}$	K_5	$\{2, 12, 13\}$
$K_{11} = \{6, 14\}$	K_1	$\{6, 14\}$
$K_{12} = \{6, 15\}$	K_1	$\{6, 15\}$

Example 2: Let $G = (V, E)$ be an n-cycle, $n \geq 4$, with $V = \{\alpha_0 = \alpha_n, \alpha_1, \ldots, \alpha_{n-1}\}$, $E = \{(\alpha_{i-1}, \alpha_i) | i = 1, \ldots, n\}$. We use the minimal triangulation $G^t = (V, E^t)$, where $E^t = E \cup \{(\alpha_0, \alpha_i) | i = 2, \ldots, n-2\}$. There exist $n-2$ cliques with three vertices in G^t. For any D-ordered clique sequence $K_1, \ldots, K_{n-2}$, $S_i = K_i \cap (\bigcup_{j=1}^{i-1} K_j)$ contains two vertices for $i = 2, \ldots, n-2$. For adjusting all cliques once, the HT method needs $4n(n-2)$ multiplications, $2n(n-2)$ divisions, and $4n^2$ subtractions and additions. If we use the Gauss–Jordan elimination for matrix inversion in the IIPS procedure, we require $60(n-2)$ multiplications, $6(n-2)$ divisions, $12(n-2)$ additions and subtractions. The advantage of the IIPS procedure manifests itself with large values of n.

4. EXPERIMENTS

In this section we report some comparisons of computing times of the IIPS procedure versus the conventional IPS procedure on modestly large Gaussian graphical models. The IPS procedures using A_c and $\tilde{A}_c$ will be denoted by IPS1 and IPS2, respectively. Our codes were written in C++ and Matlab 2007, and the programs executed in a Windows XP-based PC of AMD Athlon(tm) 64×2 Dual core Processor 4400+ 2.30 GHz with 2.00 GB memory. The codes of constructing triangulations and junction trees were written in C++, and the codes of adjusting clique marginals and exchanging potentials of nodes of junction trees in Matlab 2007. The programs were run by calling C++ from Matlab.

In our experiment, we first generate a connected graph randomly, and then generate a symmetric positive-definite matrix Σ (as described below) such that $N(0, \Sigma)$ satisfies the Markov property. After samples of size 5000 are drawn from $N(0, \Sigma)$, we use IPS1, IPS2, and the proposed IIPS procedure to compute the MLE of the covariance matrix for each sample and compare the CPU times needed for each method. In this experiment, the HT method is slower to execute than IPS1 so the CPU time of the HT method is not included in the comparison.

For given probability of adding edges p and the number of vertices n, we construct a tree with n vertices randomly and then add new edge (u, v) into the tree with probability p

for each pair of vertices u, v. For a given graph $G = (V, E)$, we generate randomly a symmetric positive-definite matrix Ω as follows:

1. Set default $\Omega_{ij} = 0$.
2. Generate $\Omega_{ii} \sim \text{Uniform}(-1, 1)$.
3. For $(i, j) \in E$ and $1 \le i < j \le \#V$, generate $\Omega_{ij} \sim \text{Uniform}(-1, 1)$.
4. Let $\Omega = \Omega + \Omega^T$.
5. For $1 \le i \le \#V$, let $\Omega_{ii} = \sum_{j=1}^{\#V} |\Omega_{ij}|$, so that Ω is a diagonal-dominant matrix. If the minimal eigenvalue of Ω is not positive, that is, Ω is not positive-definite, go to Step 1 and start over.
6. $\Sigma = \Omega^{-1}$.

We examine the cases of $p = 0.1, 0.05, 0.01, 0.005$, $n = 50, 100, 200, 300$. For each case, we recorded the CPU times of IPS1, IPS2, and IIPS on 50 randomly generated graphical models.

