Statistical Science
2005, Vol. 20, No. 4, 388–400
DOI 10.1214/088342305000000304

Experiments in Stochastic Computation for High-Dimensional Graphical Models

Beatrix Jones, Carlos Carvalho, Adrian Dobra, Chris Hans, Chris Carter and Mike West

Abstract. We discuss the implementation, development and performance of methods of stochastic computation in Gaussian graphical models. We view these methods from the perspective of high-dimensional model search, with a particular interest in the scalability with dimension of Markov chain Monte Carlo (MCMC) and other stochastic search methods. After reviewing the structure and context of undirected Gaussian graphical models and model uncertainty (covariance selection), we discuss prior specifications, including new priors over models, and then explore a number of examples using various methods of stochastic computation. Traditional MCMC methods are the point of departure for this experimentation; we then develop alternative stochastic search ideas and contrast this new approach with MCMC. Our examples range from low (12–20) to moderate (150) dimension, and combine simple synthetic examples with data analysis from gene expression studies. We conclude with comments about the need and potential for new computational methods in far higher dimensions, including constructive approaches to Gaussian graphical modeling and computation.

Key words and phrases: Decomposable models, nondecomposable models, Markov chain Monte Carlo, shotgun stochastic search, parallel implementation.

Beatrix Jones is Lecturer, Institute of Information and Mathematical Sciences, Massey University, Albany, New Zealand (e-mail: m.b.jones@massey.ac.nz). Carlos Carvalho is a student, Institute of Statistics and Decision Sciences, Duke University, Durham, North Carolina 27708, USA (e-mail: carlos@stat.duke.edu). Adrian Dobra is Assistant Research Professor, Institute for Genome Sciences and Policy and Institute of Statistics and Decision Sciences, Duke University, Durham, North Carolina 27708, USA (e-mail: adobra@stat.duke.edu). Chris Hans is a student, Institute of Statistics and Decision Sciences, Duke University, Durham, North Carolina 27708, USA (e-mail: hans@stat.duke.edu). Chris Carter is Principal Research Scientist, CSIRO Mathematical and Information Sciences, Sydney, Australia (e-mail: Chris.Carter @csiro.au). Mike West is Professor, Institute for Genome Sciences and Policy and Institute of Statistics and Decision Sciences, Duke University, Durham, North Carolina 27708, USA (e-mail: mw@stat. duke.edu).

1. INTRODUCTION

The last decade or so has witnessed a revolution in the statistical sciences, based on developments in stochastic simulation methods for scientific computation. The impact on applied Bayesian statistics has been particularly notable, with the development of Markov chain Monte Carlo (MCMC) methods that enable the application of increasingly rich and more relevant mathematical models. In tandem with model complexity is the radically increasing capacity to generate data sets that involve many, many variables. From high-frequency finance and enormous marketing data bases to gene expression studies in functional genomics, we are now faced with applied problems typified by very high-dimensional variables and/or parameter spaces. The use of stochastic computation methods to search over increasingly high-dimensional model spaces raises challenges of both statistical and computational efficiency as well as basic feasibility.

We are interested in precisely these questions—statistical and computational efficacy, and scalability with dimension—of stochastic computational methods

used to explore spaces of Gaussian graphical models. In a graphical model of a multivariate distribution, nodes represent variables and edges represent pairwise dependencies, with the edge set defining the global conditional independence structure of the distribution. Understanding the conditional independence structure is a critical element in trying to "make sense" of problems in which the number of variables far exceeds the number of observations. This is the case in analyzing gene expression data, a key example and motivating context for us. (Note that our objective is to *infer* the conditional independence structure, rather than to exploit structures implied by biology, the approach taken in Lauritzen and Sheenan, 2003.) Understanding conditional independence relationships is complementary to the approaches of Zhou, Kao and Wong (2002), which use graphical structures that represent pairwise correlations to elucidate genetic functions.

We focus on undirected graphical models, which have benefited from a good deal of interest in the computational statistics literature in recent years (e.g., Giudici and Green, 1999; Roverato, 2002; Atay-Kayis and Massam, 2006; Dellaportas, Giudici and Roberts, 2003; Wong, Carter and Kohn, 2003). Despite this attention and the need to develop methodology for increasingly large high-dimensional problems, the recent literature primarily focuses on relatively small problems (Wong, Carter and Kohn, 2003, is a notable exception). The methodological issues faced as dimension grows include questions of appropriate priors for the graphical structure, as well as the very challenging problem of searching over the space of graphs to identify high posterior regions. A number of computational methods (greedy search, simulated annealing, MCMC) have been suggested for complex variable/model selection problems such as this, but little is known about their performance and scalability as dimension increases.

We begin by reviewing some of the basic structure and recent advances for undirected graphical models, and then summarize our experiences trying to utilize stocastic computation in problems with a moderate (12–20) to large (150) number of variables. We introduce new methodology motivated by these experiences, including priors over graph space that encourage parsimonious models and a parallelizable stochastic search method for rapid traversal of spaces of the graph. The examples combine simple synthetic examples with data analysis from gene expression studies. We conclude the paper with a discussion of novel, alternative constructive approaches that are able to move to far higher dimensions, comments about the potential for theoretical advances to improve stochastic computation for these models and also discussion of hybrid approaches that combine "aggressive" moves in model space with the "local move" approaches that underlie current methods. We also comment on the need for increased development of distributed computational tools.

2. GAUSSIAN GRAPHICAL MODELS

Graphical models provide representations of the conditional independence structure of a multivariate distribution and access to efficient algorithms for computation of conditional and marginal densities (Whittaker, 1990; Lauritzen, 1996, Andersson, Madigan, Perlman and Richardson, 1999; Cowell, Dawid, Lauritzen and Spiegelhalter, 1999). The computational efficiencies arise through decompositions of the sample space into subsets of variables (graph vertices) based on their graphical relationships. The joint distribution of the variables is Markov over its graph, so likelihoods, priors and posteriors can be computed separately on the subsets of vertices and then reassembled into a likelihood or density that incorporates all variables (Hammersley and Clifford, 1971; Dawid and Lauritzen, 1993).

2.1 Graph Notation and Structure

The basic terminology and ideas for graphical models (Cowell et al., 1999), and the notation used here begin with a graph $G = \{V, E\}$, where G is defined over the set of vertices (the variables) V by the edge set E. A graph is complete if E contains all possible edges; otherwise it is incomplete. An incomplete graph G *decomposes* into disjoint subgraphs A, B and C (with $A \cup B \cup C = G$) if C is complete, and *separates* A and B (any path from a vertex in A to a vertex in B goes through C). The subgraph C is a *separator*. The decomposition is *proper* if neither A nor B is empty. If the separator C is always chosen to be minimal (so that it does not contain a proper subgraph that separates A and B), then iterative, proper decomposition of the graph G ultimately results in its *prime components*: a collection of subgraphs that cannot be further decomposed.

