Computational Aspects Related to Inference in Gaussian Graphical Models With the G-Wishart Prior *

Alex Lenkoski and Adrian Dobra

Abstract

We describe a comprehensive framework for performing Bayesian inference for Gaussian graphical models based on the G-Wishart prior with a special focus on efficiently including nondecomposable graphs in the model space. We develop a new approximation method to the normalizing constant of a G-Wishart distribution based on the Laplace approximation. We review recent developments in stochastic search algorithms and propose a new method, the mode oriented stochastic search (MOSS), that extends these techniques and proves superior at quickly finding graphical models with high posterior probability. We then develop a novel stochastic search technique for multivariate regression models and conclude with a real-world example from the recent covariance estimation literature. Supplemental materials are available online.

Key Words: Bayesian model averaging; Covariance estimation; Covariance selection; Gaussian graphical models; Multivariate regression; Stochastic search.

1 Introduction

We are concerned with a p-dimensional multivariate normal distribution $N_p(0, \Sigma)$. If the ratio between p and n – the available sample size – becomes large, the sample covariance matrix might not be positive definite, while its eigenvalues might not reflect the eigenstructure of the actual covariance matrix (Yang and Berger, 1994). Dempster (1972) proposed reducing the number of parameters that need to be estimated by setting off-diagonal elements of the precision matrix $K = \Sigma^{-1}$ to zero. A pattern of zero constraints in K is called a covariance selection model or a Gaussian graphical model as it represents a pairwise conditional independence structure (Wermuth, 1976). Among many notable contributions focused on structural learning in Gaussian graphical models we mention the regularization methods of Yuan and Lin (2007), Bickel and Levina (2008), Meinshausen and Bühlman (2006). The simultaneous confidence intervals of Drton and Perlman (2004) solve the underlying multiple comparisons problem (Edwards, 2000). These methods produce a single sparse precision matrix whose structure is used

*Alex Lenkoski is Postdoctoral Research Fellow, Department of Applied Mathematics, Heidelberg University, Heidelberg, Germany, (email: lenkoski@stat.washington.edu). Adrian Dobra is Assistant Professor, Department of Statistics, University of Washington, Seattle, WA 98195 (email: adobra@u.washington.edu).

to estimate Σ through the corresponding maximum likelihood estimates (Dempster, 1972).

By imposing suitable prior distributions for K or Σ, various Bayesian approaches have also been proposed (Leonard and Hsu, 1992; Yang and Berger, 1994; Daniels and Kass, 1999; Barnard et al., 2000; Smith and Kohn, 2002; Liechty et al., 2004; Rajaratnam et al., 2008). Inference can be performed based on the best model, i.e. the model having the highest posterior probability. Alternatively, the uncertainty related to a particular choice of zero constraints in K can be taken into account by Bayesian model averaging (Kass and Raftery, 1995). Estimation of quantities of interest is consequently performed by averaging over all possible models weighted by their posterior probabilities. MCMC techniques are key in this context since they are used to visit the space of $2^{p(p-1)/2}$ possible models (Giudici and Green, 1999; Dellaportas et al., 2003; Wong et al., 2003). As p increases, MCMC methods are likely to be slow to converge due to the sheer size of the search space and might not discover the models with the highest posterior probability. Jones et al. (2005) address this issue by proposing the shotgun stochastic search algorithm that is designed to efficiently move towards regions of high posterior probability in the model space, while Scott and Carvalho (2008) develop the feature inclusion search algorithm for decomposable Gaussian graphical models.

In this paper we develop the mode oriented stochastic search for Gaussian graphical models. Our approach builds on the ideas behind the shotgun stochastic search and aims to identify those models having the highest posterior probability. We also give the Laplace approximation to the normalizing constant of a G-Wishart distribution (Diaconnis and Ylvisaker, 1979; Roverato, 2002; Atay-Kayis and Massam, 2005) associated with nondecomposable prime graphs. The Laplace approximation works well because the mode of a G-Wishart distribution can be efficiently and accurately determined using the iterative proportional scaling algorithm (Speed and Kiiveri, 1986). In our Bayesian framework estimation of K and Σ is performed by sampling from the G-Wishart posterior distribution using the block Gibbs sampler algorithm (Asci and Piccioni, 2007). We show that the combination of our new stochastic search and our algorithm for estimating the marginal likelihood associated with a G-Wishart conjugate prior represents a computationally efficient method for rapidly exploring regions of high posterior probability of Gaussian graphical models. We subsequently extend our methodology to multivariate regression models, a connection that is rarely exploited in the Gaussian graphical models literature.

The structure of the paper is as follows. In Section 2 we introduce the G-Wishart distribution associated with a Gaussian graphical model. In Section 3 we describe the iterative proportional scaling algorithm and the block Gibbs sampler algorithm. In Section 4 we develop the Laplace approximation for the normalizing constant of a G-Wishart distribution and study its applicability. In Section 5 we describe stochastic search methods for Gaussian graphical models including our new algorithm – the mode oriented stochastic search. In Section 6 we discuss multivariate regressions associated with Gaussian graphical models with partitioned variables. We illustrate the performance of our proposed methodology in Section 7 using a dataset taken from a financial call center. We make concluding remarks in Section 8.

2 Graphical models and the G-Wishart distribution

We follow the introduction to Gaussian graphical models presented in Ch. 5 of Lauritzen (1996). Let $G = (V, E)$ with $V = \{1, 2, \ldots, p\}$ and $E \subset \{(i, j) \in V \times V : i < j\}$ be an undirected graph whose vertices are associated with a p-dimensional vector $X = X_V$ that follows a $N_p(0, \Sigma)$ distribution. The nonzero elements of $K = \Sigma^{-1}$ are associated with edges in E. A missing edge in E implies $K_{ij} = 0$ and corresponds with the conditional independence of univariate elements X_i and X_j of X given the remaining elements. The canonical parameter K is constrained to the cone P_G described by those positive definite matrices with entries equal to zero for all $(i, j) \notin E$. We define the set of indices of the free elements of $K \in P_G$ (Roverato, 2002):

$$\mathcal{V} = \{(i, j) : i \leq j \text{ with } i = j \in V \text{ or } (i, j) \in E\}. \tag{1}$$