Based on those 50 runs, we estimate the 95% confidence intervals for the ratio in CPU times of IPS1 (or IPS2) and IIPS and display them in Figure 2. These confidence intervals are normality-based and therefore meant to be approximate. From Figure 2, it is clear that the proposed IIPS procedure is computationally more efficient when $n > 100$. When we extended the case to $n = 2000$ and $p = 0.001$, we compared IPS1 and IIPS, and found the confidence interval to be (9.98, 10.43) for the ratio. Our experiment confirmed that the computational savings from the IIPS procedure can be substantial for large graphical models.

We also note that when n is small, it takes more time for the proposed IIPS procedure than IPS1 and IPS2. Recall Algorithm 2, where in Line 5 we need to judge whether c is a subset of K or not for each pair clique c in G and node K of T. Line 5 takes about half of the total computing time in the IIPS procedure, which explains why this becomes a drag for the IIPS in the speed when n is small.

5. CONCLUSION

In this article, we propose an improved IPS procedure in Gaussian graphical models by performing local computation and sharing computations on junction trees. We have shown that the IIPS procedure works more efficiently than the conventional IPS procedure for large Gaussian graphical models. We have also demonstrated the value of the proposed IIPS procedure by comparing its complexity with that of the localized method of Hara and Takemura (2010).

As pointed out by proposition 5.6 of the book by Lauritzen (1996), the MLE of a Gaussian graphical model can be computed from the MLE of marginal models on decomposed subgraphs. If we decompose the graph into its maximal prime subgraphs (cf. Malvestuto and Moscarini 2000 or Leimer 1993), then it suffices to compute the MLE

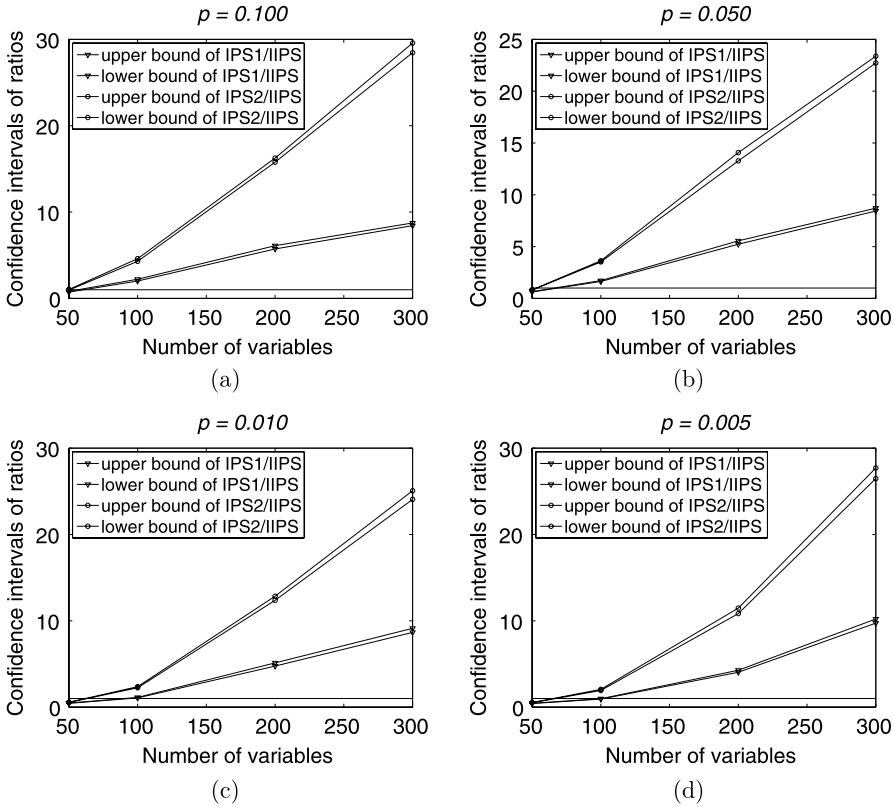

Figure 2. The curves are 95% confidence intervals for the ratio of CPU times of IPS1 and IPS2 relative to IIPS. The horizontal axis in each subpanel represents n, the number of variables. The subpanels correspond to different values of p.

on each maximal prime subgraph. The combination of such decomposition and the IIPS procedure may further speed up computation.