In a *perfect ordering* $P_1; S_2, P_2; S_3, P_3; \ldots$ of prime components (P_i's) and separators (S_i's), S_i is the intersection of P_i and all lower numbered components. We call the prime component sequence G^i and the separator sequence S^i. More than one perfect ordering may

exist for any given graph. Efficient algorithms for producing a perfect ordering of a given graph (including graphs with some incomplete prime components) were outlined by Dobra and Fienberg (2000).

If all the prime components of a graph are complete, the graph is said to be *decomposable*. Maximal complete subgraphs are called *cliques*, so the prime components of a decomposable graph are all cliques. When we refer exclusively to prime components that are cliques, we use C to denote the component rather than P. Decomposable graphs have distributional properties that make them particularly tractable, as we shall see below.

2.2 Density Factorization and Likelihood

The factorization of joint distributions that satisfy the conditional independencies implied by the edge structure of a given graph is key to the development of graphical model analyses. In general, a multivariate distribution on the specified graph G factorizes into terms that correspond to the prime components and separators of any perfect ordering for G. In the special case of a multivariate Gaussian distribution, Wermuth (1976) showed the edges of the graph correspond to nonzero elements in the precision matrix $\Omega = \Sigma^{-1}$. Dempster (1972) extensively considered this problem, referring to it as *covariance selection*. The density for a random sample of size n, $y = \{y_1, \ldots, y_n\}$, on the graph G is a function of multivariate Gaussian densities on the prime components and separators, with covariance matrices Σ_{PP} and Σ_{SS} on prime components and separators:

$$p(y|\Sigma_G) = \frac{\prod_{P \in G^i} p(y_P|\Sigma_{PP})}{\prod_{S \in S^i} p(y_S|\Sigma_{SS})}. \tag{1}$$

From a Bayesian perspective, we are interested in posterior distributions $p(G, \Sigma|y) = p(\Sigma|G, y)p(G|y)$ for specified priors $p(G, \Sigma) = p(\Sigma|G)p(G)$. (We should properly index Σ by G to indicate the constraints imposed by the graph, but we avoid that for simplicity of notation; it should be understood throughout.)

2.3 Priors and Posteriors for Covariance Matrices

Giudici (1996) discussed the major approaches to prior specification for Σ, comparing the "local priors" described by Dawid and Lauritzen (1993), and the "global priors" based on the conditional approach of Dickey (1971). These priors have the desirable property that $p(\Sigma|G)$ is consistent over graphs: the (i, j) element of Ω has the same prior whenever the graph does not constrain the (i, j) element to be zero. Roverato (2002) extended the local priors to general, nondecomposable models. Giudici (1996) suggested that the local priors encourage sparser graphs; for that reason, we use the local priors. The computational issues are similar whichever class is chosen.

The local prior $p(\Sigma|G)$ is *hyper-inverse Wishart*, HIW(G, δ, Φ), with Φ a positive definite matrix and $\delta > 0$. Like the likelihood (1), this density factors over the prime components and separators:

$$p(\Sigma|G) = \frac{\prod_{P \in G^i} p(\Sigma_{PP}|G)}{\prod_{S \in S^i} p(\Sigma_{SS}|G)}. \tag{2}$$

For each complete prime component P of G (and each separator), the corresponding submatrix of the covariance, Σ_{PP}, has an inverse Wishart(δ, Φ_{PP}) prior (as given by Giudici, 1996). Decomposable graphs consist entirely of complete prime components, so these inverse Wisharts fully define the density of Σ when we restrict consideration to decomposable graphs. The tractability of decomposable graphs is explained by the fact that while the graphical structure determines which entries of the covariance matrix appear in the density, the entries that do appear are constrained only to define full rank multivariate normal distributions on the cliques of the graph; the other entries of Σ are functions of these free entries (Grone, Johnson, de Sá and Wolkowicz, 1984).

Roverato (2002) generalized the inverse Wishart to define a density suitable for a noncomplete prime component P. This density for Σ_{PP} is obtained from a Wishart prior on Ω_{PP}, conditioned on Ω_{PP} consistent with G, by a change of variables. While based on conditioning, this prior differs from the global prior of Giudici (1996) in that the conditioning is only used within the prime components. In this density, some of the nonfree elements of Σ_{PP} will appear; consequently, the integral that defines the normalizing constant must be computed numerically. When we use the hyper-inverse Wishart prior with an unrestricted graph space, we constrain δ to be strictly greater than 2.0; it has not been shown that the prior is proper for smaller δ.

The hyper-inverse Wishart prior is conjugate in either the decomposable or unrestricted case; the posterior is HIW$(G, \delta^* = \delta + n, \Phi^* = \Phi + S_y)$, where S_y is the observed sum of products matrix, $\sum_{i=1}^n y_i y_i'$. In subsequent examples, we use $\Phi = \tau I$ for specified constants τ (other choices for Φ, such as an intraclass correlation structure, were considered by Giudici and Green, 1999). Our choice is consistent with problems in which variables represent measures of similarly defined quantities on a common scale. Choice of τ is im-

portant; simulations show increasing τ increases the marginal likelihood of high edge count graphs (data not shown). The marginal prior mode for each variance term (σ_{ii}) is $\tau(\delta+1)$; we use this quantity to set an appropriate value for τ. For example, if the data have been standardized so all the variances are 1.0, τ might be set to $1/(\delta+1)$.

2.4 Priors and Likelihoods for Graphs

A uniform prior over all graphs or all decomposable graphs assigns most of its mass to graphs with a "medium" number of edges. The mass function peaks around $|V|(|V|-1)/4$ for general graphs (where $|V|$ is the number of vertices); we have used simulation from the prior to estimate the distribution when we restrict to decomposable graphs (data not shown). In both cases the average number of edges explodes very quickly as the number of nodes increases.

We would like to represent the conditional independence structure parsimoniously and discourage the inclusion of spurious edges; in other words, we would like to encourage *sparse* graphs, especially as dimension increases. To do this we use a Bernoulli prior on each edge inclusion indicator variable with parameter $\beta = 2/(|V|-1)$. Thus a graph with $|E|$ edges has prior probability $\beta^{|E|}(1-\beta)^{\binom{|V|}{2}-|E|}$. For an unrestricted graph, this distribution has its peak at $|V|$ edges; the mode is somewhat lower when we restrict to decomposable graphs. Our approach to prior specification penalizes the number of edges; one could, of course, penalize other measures of complexity such as the maximum or average prime component size. Wong, Carter and Kohn (2003) developed an approach that equalizes the prior probability of graphs with different numbers of edges; for decomposable graphs, this requires estimating the fraction of the total number of decomposable graphs with each number of edges.

The marginal likelihood of any graph G is a simple function of the HIW prior and posterior normalizing constants, $h(G,\delta,\Phi)$ and $h(G,\delta^*,\Phi^*)$:

$$p(y|G) = (2\pi)^{-n|V|/2}\frac{h(G,\delta,\Phi)}{h(G,\delta^*,\Phi^*)}. \tag{3}$$

For a decomposable graph, the HIW normalizing constants can be computed from the normalizing constants for the inverse Wishart clique and separator densities. For nondecomposable graphs, the analogous terms from incomplete prime components do not have closed form. Monte Carlo methods for estimating these terms are discussed in Section 3.