Roverato (2002) generalizes the hyper inverse Wishart distribution of Dawid and Lauritzen (1993) to arbitrary graphs by deriving the Diaconis and Ylvisaker (1979) conjugate prior for $K \in P_G$. Letac and Massam (2007) as well as Atay-Kayis and Massam (2005) continue this development and call this distribution the G-Wishart. More specifically, the G-Wishart distribution $W_G(\delta, D)$ has density

$$p(K|G) \;=\; \frac{1}{I_G(\delta, D)} (\det K)^{(\delta-2)/2} \exp\left\{-\frac{1}{2}\langle K, D\rangle\right\}, \tag{2}$$

with respect to the Lebesgue measure on P_G. Here $\langle A, B\rangle = \text{tr}(A^T B)$ is the trace inner product. The G-Wishart distribution is a regular exponential family with canonical parameter $K \in P_G$ and canonical statistic $(-D/2)$. Having a functional form consistent with the developments of Diaconis and Ylvisaker (1979), the normalizing constant

$$I_G(\delta, D) \;=\; \int_{K \in P_G} (\det K)^{(\delta-2)/2} \exp\left\{-\frac{1}{2}\langle K, D\rangle\right\} dK, \tag{3}$$

is guaranteed to be finite if $\delta > 2$ and $D^{-1} \in P_G$. The likelihood function corresponding with a random sample $x^{(1:n)} = (x^{(1)}, \ldots, x^{(n)})$ from $N_p(0, \Sigma)$ is proportional to

$$(\det K)^{n/2} \exp\left\{-\frac{1}{2}\langle K, U\rangle\right\},$$

where $U = \sum_{i=1}^{n} x^{(i)} x^{(i)T}$. The posterior distribution of K is proportional to

$$(\det K)^{(\delta+n-2/2)} \exp\left\{-\frac{1}{2}\langle K, D + U\rangle\right\}.$$

Let us define D^* to be the positive definite matrix such that

$$\begin{cases} D^*_{ij} = (D + U)_{ij}, & \text{if } (i, j) \in \mathcal{V}, \\ (D^*)^{-1}_{ij} = 0, & \text{if } (i, j) \notin \mathcal{V}. \end{cases} \tag{4}$$

There exists a unique positive definite matrix D^* that verifies (4) (Dempster, 1972; Knuiman, 1978; Speed and Kiiveri, 1986) and which, therefore, is such that $(D^*)^{-1} \in P_G$. We note that

$$\langle K, D + U \rangle = \langle K, D^* \rangle \ .$$

Without any loss of generality, we take the posterior distribution of K to be $W_G(\delta + n, D^*)$ in order to make sure it is proper.

The marginal likelihood of the data given G is the ratio of the normalizing constants of the G-Wishart posterior and prior:

$$p(x^{(1:n)}|G) \quad = \quad I_G(\delta + n, D^*)/I_G(\delta, D). \tag{5}$$

If G is complete, $W_G(\delta, D)$ reduces to the Wishart distribution $W_p(\delta, D)$, hence its normalizing constant (3) is

$$I_G(\delta, D) \quad = \quad 2^{(\delta+p-1)p/2}\Gamma_p\left\{(\delta + p - 1)/2\right\}(\det\, D)^{-(\delta+p-1)/2}, \tag{6}$$

where $\Gamma_p(a) = \pi^{p(p-1)/4}\prod_{i=0}^{p-1}\Gamma\left(a - \frac{i}{2}\right)$ for $a > (p-1)/2$ (Muirhead, 2005).

Roverato (2002) proves that the G-Wishart distribution can be factorised according to the prime components of G and their separators. Let $P_1, \ldots, P_k$ be a perfect sequence of prime components of G and let $S_2, \ldots, S_k$ be the corresponding separators, where $S_l = \left(\cup_{j=1}^{l-1} P_j\right) \cap P_l$, $l = 2, \ldots, k$. The normalizing constant (3) is equal to (Roverato, 2002):

$$I_G(\delta, D) \quad = \quad \frac{\prod_{j=1}^{k} I_{G_{P_j}}(\delta, D_{P_j})}{\prod_{j=2}^{k} I_{G_{S_j}}(\delta, D_{S_j})}. \tag{7}$$

The subgraph G_{S_j} associated with a separator S_j is required to complete, hence $I_{G_{S_j}}(\delta, D_{S_j})$ is explicitly calculated as in (6). If a subgraph G_{P_j} happens to be complete, the computation of $I_{G_{P_j}}(\delta, D_{P_j})$ can be done in a similar manner. Therefore the primary challenge in computing $I_G(\delta, D)$ relates to the computation of the normalizing constants of the prime components whose subgraphs are not complete. This has been the main reason that many authors have restricted their attention to decomposable graphs, as such challenges are not encountered. However, there has been considerable recent progress in computing normalizing constants of nondecomposable Gaussian graphical models. Dellaportas et al. (2003) and Roverato (2002) develop importance sampling algorithms for computing such normalizing constants, while Atay-Kayis and Massam (2005) propose a simple but efficient Monte Carlo method. In Section 4 we develop the Laplace approximation method with the novel approach of using the iterative proportional scaling algorithm to compute the mode of the integrand in (3). This new method is considerably faster to compute than previously developed approximations, though it tends to be accurate only when computing the posterior normalizing constant, as discussed in Section 4.

3 The iterative proportional scaling algorithm and the block Gibbs sampler

Let G be an arbitrary undirected graph and let $C_1, \ldots, C_k$ be its set of cliques that can be determined, for example, with Algorithm 457 of Bron and Kerboscht (1973). The iterative proportional scaling algorithm and the block Gibbs sampler are cyclic algorithms that generate matrices with values in the cone of incomplete symmetric matrices X such that the submatrices $X_{C_i},\ i = 1, \ldots, k$ are positive definite and such that X can be completed in such a way that the inverse of their completion belongs to P_G.

For any given clique C of G and a given $|C| \times |C|$ positive definite matrix A, we introduce the following operator from P_G into P_G

$$M_{C,A}K = \begin{pmatrix} A^{-1} + K_{C,V\setminus C}(K_{V\setminus C})^{-1}K_{V\setminus C,C} & K_{C,V\setminus C} \\ K_{V\setminus C,C} & K_{V\setminus C} \end{pmatrix}. \tag{8}$$

which is such that $[(M_{C,A}K)^{-1}]_C = A$ (Lauritzen, 1996). Since $K_{V\setminus C}, K_{V\setminus C,C}, K_{C,V\setminus C}$ remain unchanged under this transformation, it follows that $M_{C,A}$ maps P_G into P_G.

3.1 The iterative proportional scaling algorithm

Given a $p \times p$ positive definite matrix L, we need to find the $p \times p$ matrix $K \in P_G$ such that

$$(K^{-1})_{C_j} = L_{C_j},\ \text{for } j = 1, \ldots, k. \tag{9}$$

If $L = U/n$, then (9) is the system whose solution is the maximum likelihood estimate of the precision matrix of a covariance selection model (Lauritzen, 1996). For $L = D + U$, we obtain the system (5) whose solution is the inverse of D^* of the posterior distribution $W_G(\delta + n, D^*)$. If we set $L = D/(\delta - 2)$, we obtain the system whose solution is the mode of the G-Wishart distribution $W_G(\delta, D)$ – see Section 4.

Speed and Kiiveri (1986) proposed the iterative proportional scaling (IPS) algorithm that finds the unique matrix in P_G that satisfies (9). The IPS algorithm proceeds as follows:

Step 1. Start with $K^0 = I_p$, the p-dimensional identity matrix.