For graphical models with contingency tables, Badsberg and Malvestuto (2001) implemented the IPS procedure on (fast) Markovian propagation trees. In their Markovian propagation algorithm, after one clique marginal c of G contained in a node of Markovian propagation tree is adjusted, they send messages to all the nodes of the Markovian propagation tree such that the probabilistic database $\mathbf{Q}$ is consistent. However, their propagation schedule is not very efficient, because it computes the marginal distributions of all nodes, when only one marginal distribution is required in the next proportional scaling step. We can improve efficiency of their algorithm via sharing computations similarly to the IIPS procedure when computing the maximum likelihood estimators for graphical models with contingency tables, that is, we can compute discrete conditional probability tables instead of regression coefficients and conditional covariance matrices. The same idea may work for graphical models with mixed data, but further investigations are needed. Finally, some of the ideas pursued here may also be useful for sampling from graphical models as discussed by Carvalho, Massam, and West (2007); we hope to explore this direction in future work.

SUPPLEMENTAL MATERIALS

Appendix: Theorem 1 and the validity of the IIPS procedure are proven. (appendix.pdf)

Online Supplement: The supplement contains computer codes and illustrations how to use the codes. It can be obtained via a single download. (Supplement.zip)

ACKNOWLEDGMENTS

The authors thank an associate editor and two referees for helpful comments and for bringing to our attention the work of Hara and Takemura. The authors also thank Dr. Na Shan for helpful discussions. This research was supported by the National Natural Science Foundation of China (grants 10701022, 10871038, 10828102, and 10926186), the National 973 Key Project of China (2007CB 311002), and the U.S. National Science Foundation award DMS-0630950.

[Received March 2009. Revised May 2010.]

REFERENCES

Anderson, T. W. (2003), *Introduction to Multivariate Statistical Analysis* (3rd ed.), New York: Wiley. [417]

Badsberg, J. H., and Malvestuto, F. M. (2001), "An Implementation of the Iterative Proportional Fitting Procedure by Propagation Trees," *Computational Statistics & Data Analysis*, 37, 297–322. [418,429]

Berry, A., Blair, J. R. S., Heggernes, P., and Peyton, B. W. (2004), "Maximum Cardinality Search for Computing Minimal Triangulations of Graph," *Algorithmica*, 39, 287–298. [425]

Blair, J. R. S., and Peyton, B. W. (1993), "An Introduction to Chordal Graphs and Clique Trees," in *Graph Theory and Sparse Matrix Computation. IMA Volumes in Mathematics and Its Applications*, Vol. 56, New York: Springer-Verlag, pp. 1–30. [425]

Bron, C., and Kerbosch, J. (1973), "Algorithm 457: Finding All Cliques of an Undirected Graph," *Communications of the ACM*, 16, 575–577. [425]

Carvalho, C. M., Massam, H., and West, M. (2007), "Simulation of Hyper-Inverse Wishart Distributions in Graphical Models," *Biometrika*, 94 (3), 647–659. [429]

Cowell, R. G., David, A. P., Lauritzen, S. L., and Spiegelhalter, D. J. (1999), *Probabilistic Networks and Expert Systems*, New York: Springer. [419]

Csiszár, I. (1975), "I-Divergence Geometry of Probability Distributions and Minimization Problems," *The Annals of Probability*, 3, 146–158. [417]

——— (1989), "A Geometric Interpretation of Darroch and Ratcliff's Generalized Iterative Scaling," *The Annals of Statistics*, 17, 1409–1413. [417]

Darroch, J. N., and Ratcliff, D. (1972), "Generalized Iterative Scaling for Log-Linear Models," *The Annals of Mathematical Statistics*, 43, 1470–1480. [417,418]

Deming, W. E., and Stephan, F. F. (1940), "On a Least Squares Adjustment of a Sampled Frequency Table When the Expected Marginal Totals Are Known," *The Annals of Mathematical Statistics*, 11, 427–444. [417]