3. COMPUTING LIKELIHOODS FOR NONDECOMPOSABLE MODELS

For nondecomposable models, the contribution to (3) from an incomplete prime component P does not have a closed form expression. (To simplify notation throughout this section, we assume that P constitutes the whole graph, so subscripting by P can be omitted.) The term is the normalizing constant of a constrained inverse Wishart distribution and can be expressed as an integral over the space of Ω^E, the nonzero elements of Ω as dictated by the edge set E. To estimate this integral, we use the method presented by Atay-Kayis and Massam (2006). They exploit two changes of variables: from Ω^E to ϕ^E, the free elements of the upper triangular matrix produced by the Cholesky decomposition of Ω, and from ϕ^E to ψ^E, where $\psi = \phi T^{-1}$ and $T'T$ is the Cholesky decomposition of Φ. After this second change, the free elements of ψ are independent normals and square roots of χ^2 random variables, and thus are easily generated; the nonfree elements can be straightforwardly computed from the free elements. The relevant integral is then

$$h(P,\delta,\Phi) = C\mathrm{E}_{\psi^E}(f_T(\psi^E)), \tag{4}$$

where C is a constant further discussed in Section 7.1, E_{ψ^E} denotes expectation with respect to the distribution of ψ^E and

$$f_T(\psi^E) = \exp\left\{-\tfrac{1}{2}\sum_{(i,j)\notin E, i<j}\psi_{ij}^2\right\}. \tag{5}$$

Values of ψ^E can be easily sampled, so it is straightforward to estimate the expectation of (5) by Monte Carlo. Note that when P is a clique, (5) evaluates to 1 and (4) simplifies to an inverse Wishart normalizing constant.

Roverato (2002) and Dellaportas, Giudici and Roberts (2003) presented alternative methods for estimating this normalizing constant based on generating Ω from approximations to its actual distributions; the relevant integral is then estimated by importance sampling. We prefer (and use throughout our examples) the method of Atay-Kayis and Massam (2006) because it avoids worries about the efficiency of the importance sampler, that is, how far the sampling distribution of the Ω's differs from the desired distribution.

4. LOCAL UPDATES FOR DECOMPOSABLE MODELS

In addition to having analytical expressions for their normalizing constants, decomposable graphs have

computationally efficient "local updates" in model search based on comparing decomposable graphs G and G' that differ by one edge only. Computing the likelihood ratio $p(Y|G)/p(Y|G')$ requires far less effort than computating either likelihood. This property was exploited by Giudici and Green (1999), and more fully explained by Armstrong, Carter, Wong and Kohn (2005). Suppose G is produced from G' by deleting edge $\{a, b\}$. The fact that both are decomposable implies that in G', $\{a, b\}$ lies in a single clique, C_q. At most one of a and b lies in the separator S_q. Armstrong, Carter, Wong and Kohn (2005) showed that if $a \notin S_q$, the perfect ordering of G is identical to that of G' except C_q is replaced with consecutive cliques, $C_{q_1} = C_q/a$ and $C_{q_2} = C_q/b$, with separator $S_{q_2} = C_q/\{a, b\}$. Consequently the likelihood ratio simplifies to an expression that involves only C_q, C_{q_1}, C_{q_2} and S_{q_2}.

In contrast, when we do not restrict ourselves to decomposable graphs, there is no guarantee of significant cancellations in the likelihood ratio between graphs that differ by one edge. Imagine starting with a graph where all the nodes are connected in a chain and then adding the edge that completes the full cycle. The single edge change moves us from a situation with $p - 1$ prime components to a single prime component; there is no cancellation in the likelihood ratio.

5. MARKOV CHAIN MONTE CARLO ALGORITHMS

The MCMC is a much used tool for exploring the space of graphical structures (e.g., Madigan and York, 1995; Dellaportas and Forster, 1999; Giudici and Castelo, 2003). In the context of Gaussian graphical models, Armstrong, Carter, Wong and Kohn (2005) used their results to construct a fixed scan Gibbs sampler for decomposable graphs; their results are also easily exploited in a Metropolis–Hastings sampler. We constructed three samplers to traverse the space of decomposable graphs: fixed scan Gibbs, Metropolis–Hastings where the edge to be updated was picked at random, and Metropolis–Hastings where the choice to add or delete an edge was made and then an edge was selected at random from those appropriate for that type of move. There was no noticeable difference in performance between these closely related MCMC algorithms; the results presented are from the add–delete Metropolis–Hastings sampler.

We also implemented the add–delete Metropolis–Hastings sampler for an unrestricted search of graph space. When evaluating a proposal that involves a nondecomposable graph, the algorithm described in Section 3 is used to evaluate the marginal likelihood. This adds considerable computational burden; see Tables 1 and 2. In addition, because the local computation properties described in Section 4 no longer hold, we recompute the perfect ordering and entire likelihood for each proposed graph.

For problems with even a moderate number of variables (either in the decomposable or unrestricted space), the space to be explored is so large that a graph's frequency in the sample of graphs produced cannot be viewed as reflecting its posterior probability. Indeed, many graphs are not revisited after the chain leaves them. Posterior graph probability estimates must be based on normalizing the posterior mass function using the visited graphs, and these quantities will reflect the true posterior mass only to the extent that the majority of the mass has been visited. However, the frequencies of other quantities, such as the marginal probabilities of edge inclusion, can be viewed as posterior probabilities.

TABLE 1
Comparison between algorithms of run time and quality of best graph found for the 12 *node example*

Method[a]	Run time (sec)	Max log posterior	Graphs to first top graph visit	Time to first top graph visit
MH-d	36	−2591.18	912	1
SSS-d	183	−2591.18	792	2
MH-u	15,220	−2590.94	415	2
SSS-u	2773	−2590.94	13,266	5

[a]MH-d (-u) refers to the Metropolis–Hastings algorithm on decomposable (unrestricted) models, while SSS-d (-u) refers to the shotgun stochastic search method on decomposable (unrestricted) models.

TABLE 2
Comparison between algorithms of run time and quality of best graph found for the 15 *node example*

Method[a]	Run time (sec)	Max log posterior	Graphs to first top graph visit	Time to first top graph visit
MH-d	93	15633.76	349,484	36
SSS-d	234	15633.76	33,495	9
MH-u	513,077	15633.83	666,425	309, 222
SSS-u	5930	15636.38	82,845	112

[a]MH-d (-u) refers to the Metropolis–Hastings algorithm on decomposable (unrestricted) models, while SSS-d (-u) refers to the shotgun stochastic search method on decomposable (unrestricted) models.

6. SHOTGUN STOCHASTIC SEARCH ALGORITHMS

If Markov chain Monte Carlo is viewed merely as a tool for visiting high probability regions of graph space, there are certainly competing algorithms. The following algorithm is attractive because step 2 (which contains most of the computational burden) can be easily parallelized.

1. Start with a graph G.
2. Select at random X_1 graphs that differ by one edge (neighbors), compute their unnormalized posterior mass and retain the top $X_2 \leq X_1$.
3. From among the X_2 top neighbors, propose the ith graph G_i as a new starting graph with probability proportional to p_i^α, where p_i is the unnormalized posterior probability of graph i and α is an annealing parameter.
4. Return to step 2 and iterate. Maintain a list of the overall best X_3 graphs visited.