Step 2. At iteration $r = 0, 1, \ldots$ do

Step 2A. Set $K^{r+(0/k)} = K^r$.

Step 2B. For each $j = 1, \ldots, k$, set $K^{r+(j/k)} = M_{C_j,L_{C_j}}K^{r+((j-1)/k)}$.

Step 2C. Set $K^{r+1} = K^{r+(k/k)}$.

The sequence $(K^r)_{r\geq 0} \subset P_G$ converges to the solution of (9) (Speed and Kiiveri, 1986). In particular, when G is decomposable with cliques $\{C_1, \ldots, C_k\}$ arranged in a perfect sequence

and separators $\{S_2, \ldots, S_k\}$, the IPS algorithm converges after one iteration to the solution of (9) that is readily available through the following formula (Lauritzen, 1996):

$$\sum_{j=1}^{k} [(L_{C_j})^{-1}]^0 - \sum_{j=2}^{k} [(L_{S_j})^{-1}]^0, \tag{10}$$

where $[A]^0$ is the matrix whose C-submatrix coincides with A and has zero entries everywhere else.

3.2 The block Gibbs sampler

The IPS algorithm can be transformed into a sampling procedure from the G-Wishart distribution by replacing the deterministic updates associated with each clique of G with random updates drawn from appropriate Wishart distributions. This sampling method is called the block Gibbs sampler and was originally discussed in Piccioni (2000). Asci and Piccioni (2007) develop it explicitly by exploiting the theory of exponential families with cuts. Here a cut is a clique of G.

If $K \in P_G$ is distributed $W_G(\delta, D)$ and C is a clique of G then from Corollary 2 of Roverato (2002) we know that $[(K^{-1})_C]^{-1}$ has a Wishart distribution $W_{|C|}(\delta, D_C)$, also written $W_{|C|}(\delta + |C| - 1, (D_C)^{-1})$ in the notation used in Ch. 3 of Muirhead (2005). This chapter also describes a method to sample from a Wishart distribution through the Bartlett decomposition.

The block Gibbs sampler is obtained by replacing Step 2B of the IPS algorithm with

For each $j = 1, \ldots, k$, simulate A from $W_{|C_j|}(\delta, D_{C_j})$ and set $K^{r+(j/k)} = M_{C_j, A^{-1}} K^{r+((j-1)/k)}$.

The other steps remain unchanged. The sequence of matrices $(K^r)_{r \geq r_0}$ generated by the block Gibbs sampler are random samples from $W_G(\delta, D)$ after a suitable burn-in time r_0 (Piccioni, 2000; Asci and Piccioni, 2007). Carvalho et al. (2007) propose another sampler that is based on decomposing G into its sequence of prime components.

4 Computing the marginal likelihood of a graph

We develop the Laplace approximation of Tierney and Kadane (1986) to estimate the normalizing constant $I_G(\delta, D)$ in (3). The indices of the free elements of $K \in P_G$ are given by $\mathcal{V}$ defined in (1). We write

$$I_G(\delta, D) = \int_{K \in P_G} \exp\left(h_{\delta,D}(K)\right) \prod_{(i,j) \in \mathcal{V}} dK_{ij},$$

where

$$h_{\delta,D}(K) = -\frac{1}{2}\left[\mathrm{tr}(K^T D) - (\delta - 2)\log(\det K)\right].$$

The Laplace approximation to $I_G(\delta, D)$ is

$$\widehat{I_G(\delta, D)} = \exp\left(h_{\delta,D}(\widehat{K})\right)(2\pi)^{|\mathcal{V}|/2}[\det H_{\delta,D}(\widehat{K})]^{-1/2},$$

where $\widehat{K} \in P_G$ is the mode of $W_G(\delta, D)$ and $H_{\delta,D}$ is the $|\mathcal{V}| \times |\mathcal{V}|$ Hessian matrix associated with $h_{\delta,D}$.

For $(i, j) \in \mathcal{V}$, the first derivative of $h_{\delta,D}$ is (see, for example, Hartville (1997))

$$\frac{dh_{\delta,D}(K)}{dK_{ij}} = -\frac{1}{2}\text{tr}\left\{\left[D - (\delta - 2)K^{-1}\right]\ (1_{ij})^0\right\},$$

where $(1_{ij})^0$ is a $p \times p$ matrix that has a 1 in the (i, j) and (j, i) positions and zero elsewhere. By setting the first derivatives to zero, i.e. $\frac{dh_{\delta,D}(K)}{dK_{ij}} = 0$, $(i, j) \in \mathcal{V}$, we obtain that the mode $\widehat{K}$ has to satisfy the system of equations (9) for $L = D/(\delta - 2)$. If G is decomposable, $\widehat{K}$ is given by (10) with $L = D/(\delta - 2)$. If G is nondecomposable, $\widehat{K}$ can be efficiently determined using the IPS algorithm.

For $(i, j), (l, m) \in \mathcal{V}$, the $((i, j), (l, m))$ entry of $H_{\delta,D}$ is given by

$$\frac{d^2h_{\delta,D}(K)}{dK_{ij}dK_{lm}} = -\frac{\delta - 2}{2}\text{tr}\left\{K^{-1}(1_{ij})^0K^{-1}(1_{lm})^0\right\}.$$

Kass and Raftery (1995) note that the accuracy of the Laplace approximation depends on the degree to which the density resembles a Gaussian distribution. In order to empirically illustrate the applicability of the Laplace approximation, we used the block Gibbs sampler to draw $10,000$ samples from the G-Wishart distributions $W_{C_5}(\delta, (\delta - 2)K^{-1})$, for $\delta \in \{3, 13, 28\}$. Here C_5 is the five-dimensional cycle graph, and the elements of K are $K_{i,i} = 1$, $K_{i,i-1} = K_{i-1,i} = 0.5$, $K_{1,5} = K_{5,1} = 0.4$ and $K_{i,j} = 0$ otherwise. Figure 1 shows the empirical marginal distributions of K_{12} corresponding with these three G-Wishart distributions. The fact that the distribution of K_{12} is so much more diffuse when $\delta = 3$ as opposed to 13 or 28 implies that, for a G-Wishart distribution $W_G(\delta, D)$, $\widehat{I_G(\delta, D)}$ approximates $I_G(\delta, D)$ better for larger values of δ. The Monte Carlo method of Atay-Kayis and Massam (2005) requires an increased number of iterations to converge for datasets with a larger sample size, hence using it for computing $I_G(\delta_0 + n, D^*)$ is accurate but computationally demanding. This same Monte Carlo method converges quickly when computing the prior normalizing constant $I_G(\delta_0, D_0)$ with $\delta_0 = 3$ (small) and D_0 set to the identity matrix. As such, we evaluate the marginal likelihood (5) of a nondecomposable graph G in an efficient and accurate manner by employing the Laplace approximation for $I_G(\delta_0 + n, D^*)$ and the Monte Carlo method for $I_G(\delta_0, D_0)$.