Dempster, A. P. (1972), "Covariance Selection," *Biometrics*, 28, 157–175. [417]

Fienberg, S. E. (1970), "An Iterative Procedure for Estimation in Contingency Tables," *The Annals of Mathematical Statistics*, 41, 907–917. [417,418]

Gao, W., Shi, N. Z., Tang, M. L., Fu, L. Y., and Tian, G. L. (2010), "Unified Generalized Iterative Scaling and Its Applications," *Computational Statistics and Data Analysis*, 54, 1066–1078. [418]

Hara, H., and Takemura, A. (2010), "A Localization Approach to Improve Iterative Proportional Scaling in Gaussian Graphical Models," *Communications in Statistics—Theory and Methods*, 39, 1643–1654. [418,421,428]

Jensen, F. V., and Jensen, F. (1994), "Optimal Junction Trees," in *Proceedings of the Tenth Annual Conference on Uncertainty in Artificial Intelligence*, San Mateo, CA: Morgan Kaufmann, pp. 360–366. [425]

Jirousek, R., and Preucil, S. (1995), "On the Effective Implementation of the Iterative Proportional Fitting Procedure," *Computational Statistics & Data Analysis*, 19, 177–189. [418]

Kjærulff, U. (1992), "Optimal Decomposition of Probabilistic Networks by Simulated Annealing," *Statistics and Computing*, 2, 7–17. [425]

Kullback, S. (1968), "Probability Densities With Given Marginals," *The Annals of Mathematical Statistics*, 39, 1236–1243. [417]

Larrañaga, P., Kuijpers, C. M. H., Poza, M., and Murga, R. (1997), "Decomposing Bayesian Networks: Triangulation of the Moral Graph With Genetic Algorithms," *Statistic and Computing*, 7, 19–34. [425]

Lauritzen, S. L. (1996), *Graphical Models*, Oxford: Clarendon Press. [417-420,428]

Lauritzen, S. L., and Jensen, F. (2001), "Stable Local Computation With Conditional Gaussian Distributions," *Statistics and Computing*, 11, 191–203. [421]

Leimer, H. G. (1993), "Optimal Decomposition by Clique Separators," *Discrete Mathematics*, 113, 99–123. [428]

Li, H., and Gui, J. (2006), "Gradient Directed Regularization for Sparse Gaussian Concentration Graphs, With Applications to Inference of Genetic Networks," *Biostatistics*, 7, 302–317. [417]

Ma, S., Gong, Q., and Bohnert, H. J. (2007), "An Arabidopsis Gene Network Based on the Graphical Gaussian Model," *Genome Research*, 17, 1614–1625. [417,426]

Malvestuto, F. M., and Moscarini, M. (2000), "Decomposition of a Hypergraph by Partial-Edge Separators," *Theoretical Computer Science*, 237, 57–79. [428]

Rüschendorf, L. (1995), "Convergence of the Iterative Proportional Fitting Procedure," *The Annals of Statistics*, 23, 1160–1174. [418]

Speed, T. P., and Kiiveri, H. T. (1986), "Gaussian Markov Distribution Over Finite Graphs," *The Annals of Statistics*, 14, 138–150. [417,418,420]

Stoer, J., and Bulirsch, R. (2002), *Introduction to Numerical Analysis*, New York: Springer. [426]

Wang, X. F., and Guo, J. H. (2008), "Junction Trees of General Graphs," *Frontiers of Mathematics in China*, 3, 399–413. [419]

Wen, W. X. (1990), "Optimal Decomposition of Belief Networks," in *Proceedings of the Sixth Annual Conference on Uncertainty in Artificial Intelligence*, New York: Elsevier, pp. 209–224. [425]

Whittaker, J. (1990), *Graphical Models in Applied Multivariate Statistics*, Chichester: Wiley. [417,418,420,421]

Xu, P. F., and Guo, J. H. (2010), "A New Algorithm for Decomposition of Graphical Models," *Acta Mathematicae Applicatae Sinica*, to appear. [425]