In experimenting with this approach, we have typically used $X_1 = X_2 = \binom{|V|}{2}$, so all the neighbors are examined at each stage. We refer to this as a *shotgun stochastic search* (SSS) method; at each step we generate a large number of candidate models, "shooting out" candidates in all directions and then following one (or, in a variant of the above, more than one) plausible candidate. Algorithms of this type can accommodate either unrestricted or decomposable graphs. When restricted to decomposable graphs, step 2 contains a check for decomposability; nondecomposable graphs are considered to have zero posterior probability.

Posterior probabilities can be normalized only within the list of the top X_3 graphs; this reflects their true posterior probability to the extent that they contain most of the posterior mass. Similarly, estimated edge probabilities can be viewed as posterior probabilities only to the extent that the whole posterior mass is captured in the top X_3 graphs.

7. SIMULATED EXAMPLES

We first consider two simulated examples where the true underlying graph is known. The first graph, pictured in Figure 1, has 15 nodes and is decomposable. The second graph, pictured in Figure 2 consists of 12 nodes in a single noncomplete prime component. Each data set consists of 250 observations. The first simulated data set was inspired by patterns of daily currency exchange fluctuations against the U.S. dollar. Consequently, the data range approximately between $\pm 2\%$.

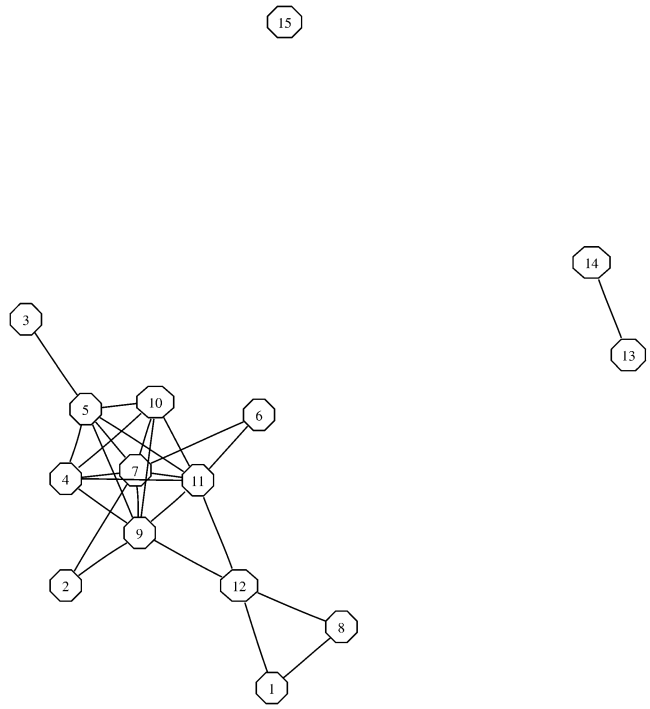

FIG. 1. *The true underlying decomposable graph on $p = 15$ nodes—the first simulated example.*

We assume this range is about 2 standard deviations, so $\sigma_{ii} \approx 0.0001$. We choose $\delta = 3$ and $\tau = 0.0004$. For the second data set, Σ is actually a random draw from the inverse Wishart$(I, 3)$ constrained to obey the graph; thus we use $\tau = 1$ and $\delta = 3$. In both cases the prior

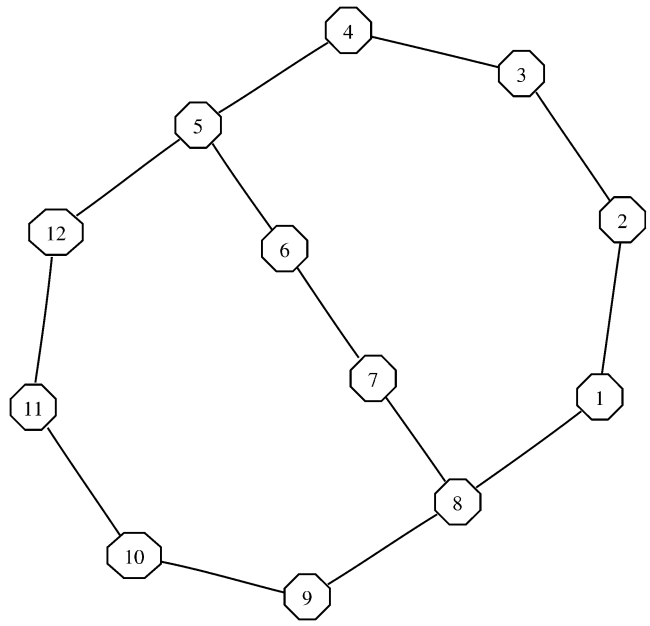

FIG. 2. *The true underlying nondecomposable graph on $p = 12$ nodes—the second simulated example.*

over graphs is the sparsity-encouraging prior suggested in Section 4. For the shotgun stochastic search, the annealing parameter was set at 1.0 for simplicity. Performance of the algorithm in larger examples is very sensitive to the annealing parameter; see Section 8 for details.

7.1 Difficulties Evaluating Nondecomposable Models

To search the unrestricted model space, we must specify the number of random draws that will be used to estimate the normalizing constants for nondecomposable prime components. Initial runs with 1000 draws, regardless of prime component size, revealed an important and, we believe, both generic and limiting problem: high variance estimates of the marginal likelihood (standard deviation on the order of 2 units of log likelihood) created artificial local modes, greatly inhibiting the algorithms' movement.

To explore the behavior of the normalizing constant estimates, we examined noncomplete prime components with different numbers of nodes. Two examples for each size were selected from those that occurred during the Metropolis–Hastings model search for the 15 variable data set. Because our search strategies depend on likelihood ratios, it is the variances of the log normalizing constants that are relevant. Figure 3A and 3B shows the variances of the estimated log of the prior and posterior normalizing constants (where the estimate is based on 100 random draws). The plotted variances are of course estimates themselves, each based on 1000 separate normalizing constant estimations.

The estimates of the log prior normalizing constants have systematically smaller variances than the corresponding estimates for the posterior; there is also a tendency for variance to increase with component size. This can be partially explained by examining the form of ψ, the sampled matrix from which the estimate is computed. The variance of diagonal entries increases as one moves down the diagonal, so larger components are more variable; similarly, the parameters of the HIW posterior dictate that the diagonal entries have larger variance in the posterior.

We also note that the ordering of the variables used when setting up ψ affects the variance of the log normalizing constant. Each prime component considered in Figure 3C is a cycle; in the "optimal" configuration, each variable, except the first and the last, has exactly one neighbor preceding it in the rows of ψ. The "worst" configuration has the first $|P|/2$ variables each with both neighbors occurring further down in the matrix.