5 Stochastic search algorithms

We denote by $\mathcal{G}$ a set of competing graphs. We associate with each $G \in \mathcal{G}$ a neighborhood $\text{nbd}(G) \subset \mathcal{G}$. The neighborhoods are defined with respect to the class of graphs considered

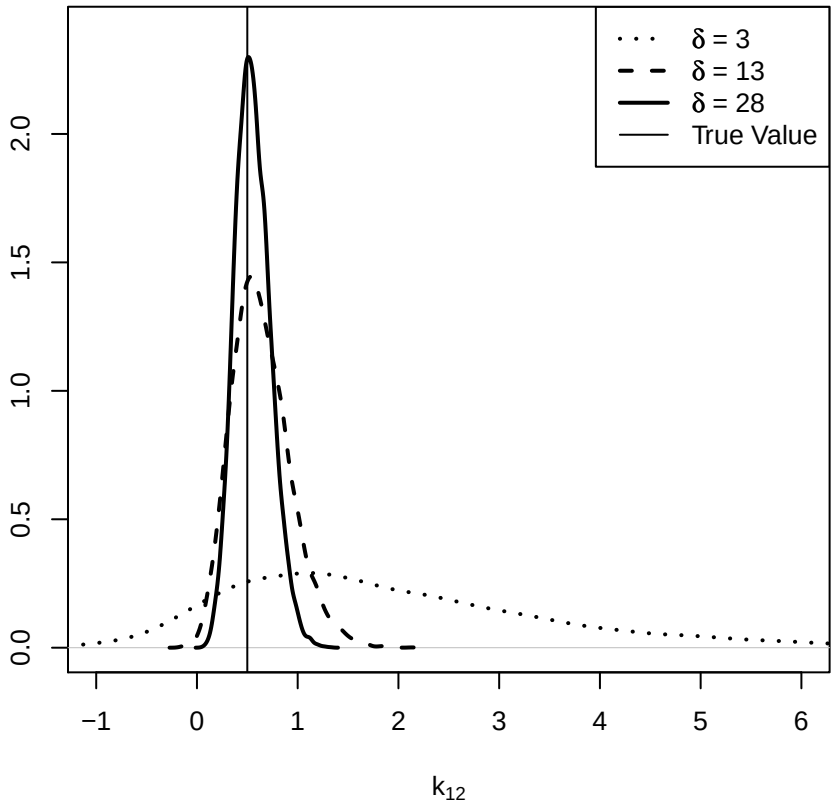


Figure 1: Marginal distributions of K_{12} based on $10,000$ samples from the G-Wishart distributions $W_{C_5}(\delta, (\delta-2)K^{-1})$, for $\delta \in \{3, 13, 28\}$. The vertical line $x = 0.5$ shows the true value of K_{12}.

such that any two graphs are connected by a sequence of graphs. Two consecutive graphs in this sequence are neighbors of each other. If $\mathcal{G}$ represents the 2^r graphs with p vertices where $r = p(p-1)/2$, nbd(G) contains those graphs that are obtained by adding or deleting one edge from G. If $\mathcal{G}$ is restricted to decomposable graphs, nbd(G) contains only the decomposable graphs that are one edge away from G.

The absence of any prior beliefs related to the structure and the complexity of the graphs can be modeled with a uniform prior $p(G) \propto 1$ over $\mathcal{G}$. This represents our choice of prior for the real-world example from Section 7, but alternatives to the uniform prior have also been developed. For instance, Wong et al. (2003) propose a uniform prior on graphs with the same number of edges, while Jones et al. (2005) discuss priors that encourage sparse graphs by penalizing for the inclusion of additional edges:

$$p(G) \propto \beta^k (1-\beta)^{r-k},$$

where k is the number of edges of G and $\beta \in (0,1)$ is the probability of the inclusion of an edge in G. A uniform prior for β leads to a prior with desirable multiple testing correction properties (Scott and Berger, 2006; Scott and Carvalho, 2008):

$$p(G) \propto \left[(r+1) \binom{r}{k} \right]^{-1}. \tag{11}$$

Our goal is ultimately to form a model averaged estimate of the precision matrix K based on using the posterior probabilities

$$p(G|x^{(1:n)}) \propto p(x^{(1:n)}|G)p(G) = p^*(G|x^{(1:n)})$$

of each graph considered. Henceforth we refer to $p^*(G|x^{(1:n)})$ as the graph's score. Clearly, even in relatively low dimensions, the entire space of graphical models cannot be enumerated and stochastic search methods have long been recognized as integral to forming high-dimensional model averaged estimates. Early work in this area focused on MCMC techniques, which initially appear desirable as they theoretically allow one to sample from the posterior model probability space given suitable burn in. Along these lines, Giudici and Green (1999) and Wong et al. (2003) develop reversible jump MCMC samplers (Green, 1995) for decomposable and arbitrary Gaussian graphical models, respectively. However, as p gets larger, MCMC methods in which the chains are run over the product space of $\mathcal{G}$ and other model parameters require a considerable number of iterations to achieve convergence due to the high dimensionality of the state space. Furthermore, as discussed in Section 4, by using the G-Wishart distribution graphical models can be compared directly without requiring simultaneous sampling from the parameter space. The Markov chain model composition algorithm (MC3) of Madigan and York (1995) can be used in conjunction with this property to construct an irreducible chain only on $\mathcal{G}$. If the chain is currently in state G, a candidate graph G' is randomly drawn uniformly from nbd(G). The chain moves to G' with probability

$$\min \left\{ 1, \frac{p(G'|x^{(1:n)})/|\text{nbd}(G')|}{p(G|x^{(1:n)})/|\text{nbd}(G)|} \right\}, \tag{12}$$

The acceptance probability (12) simplifies to $\min\{1, p(G'|x^{(1:n)})/p(G|x^{(1:n)})\}$ in the space of arbitrary graphs since $|\text{nbd}(G)| = r$ for all $G \in \mathcal{G}$. This is not true for the smaller space of decomposable graphs since the number of decomposable graphs that differ with one edge from a given graph is not constant.

A typical application of the MC3 algorithm then estimates model probabilities by the frequency of times the chain visited a given graph G. However, in the framework we consider here MC3 is, in reality, only a device to explore the graph space, since the posterior probability of each graph G that has been visited is readily available up to the normalizing constant

$$\left[\sum_{G\in\mathcal{G}} p(G|x^{(1:n)})\right]^{-1},$$

hence there is no explicit need to repeatedly revisit G to determine $p(G|x^{(1:n)})$ within the collection of graphs visited. Recognizing this fact, Jones et al. (2005), consider an alternative algorithm, the shotgun stochastic search (SSS) which extends the concept of the MC3 algorithm. In SSS, there is still a current state G, but instead of randomly sampling and scoring one neighbor of G, all the neighbors of G are scored and these scores are retained. A new state G' is then sampled from these neighbors, with probability proportional to the scores within this neighborhood. While this technique no longer retains the MCMC properties of the MC3 algorithm, Jones et al. (2005) show that scoring all graphs in the neighborhood of the current G dramatically increases the speed with which high probability graphs are discovered.