The cause of this phenomenon can be seen by factoring equation (4) into the constant C and the part estimated by Monte Carlo, $\mathrm{E}(M)$, where

$$C = \left(\prod_{i=1}^{|P|} 2^{(\delta+\nu_i)/2} (2\pi)^{\nu_i/2} \cdot \Gamma\left(\frac{\delta+\nu_i}{2}\right) T_{ii}^{(\delta+b_i-1)/2} \right), \tag{6}$$

$$\mathrm{E}(M) = \mathrm{E}_{\psi^E}(f_T(\psi^E)). \tag{7}$$

Recall that T_{ii} are the entries of the Cholesky decomposition of the HIW parameter Φ (Φ^* for the posterior), ν_i is the number of neighbors of node i subsequent to it in the ordering of vertices and b_i is the total number of neighbors of node i, plus 1. We list the variables of a prime component in an arbitrary order, but the relative sizes of C and $\mathrm{E}(M)$ clearly depend on the ordering of the variables (although their product is constant, the expression is valid for any ordering). In our experiments the variance of $\bar{M}$ was (roughly) unaffected by ordering; however, we are interested in the variance of $\log(\bar{M}) \approx \mathrm{Var}(\bar{M})/\bar{M}^2$. Thus, orderings that increase C and decrease $\mathrm{E}(M)$ increase the variance of $\log(\bar{M})$. The "optimal" ordering for cycles discussed above minimizes C for the HIW prior; the "worst" ordering maximizes it. Similar multiplicative differences due to different orderings were observed in estimates of the log posterior normalizing constant; however, because of the appearance of the data (through the T_{ii}) in the expression for C, the optimal ordering depends on the data as well as the graph structure. We will not try to optimize the ordering of variables, but rather develop a scheme that will produce adequate estimates for any ordering.

The highest variance samples in Figure 3B represent very low likelihood graphs, which have small $\mathrm{E}(M)$—and high variance of $\log(\mathrm{E}(M))$—regardless of ordering. Figure 3D, a plot of variances of log posterior normalizing constants for prime components in graphs *accepted* during the Metropolis–Hastings search, is more consistent with the variance trends in the log prior normalizing constants. The variance of the "worst case" for each component size seems to be a function of the size of the component considered, $|P|$. Based on this, we used $1.5|P|^3$ samples for the posterior normalizing constants and $0.5|P|^3$ for the prior normalizing constants. This scheme solved the problem

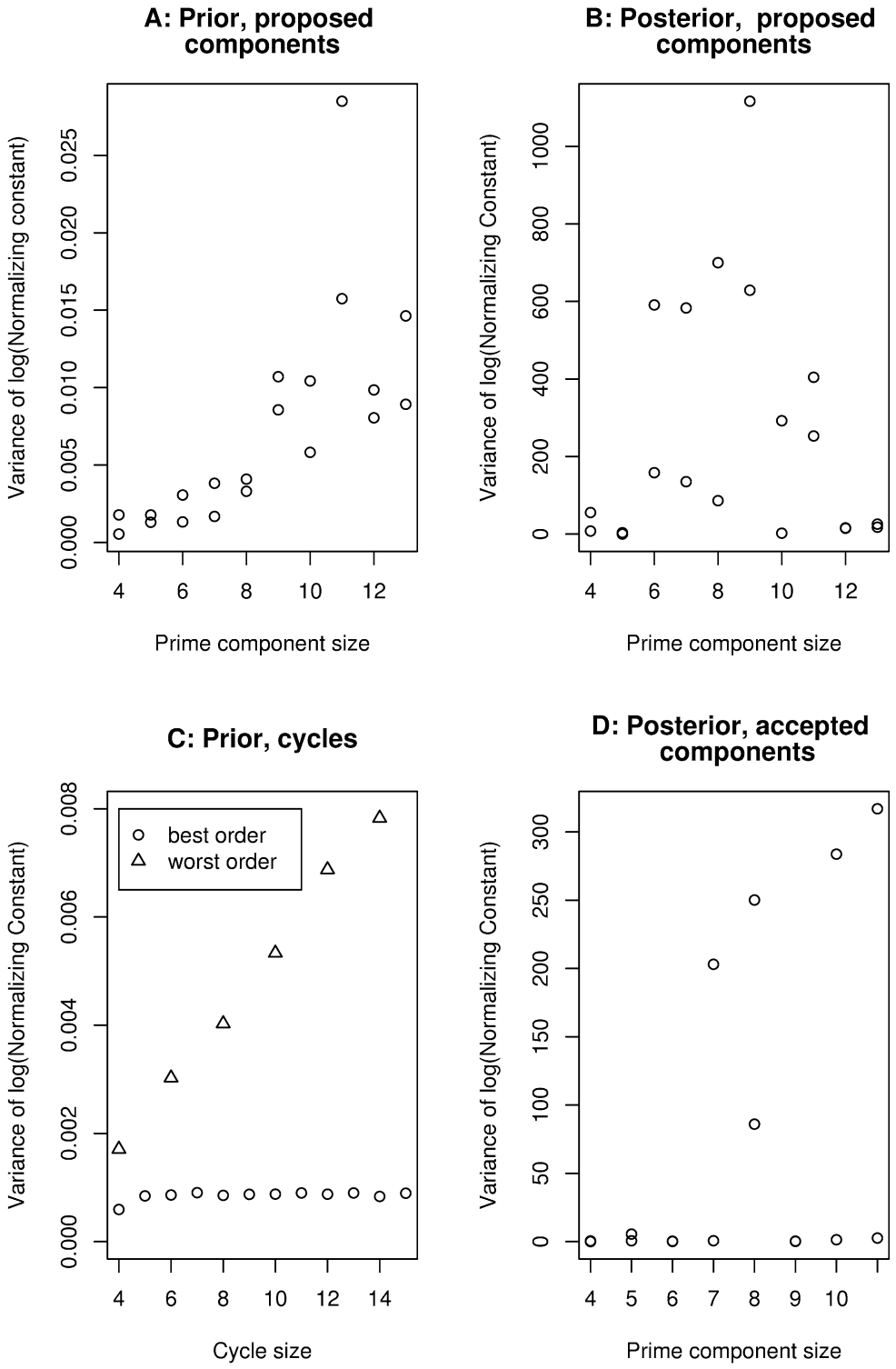

FIG. 3. *Relationship between the variance of the estimated normalizing constants, based on* 100 *samples, and the size of the prime component. Four cases are considered*: A, *the prior normalizing constant for components proposed during the unrestricted Metropolis–Hastings search for the* 15 *node data set*; B, *the posterior normalizing constants for these components*; C, *prior normalizing constants for cycles, using different variable orderings*; D, *posterior normalizing constants for components considered during the unrestricted model search and subsequently accepted by the Metropolis–Hastings algorithm*.

with chain mobility discussed at the beginning of this section. At the end of our run, all graphs with a log posterior within 2.0 of the top log posterior were reexamined with enough Monte Carlo runs to ensure the graph listed as "best" did indeed have the highest log posterior.

7.2 Results

For each example, the add–delete Metropolis was run for 10,000 $\times$ $\binom{|V|}{2}$ steps [where $\binom{|V|}{2}$ is the number of possible one edge moves in the unrestricted case]. The shotgun stochastic search algorithm was run with 10,000 iterations; at each iteration it considers all possible (unrestricted) one edge moves, so it performs the same number of graph comparisons as the Metropolis–Hastings algorithm. The searches were each started at the empty graph; both unrestricted and decomposable-only spaces were considered.