While the SSS algorithm reveals the benefit of scoring all neighbors of a current graph, it shares a potential shortcoming with the MC3 algorithm. Both algorithms have the property that the graph at a given iteration is selected from the neighbors of the graph in the previous iteration. This implies that if the search has moved along a path to low probability graphs, this path will have to be retraced in order to move back to high probability graphs. Therefore, it is sensible to assume that higher posterior probability graphs can be reached faster by exploring the neighborhood of a graph that has been identified at any previous iteration, in accordance to each graph's posterior probability. This concept is discussed by Berger and Molina (2005). Finally, Scott and Carvalho (2008) discuss the concept of "global moves", in which a graph in the current state is proposed that has not necessarily been visited previously in the search. This enables the search to move quickly to alternative areas of the graph space and reduces dependence on the particular path the search has taken up to the current iteration.

We combine the concepts of assessing a graph's entire neighborhood with the model proposal concepts of Berger and Molina (2005) into a novel stochastic search algorithm called the mode oriented stochastic search (MOSS). Our algorithm proceeds by retaining a large list of graphs $\mathcal{L} \subset \mathcal{G}$ that is updated at each iteration. We define the probability of a graph $G \in \mathcal{L}$ with respect to $\mathcal{L}$ as $p_{\mathcal{L}}(G) = \left[\sum_{G'\in\mathcal{L}} p^*(G'|x^{(1:n)})\right]^{-1} p^*(G|x^{(1:n)})$. The probability of an edge $(i,j) \in V \times V$ with respect to $\mathcal{L}$ is the sum of all $p_{\mathcal{L}}(G)$ such that (i,j) appears in G. The median graph $G_{\mathcal{L}}$ with respect to $\mathcal{L}$ is given by all the edges having a probability with respect to $\mathcal{L}$ above 0.5. By continually updating the median graph associated with the current state of the

search and adding it as a candidate for evaluation, we also incorporate the concept of allowing for global moves in the model space. A graph is called explored if all its neighbors have been visited. We keep track of the graphs currently in $\mathcal{L}$ that have been explored. MOSS has two parameters: m – the maximum number of graphs to be recorded in $\mathcal{L}$ and k – the maximum number of iterations to be performed. Our algorithm proceeds as follows:

`procedure MOSS(`m`,`k`)`
► Initialize the starting list of graphs $\mathcal{L} \subset \mathcal{G}$. For each graph $G \in \mathcal{L}$, calculate and record its score $p^*(G|x^{(1:n)})$ and mark it as unexplored.
For $l = 1, \ldots, k$ do:

- Let $\mathcal{L}_U$ be the unexplored graphs in $\mathcal{L}$. If $\mathcal{L}_U = \emptyset$, STOP.
- Sample a graph $G \in \mathcal{L}_U$ according to probabilities $p_{\mathcal{L}_U}(G)$. Mark G as explored.
- For each $G' \in \text{nbd}(G)$ do the following: if $G' \notin \mathcal{L}$, include G' in $\mathcal{L}$, evaluate and record its score $p^*(G'|x^{(1:n)})$, and mark G' as unexplored.
- Delete from $\mathcal{L}$ and $\mathcal{L}_U$ the lowest posterior probability graphs such that $\mathcal{L}$ contains a maximum of m graphs.
- Determine the median graph $G_{\mathcal{L}}$. If $G_{\mathcal{L}} \notin \mathcal{L}$, calculate its score $p^*(G_{\mathcal{L}}|x^{(1:n)})$, mark it as unexplored and add it to $\mathcal{L}$.

`end procedure`

A feature of MOSS is that it can end before completing k iterations if $\mathcal{L}$ does not contain any unexplored graphs. Instead of stopping MOSS, we could choose to mark all the models in $\mathcal{L}$ as unexplored and continue the procedure until the specified number of iterations has been reached. MOSS can be started with a list $\mathcal{L}$ of randomly generated graphs as we do in Section 7. Alternatively, MOSS could be started from a list of graphs determined from other covariance selection methods.

We recommend running several separate instances of MOSS to make sure that it reaches the same highest score graphs. If this does not happen, MOSS should be run for an additional number of iterations.

6 Multivariate regressions

We consider the case when the random vector $X \sim N_p(0, \Sigma)$ is partitioned into a set of response variables X_R and a set of explanatory variables $X_{V\setminus R}$. It follows that the conditional distribution of X_R given $X_{V\setminus R} = x_{V\setminus R}$ is $N_{|R|}(\Gamma_{R|V\setminus R} x_{V\setminus R}, \Sigma_{R|V\setminus R})$, where $\Gamma_{R|V\setminus R} = \Sigma_{R,V\setminus R}(\Sigma_{V\setminus R})^{-1}$ and $\Sigma_{R|V\setminus R} = \Sigma_R - \Sigma_{R,V\setminus R}(\Sigma_{V\setminus R})^{-1}\Sigma_{V\setminus R,R}$. The marginal distribution of $X_{V\setminus R}$ is $N_{|V\setminus R|}(0, \Sigma_{V\setminus R})$. The regression parameters $(\Gamma_{R|V\setminus R}, \Sigma_{R|V\setminus R})$ are independent of $\Sigma_{V\setminus R}$ (Muirhead, 2005). As such, inference for the conditional $p(X_R|X_{V\setminus R})$ can be performed independently from inference on the marginal $p(X_{V\setminus R})$.

The zero constraints $K_{ij} = 0$ for the elements of the precision matrix $K = \Sigma^{-1}$ are classified as (a) conditional independence of two response variables given the rest, i.e. $i, j \in R$;

(b) conditional independence of a response variable and an explanatory variable, i.e. $i \in R$ and $j \in V \setminus R$; and (c) conditional independence of two explanatory variables, i.e. $i, j \in V \setminus R$. Note that condition (b) is equivalent with the absence of X_j from the regression of X_i given $X_{V\setminus R}$. The zero constraints of type (a) and (b) are associated with the conditional $p(X_R|X_{V\setminus R})$, while the constraints of type (c) involve the marginal $p(X_{V\setminus R})$. Define the set of graphs $\mathcal{G}^{[V\setminus R]}$ with vertices V such that their $V \setminus R$ subgraph $G_{V\setminus R}$ is complete. A graph $G \in \mathcal{G}^{[V\setminus R]}$ embeds only constraints of type (a) and (b) for K, hence it is representative of the conditional independence relationships in $p(X_R|X_{V\setminus R})$.

As before, we assume a G-Wishart prior $W_G(\delta_0, D_0)$ for $K \in P_G$. The induced prior for $\Sigma = K^{-1}$ is hyper inverse Wishart $\text{HIW}_G(\delta_0, D_0)$ which is strong hyper Markov – see Corollary 1 of Roverato (2002). From Proposition 5.6 of Dawid and Lauritzen (1993) it follows that the marginal data-distribution of X is Markov. Since G is collapsible onto $V \setminus R$ and $(R, V \setminus R)$ is a decomposition of G, it follows that the marginal likelihood of the regression of X_R on $X_{V\setminus R}$ with constraints induced by G is the ratio of the marginal likelihoods of G and $G_{V\setminus R}$:

$$p\left(x_R^{(1:n)}|x_{V\setminus R}^{(1:n)}, G\right) = p\left(x^{(1:n)}|G\right) / p\left(x_{V\setminus R}^{(1:n)}|G_{V\setminus R}\right). \tag{13}$$

Here $x_A^{(1:n)}$, $A \subset V$, represents the subset of the data $x^{(1:n)}$ corresponding with X_A. Since $G_{V\setminus R}$ is complete, $p\left(x_{V\setminus R}^{(1:n)}|G_{V\setminus R}\right)$ is the ratio of the normalizing constants of the Wishart distributions $W_{|G\setminus R|}(\delta_0+n, D^*_{V\setminus R})$ and $W_{|G\setminus R|}(\delta_0+n, D_{V\setminus R})$ – see (6). This means that $p\left(x_{V\setminus R}^{(1:n)}|G_{V\setminus R}\right)$ is constant for all $G \in \mathcal{G}^{[V\setminus R]}$.

Only a subset of explanatory variables might be connected with the response variables in any given graph $G \in \mathcal{G}^{[V\setminus R]}$. Let $\text{bd}_G(R) \subset V \setminus R$ be the boundary of the response variables in G, that is, $j \in \text{bd}_G(R)$ if there exists an $i \in R$ such that (i, j) is an edge in G. If there exists at least one explanatory variable that is not linked with any response variable (i.e., $V \setminus (R \cup \text{bd}_G(R)) \neq \emptyset$) then $V \setminus R$ is a clique in G. Therefore X_R is independent of $X_{V\setminus(R\cup\text{bd}_G(R))}$ given $X_{\text{bd}_G(R)}$, thus $p(X_R|X_{V\setminus R}) = p(X_R|X_{\text{bd}_G(R)})$. Since the marginal data-distribution of X is Markov, (13) reduces to:

$$p\left(x_R^{(1:n)}|x_{V\setminus R}^{(1:n)}, G\right) = p\left(x_R^{(1:n)}|x_{\text{bd}_G(R)}^{(1:n)}, G_{R\cup\text{bd}_G(R)}\right) = \frac{p\left(x_{R\cup\text{bd}_G(R)}^{(1:n)}|G_{R\cup\text{bd}_G(R)}\right)}{p\left(x_{\text{bd}_G(R)}^{(1:n)}|G_{\text{bd}_G(R)}\right)}. \tag{14}$$

This property is especially important if $|R|$ is much smaller than $|V \setminus R|$, hence it is likely that only a few explanatory variables will be connected with the responses.

The stochastic search methods from Section 5 can be employed to determine high posterior probability graphs in $\mathcal{G}^{[V\setminus R]}$. The size of this search space is $2^{|R|(|R|-1)/2} + 2^{|R|\cdot|V\setminus R|}$ which is significantly smaller than $2^{p(p-1)/2}$. The admissible set of neighbors of a graph needs to be modified so that no two vertices in $V \setminus R$ become disconnected during the search. For the purpose of identifying graphs with a more parsimonious structure, we could restrict the search to the subset $\mathcal{G}_q^{[V\setminus R]}$ of $\mathcal{G}^{[V\setminus R]}$ containing graphs with the maximum size of their cliques at most

q, where q is a small integer. Such graphs have a better chance of inducing good predictive models for X_R. The size of the boundary of X_R will be at most $q-1$ and hence no more than $q-1$ explanatory variables can enter each multivariate regression model. The graphs in $\mathcal{G}_q^{[V\setminus R]}$ are connected not only by the addition and removal of edges, but also by substituting edges and explanatory variables. Let $G=(V,E)\in\mathcal{G}_q^{[V\setminus R]}$ and denote by $\mathcal{E}=\{(i,j)\in(V\times V)\setminus[(V\setminus R)\times(V\setminus R)]:i<j\}$ the subset of edges that can be changed in G. These are edges that connect two response variables or a response variable with an explanatory variable. The neighbors of G are obtained by: (i) including in E an edge from $\mathcal{E}\setminus E$, (ii) deleting an edge from E, (iii) replacing an edge in E with any other edge in $\mathcal{E}\setminus E$ and (iv) replacing an explanatory variable currently in $\text{bd}_G(R)$ with any another explanatory variable currently in $V\setminus(R\cup\text{bd}_G(R))$.

7 Real-world example: the call center data

We analyze a large scale dataset originally described in Shen and Huang (2005) and further studied in Huang et al. (2006), Bickel and Levina (2008) and Rajaratnam et al. (2008). The number of calls n_{ij} for $i=1,\ldots,239$ days and $j=1,\ldots,102$ ten-minute daily time intervals were recorded in 2002 from the call center of a major financial institution. A transformation $x_{ij}=(n_{ij}+0.25)^{1/2}$ was subsequently employed to assure the normality assumption. The call center data have been used to predict the volume of calls in the second half of the day given the volume of calls from the first half of the day based on an estimate of Σ, the covariance of calls in a given day. The dataset is then divided into a training set (first 205 days) and a test set (the remaining 34 days). The training data are then used to learn the dependency structure among the 102 variables associated with each time interval and to further estimate the covariance matrix Σ. Previous assessments of these data have shown them to be an interesting case study in the benefits of using sparse models to perform prediction. In particular, Rajaratnam et al. (2008) indicate that this sparsity effect is most pertinent for the "response variables" – the 51 variables for which values will be predicted.

Since the modeling problem presents a clear separation between response and predictor variables, we restrict our attention to the multivariate regression graphs outlined in Section 6 with $R=\{52,\ldots,102\}$. In order to assess the degree to which sparsity affects prediction error we conduct searches in which the maximum clique size q allowed in the graph space is set to 3, 4 and 5. We are also interested in comparing the relative performance of the MOSS algorithm to other stochastic search methods, in particular the MC3 and SSS algorithms. To do this we first run five instances of the MOSS algorithm from different random starting points, with $m=1000$, and $k=2500$ and recorded the total number of graphs evaluated, while also inspecting the chains to ensure that they had converged. Figure 2 displays the log of the top graph score by log iteration for each chain run when $q=5$, showing satisfactory agreement between chains regarding the top graph. The corresponding figures for $q=3$ and $q=4$ look similar. We then ran five instances of both the MC3 and SSS algorithms, such that each instance evaluates the same total number of graphs as in the MOSS search and recorded the top 1000

Figure 2: Score of the top graph in each of five different instances of the MOSS algorithm for $q = 5$. By 2500 iterations of the MOSS algorithm, all five instances agree on the top graph.