The algorithms clearly use a similar amount of computing resources, as they evaluate the same number of comparisons between current and proposed graphs. However, the stochastic search algorithm is parallelizable. The run times for both types of algorithm are given in Tables 1 and 2, demonstrating the advantage of being able to exploit multiple processors. The Metropolis–Hastings was run on a Dell PC with a 1.8 MHz Xeon processor in a Linux environment, and the shotgun stochastic search was run on a Beowulf cluster with 26 dual processor, 1.4 MHz nodes. The C++ implementations used are available at www.isds.duke.edu under the *software* link.

The "top" decomposable graphs—those identified with highest posterior probability—are pictured in Figures 4 and 5; the top graphs from the unrestricted search appear in Figures 6 and 7. Likelihood comparison with true graphs show that each of these graphs has greater likelihood (and posterior) support than the true graph. For the maximum posterior probability graphs, the edges included generally have higher estimated posterior probability than those not included. The 15 node decomposable graph includes the only observed exception to this: the lowest probability included edge has probability 0.58, while the highest probability excluded edge has probability 0.60. Thus aggregating high probability edges into a graph does not result in dramatically different graphs than taking the best graph found. The most probable graph found in the 12 node case and the decomposable cases was insensitive to the starting point: the same graph was found starting at the complete graph. The unrestricted search for the 15 node case starting at the complete graph did not attain the likelihood for the top graph shown in the table.

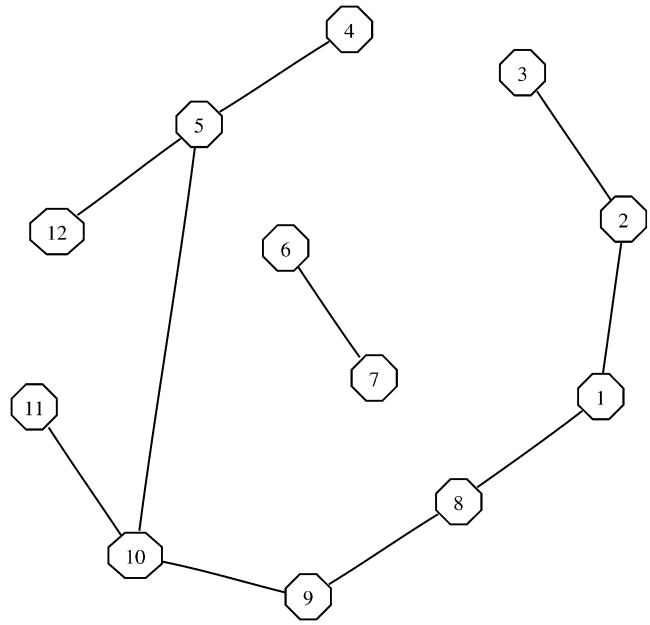

FIG. 4. *Highest log posterior graph for the* 12 *node example when the search is restricted to decomposable models.*

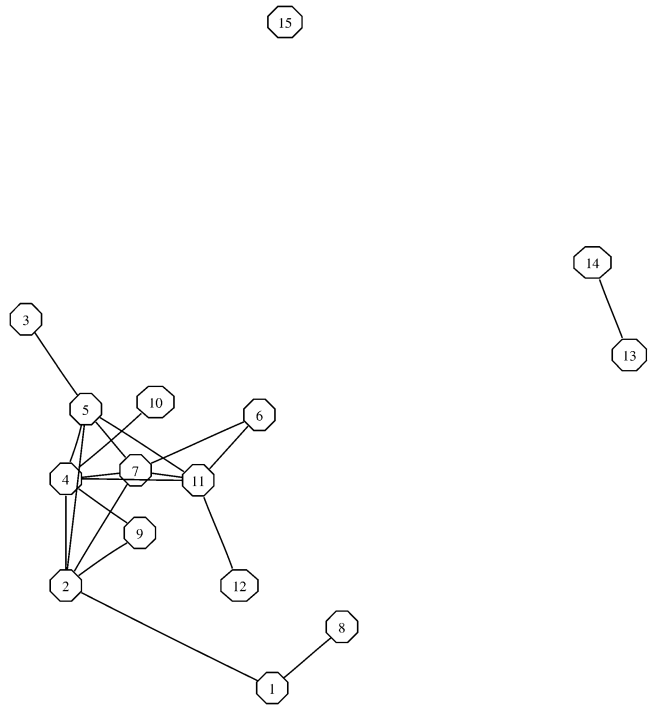

FIG. 5. *Highest log posterior graph for the* 15 *node example when the search is restricted to decomposable models.*

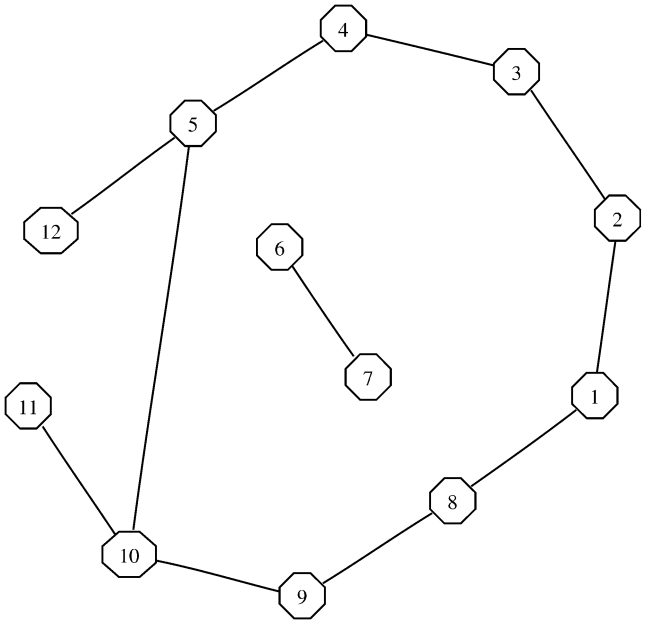

FIG. 6. *Highest log posterior graph for the* 12 *node example when the search is unrestricted.*

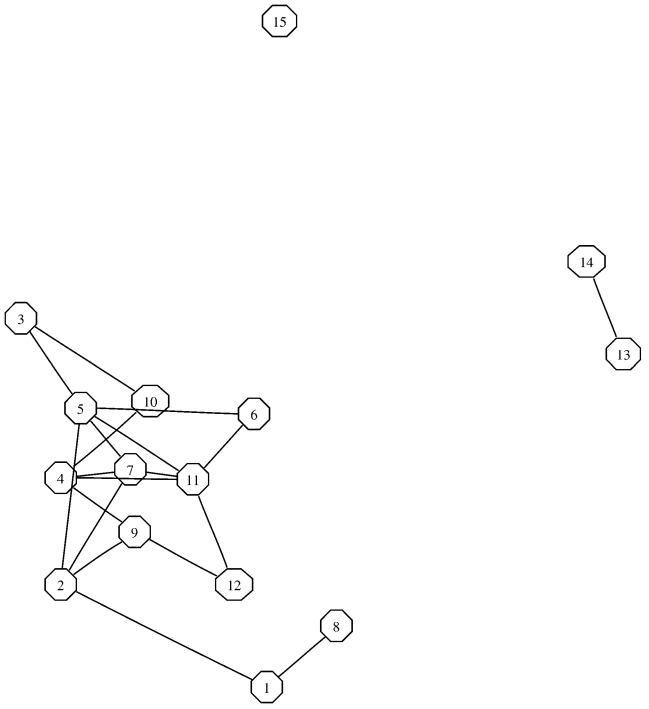

FIG. 7. *Highest log posterior graph for the* 15 *node example when the search is unrestricted.*