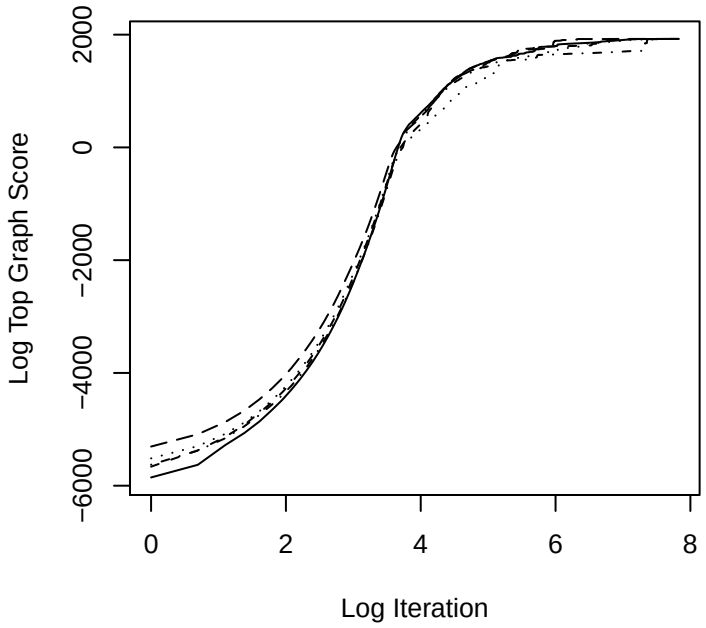


Table 1: Mean absolute prediction error for predictions of the call center data by stochastic search method and maximum clique size q. For reference, the mean absolute prediction error associated with the sample covariance matrix is 1.47.

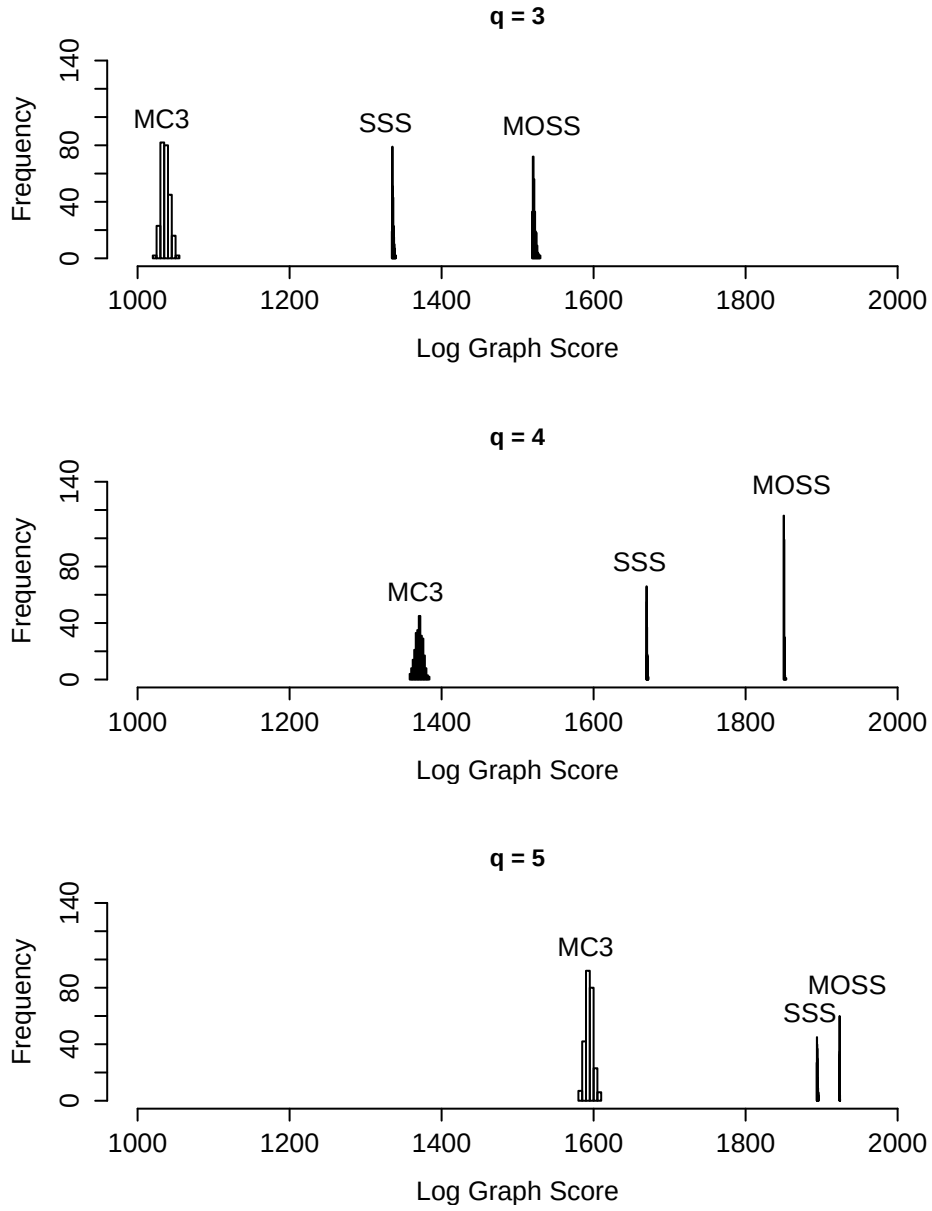
	$q = 3$	$q = 4$	$q = 5$
MOSS	1.043	1.211	1.312
SSS	1.073	1.214	1.338
MC3	1.129	1.241	1.375

graphs found by each of these methods. For reference, the MOSS searches took 16, 17 and 19 hours when q was set to $3, 4$ and 5 respectively, when run on a Mac Pro with 8 GB ram and a 2GHz processor. Computation time was similar for the MC3 and SSS searches at equivalent levels of q, as the same number of graphs were evaluated and there is little difference in the computational overhead of the bookkeeping associated with each algorithm.

Figure 3 shows the distribution of the log score for each of the top graphs, by search type and the value at which q is set. From Figure 3 we see a clear ordering of the graph scores by algorithm. For each value of q, MC3 returns the lowest scoring graphs, SSS finds somewhat higher scoring graphs and MOSS finds the highest scoring graphs.

Table 1 shows that higher posterior probability graphs actually translate into improved predictive performance. We see that the graphs returned by MOSS consistently give lower predictive error compared to the other stochastic search algorithms. Table 1 also shows that the prediction error declines as the level of sparsity increases, which is in line with the results of Rajaratnam et al. (2008). Figure 4 displays the mean absolute prediction error for each time point at varying levels of q, using the graphs identified by MOSS. For comparison, the mean absolute error associated with the full covariance matrix is also displayed. We see that in the first few time periods, denser graphs perform well. However, in the middle time periods, there is a significant improvement in the use of sparse graphs with $q \leq 3$.