8. 150 NODE EXAMPLE: GENE EXPRESSION DATA

A more challenging problem is analysis of expression data from $p = 150$ genes associated with the estrogen receptor pathway, taken from $n = 49$ individuals; the data come from the study of West et al. (2001). The data were standardized and the prior was specified with $\delta = 3$, $\tau = 4$. In this context our sparsity-encouraging prior can be interpreted as a belief that, on average, each gene has major interactions with a relatively small number of other genes. In this large example, we add to the prior over graphs the restriction that the prime component/clique size not exceed $n - 1$, so as to maintain identifiability of the model.

The results from three algorithms are shown in Table 3. Times are now given in hours. Because the unrestricted search Metropolis–Hastings showed such poor performance, it was not used. The best results for the shotgun search algorithm were obtained with an annealing parameter of 50: essentially deterministic hill climbing. In this large example we see that even in the decomposable case, the shotgun stochastic search algorithm finds much more probable graphs.

TABLE 3
Comparison between algorithms of run time and quality of best graph found for the gene expression example

Method[a]	Run time (hrs)	Max log posterior	Graphs to first top graph visit	Time to first top graph visit
MH-d	18.02	−9417.97	100,466,818	6.51
SSS-d	0.03	−9260.84	1,698,600	0.03
SSS-u	6.29[b]	−9227.68	44,700	3.39

[a]MH-d refers to the Metropolis–Hastings algorithm on decomposable models, while SSS-d (-u) refers to the shotgun stochastic search method on decomposable (unrestricted) models.

[b]Starting from the best decomposable graph found.

A large annealing parameter was also used for the shotgun stochastic search in the unrestricted case. However, in this case the annealing does not eliminate the stochasticity of the search, as the marginal likelihoods are estimated with substantial error. Increasing the number of iterations enough to get a sharp evaluation of the likelihood was infeasible; settling for a standard deviation of the log likelihood of 1.0 resulted in one cycle of neighbors evaluations (a single step in our stochastic search procedure) taking up to 40 computer days (1 day on a 40 node cluster). Using this procedure, starting from the empty graph and running until the estimated log posterior stopped improving, the best graph found had log posterior −9364.67, worse than the best decomposable graph. This graph may represent a local mode not present in the decomposable framework or may be the result of suboptimal moves that result from the imprecise likelihood evaluation. The table shows the best graph found by starting at the best decomposable graph (the final estimate of the log posterior

for this graph was run with enough iterations to put the standard deviation below 0.1). A total of 10 cycles of evaluating all neighbors was done. Because these graphs were "close" to decomposable graphs, the evaluation time was reduced versus graphs with similar numbers of edges produced by the search starting at the empty graph.

9. DISCUSSION AND RECENT DEVELOPMENTS

Fitting decomposable Gaussian graphical models using local move methods is feasible for large numbers of variables, certainly up to a few hundred. Exploration of model space to find high posterior probability graphs can be successfully carried out using direct search such as with our shotgun stochastic search method; traditional MCMC is competitive for relatively small graphs. Unrestricted (nondecomposable and/or decomposable) model search is much more problematic; it is easily accomplished for 15 variables, but becomes very challenging quickly thereafter. Large prime components induce a major computational burden via the Monte Carlo estimation of the needed normalizing constants. Other methods are needed to deal with this computational problem. Local search of unrestricted graphs around "good" decomposable graphs or other candidate graphs is possible for 150 variables and represents a promising strategy. For unrestricted models the method of choice is never a Markov chain Monte Carlo algorithm, but rather the shotgun stochastic search that rapidly traverses graph model space around sequences of "promising" models. The specific stochastic search algorithm we have introduced and exemplified here is easily parallelizable and, indeed, designed for distributed implementation. More experimentation with the annealing schedules is needed to find optimal strategies for different situations. For the 150 node decomposable model search presented as an example here, deterministic hill climbing produced the best results in terms of rapid identification of high probability graphs.

In the case of unrestricted search, new theoretical insights and methods are needed to improve the capacity to estimate the normalizing constants associated with noncomplete prime components. One potential direction for research that would have immediate payoff involves a characterization of the changes in prime component structure when one edge moves are made from a current graph. Flores, Gámez and Olesen (2003) addressed this problem in the context of directed graphs; their results could be applied to provide characterization of prime component changes analogous to the results for clique changes in decomposable graphs used in Giudici and Green (1999). Correlating the marginal likelihood estimates of graphs that are to be compared by using the same random number draws to estimate the normalizing constants involved may also improve computational efficiency.

A rather different view—developed since this work was completed—was described by Dobra et al. (2004) and Dobra and West (2004). In these methods, the full joint distribution is derived using a triangular set of regressions that represent the relationships between variables. This methodology is related to both the dependency network framework of Heckerman et al. (2000) and approaches that model structure in the Cholesky decomposition of variance matrices; it is innovative in the creation of an approach that scales with dimension, encourages graph sparsity, utilizes priors consistent across graphs and generates many candidate graphs via MCMC methods for variable selection in the composing regressions. These methods can handle large sets of variables, partly by using a prescreening procedure that limits which variables will be considered possible predictors of others.

This type of constructive method generates graphs that are potentially far more widely separated than by one edge moves. An appealing concept is to integrate methods of this sort with the local-move methods described in this paper. Research to understand the theoretical differences, in terms of prior and model specifications, between such constructive approaches and the undirected Gaussian graphical models considered in our framework is necessary: our attempts to use the method of Dobra et al. (2004) to generate conditional independence structures that had high posterior probability under our model or were good search starting points for our algorithms were not successful. The constructive approaches based on regressions yield models that correspond to directed acyclic graphs, which clearly impact the regions of model space visited. However, our experiments with "local-move" methods lead us to conclude that a constructive approach of some form is needed to scale beyond moderate dimensions. While questions of search adequacy still exist, the constructive approaches are at least implementable for very large sets of variables: the example in Dobra and West (2004) concerns gene expression data on over 12,000 genes. Their example also seems to identify graphs that are interpretable and consonant with known biology.

Modeling of discrete (or discretized) data is an important alternative to the Gaussian formulation; these models can (at the expense of additional parameters) approximate nonlinear relationships between variables. Examples in the context of microarray data include Yu et al. (2004) and Friedman, Linial, Nachman and Pe'er (2000); feasible methods have not yet been demonstrated for a full chip's worth of gene expression measurements (several thousand). While the mechanics of evaluating a conditional independence structure for discrete data differ, the general lessons about search strategies over model space apply, and some of our current work is focused on large-scale discrete problems.

It is apparent that radical, near-term progress in model and variable selection/search in the face of increasing dimension is unlikely if computations are restricted to serial, single processors. Our experiments have heavily utilized a Beowulf cluster, and distributed computation is essential to the development of search and constructive methods beyond moderate dimensions. With increasing access to larger clusters for distributed computing, the computational statistics research community has come to an opportune time to substantially advance our ability to explore complex, high-dimensional model spaces by embracing this technology and integrating it into day-to-day research.

ACKNOWLEDGMENTS AND SUPPORTING MATERIAL

This work was developed during 2003 as part of the SAMSI program Stochastic Computation. The authors acknowledge the support of the National Science Foundation through the SAMSI Grant DMS-01-12069, and Grants DMS-01-02227 and DMS-01-12340 to Duke University, and grants from the Keck Foundation and NIH. Graphical displays (Figures 1–2, 4–7) are based on the AT&T Labs GraphViz software (available at `www.research.att.com/sw/tools/graphviz/`). The authors thank the editors and anonymous reviewers whose comments helped substantially in the revisions.

Interested readers can find an original, longer version of this paper—that includes additional graphical model review material, appendix material on computation of junction trees and marginal likelihood functions, as well as computer code for the analyses presented here—at the ISDS web site www.isds.duke.edu under the *software* link. Additional material of interest there includes the *GraphExplore* software for visualization and exploration of graphical models, and the HdBCS software that implements more recent methods for exploration of the graphical models of Dobra and West (2004).

REFERENCES

ANDERSSON, S. A., MADIGAN, D., PERLMAN, M. D. and RICHARDSON, T. (1999). Graphical Markov models in multivariate analysis. In *Multivariate Analysis, Design of Experiments and Survey Sampling* (S. Ghosh, ed.) 187–229. Dekker, New York.

ARMSTRONG, H., CARTER, C. K., WONG, K. F. and KOHN, R. (2005). Bayesian covariance matrix estimation using a mixture of decomposable graphical models. Unpublished manuscript.

ATAY-KAYIS, A. and MASSAM, H. (2006). The marginal likelihood for decomposable and non-decomposable graphical Gaussian models. *Biometrika*. To appear.

COWELL, R. G., DAWID, A. P., LAURITZEN, S. L. and SPIEGELHALTER, D. J. (1999). *Probabilistic Networks and Expert Systems*. Springer, New York.

DAWID, A. P. and LAURITZEN, S. L. (1993). Hyper-Markov laws in the statistical analysis of decomposable graphical models. *Ann. Statist.* **21** 1272–1317.

DELLAPORTAS, P. and FORSTER, J. J. (1999). Markov chain Monte Carlo model determination for hierarchical and graphical log-linear models. *Biometrika* **86** 615–633.

DELLAPORTAS, P., GIUDICI, P. and ROBERTS, G. (2003). Bayesian inference for nondecomposable graphical Gaussian models. *Sankhyā* **65** 43–55.

DEMPSTER, A. P. (1972). Covariance selection. *Biometrics* **28** 157–175.

DICKEY, J. M. (1971). The weighted likelihood ratio, linear hypotheses on normal location parameters. *Ann. Math. Statist.* **42** 204–223.

DOBRA, A. and FIENBERG, S. E. (2000). Bounds for cell entries in contingency tables given marginal totals and decomposable graphs. *Proc. Natl. Acad. Sci. U.S.A.* **97** 11,885–11,892.

DOBRA, A., HANS, C., JONES, B. NEVINS, J., YAO, G. and WEST, M. (2004). Sparse graphical models for exploring gene expression data. *J. Multivariate Anal.* **90** 196–212.

DOBRA, A. and WEST, M. (2004). Bayesian covariance selection. Available as Discussion Paper 04-23 at `www.isds.duke.edu`.

FLORES, M. J., GÁMEZ, J. A. and OLESEN, K. G. (2003). Incremental compilation of Bayesian networks. In *Proc. 19th Annual Conference on Uncertainty in Artificial Intelligence* 233–240. Morgan Kaufmann, San Francisco.

FRIEDMAN, N., LINIAL, M., NACHMAN, I. and PE'ER, D. (2000). Using Bayesian networks to analyze expression data. *J. Computational Biology* **7** 601–620.

GIUDICI, P. (1996). Learning in graphical Gaussian models. In *Bayesian Statistics 5* (J. M. Bernardo, J. O. Berger, A. P. Dawid and A. F. M. Smith, eds.) 621–628. Oxford Univ. Press, London.

GIUDICI, P. and CASTELO, R. (2003). Improving Markov chain Monte Carlo model search for data mining. *Machine Learning* **50** 127–158.

GIUDICI, P. and GREEN, P. J. (1999). Decomposable graphical Gaussian model determination. *Biometrika* **86** 785–801.

GRONE, R., JOHNSON, C. R., DE SÁ, E. M. and WOLKOWICZ, H. (1984). Positive definite completions of partial Hermitian matrices. *Linear Algebra Appl.* **58** 109–124.

HAMMERSLEY, J. M. and CLIFFORD, P. E. (1971). Markov fields on finite graphs and lattices. Unpublished manuscript.

HECKERMAN, D., CHICKERING, D. M., MEEK, C., ROUNTHWAITE, R. and KADIE, C. M. (2000). Dependency networks for inference, collaborative filtering, and data visualization. *J. Machine Learning Research* **1** 49–75.

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

LAURITZEN, S. L. and SHEEHAN, N. A. (2003). Graphical models for genetic analyses. *Statist. Sci.* **18** 489–514.

MADIGAN, D. and YORK, J. (1995). Bayesian graphical models for discrete data. *Internat. Statist. Rev.* **63** 215–232.

ROVERATO, A. (2002). Hyper-inverse Wishart distribution for non-decomposable graphs and its application to Bayesian inference for Gaussian graphical models. *Scand. J. Statist.* **29** 391–411.

WERMUTH, N. (1976). Model search among multiplicative models. *Biometrics* **32** 253–263.

WEST, M., BLANCHETTE, C., DRESSMAN, H., HUANG, E., ISHIDA, S., SPANG, R., ZUZAN, H., OLSON, JR., J. A., MARKS, J.R. and NEVINS, J. R. (2001). Predicting the clinical status of human breast cancer by using gene expression profiles. *Proc. Natl. Acad. Sci. U.S.A.* **98** 11,462–11,467.

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

WONG, F., CARTER, C. and KOHN, R. (2003). Efficient estimation of covariance selection models. *Biometrika* **90** 809–830.

YU, J., SMITH, V., WANG, P., HARTEMINK, A. and JARVIS, E. (2004). Advances to Bayesian network inference for generating causal networks from observational biological data. *Bioinformatics* **20** 3594–3603.

ZHOU, X., KAO, M. J. and WONG, W. H. (2002). Transitive functional annotation by shortest-path analysis of gene expression data. *Proc. Natl. Acad. Sci. U.S.A.* **99** 12,783–12,788.