Figure 3: Distribution of the log graph scores for the top 1000 graphs returned by MOSS, SSS and MC3 when run for an equal number of graph evaluations at different levels of q, the maximum clique size allowed.



8 Conclusions

The implicit claim that sampling from the posterior distribution on the space of graphs should reveal the highest posterior probability graphs in a computationally efficient manner holds only if the cumulative posterior probability of these graphs is not close to zero. For high-dimensional applications this condition is no longer true: the number of possible models is extremely large while the sample size tends to be small. As such, the uncertainty in selecting graphs is very high and methods that sample from the posterior distribution on the space of graphs spend most of their time around graphs having low posterior probability because their cumulative posterior probability approaches one. The implication is that determining the highest posterior probability graphs and sampling from the corresponding posterior distribution become two separate problems that need distinct solutions.

MOSS is designed to efficiently move towards high posterior probability graphs without attempting to sample. We have shown empirically that it outperforms MC3 and SSS. Comparisons with the feature-inclusion algorithm of Scott and Carvalho (2008) are obstructed by its implementation that is restricted to decomposable graphs and a much different choice of priors.

We have also taken on the challenge of efficiently including nondecomposable graphs in the search space. The Laplace approximation has proved crucial to this end. While the Laplace approximation has been widely used for over two decades, the particular approximation derived in this paper for Gaussian graphical models is novel, in that it incorporates the IPS algorithm to quickly find the mode of the integrand.

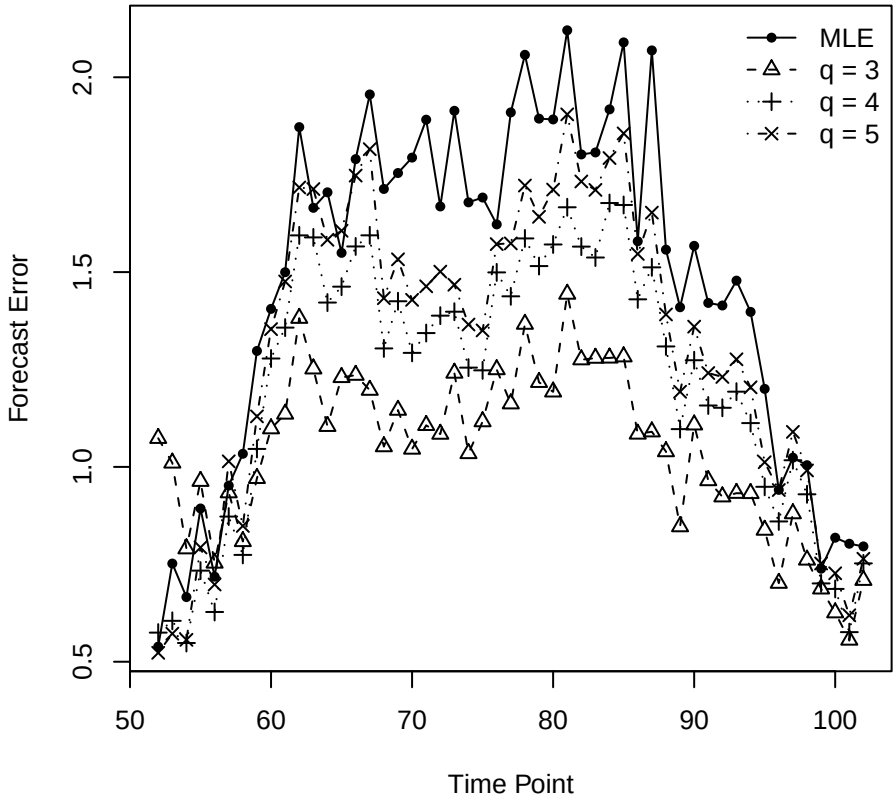
The computational challenges involved in this study were considerable. In particular, we have found there to be a relative dearth of fast clique decomposition algorithms for sparse graphs, which are required both to compute the Laplace approximation and run the block Gibbs sampler. The development of an algorithm that can quickly determine the clique decomposition of a given graph based on the cliques of one of its neighbors would be extremely helpful to statisticians working with general graphical models.

Finally, we have combined the concepts of graphical models and multivariate regressions into a consistent framework, which we believe dramatically increases the number of potential applications of this methodology. We have shown that in the call center data, a multivariate regressions framework reduced search burden and allowed for a sparse search that ultimately yields low predictive error.

Supplemental Material

Computer Code and Data: Supplemental materials for this article are contained in a single tar archive and can be obtained in a single download. This archive contains the call center data (in a comma separated text file) as well as the C++ source code to run the algorithms described in this article. A detailed description of the files contained in this archive is contained in a README.txt file enclosed in the archive. (lenkoski-dobra.tar.gz, GNU zipped tar file)

Figure 4: Mean absolute prediction error by time period using the sample covariance matrix ("MLE") and the model averaged estimates returned by MOSS with different levels of q, the maximum allowed clique size.



Acknowledgments

The work of Alex Lenkoski and Adrian Dobra was supported in part by a seed grant from the Center of Statistics and the Social Sciences, University of Washington. The authors are grateful to Hélène Massam for helpful discussions on an earlier version of the manuscript. The authors thank Jianhua Huang for sharing the Call Center data.

References

Asci, C. and Piccioni, M. (2007). "Functionally compatible local characteristics for the local specification of priors in graphical models." *Scand. J. Statist.*, 34, 829–40.

Atay-Kayis, A. and Massam, H. (2005). "A Monte Carlo method for computing the marginal likelihood in nondecomposable Gaussian graphical models." *Biometrika*, 92, 317–35.

Barnard, J., McCulloch, R., and Meng, X. (2000). "Modeling covariance matrices in terms of standard deviations and correlations, with application to shrinkage." *Statist. Sinica*, 10, 1281–311.

Berger, J. O. and Molina, G. (2005). "Posterior model probabilities via path-based pairwise priors." *Statistica Neerlandica*, 59, 3–15.

Bickel, P. J. and Levina, E. (2008). "Regularized estimation of large covariance matrices." *Ann. Statist.*, 36, 199–227.

Bron, C. and Kerboscht, J. (1973). "Algorithm 457: finding all cliques of an undirected graph." *Communications of the ACM archive*, 16, 575–577.

Carvalho, C. M., Massam, H., and West, M. (2007). "Simulation of hyper-inverse Wishart distributions in graphical models." *Biometrika*, 7, 269–81.

Daniels, M. and Kass, R. (1999). "Nonconjugate Bayesian estimation of covariance matrices." *J. Am. Statist. Assoc.*, 94, 1254–63.

Dawid, A. P. and Lauritzen, S. L. (1993). "Hyper Markov laws in the statistical analysis of decomposable graphical models." *Ann. Statist.*, 21, 1272–317.

Dellaportas, P., Giudici, P., and Roberts, G. (2003). "Bayesian inference for nondecomposable graphical Gaussian models." *Sankhya*, 65, 43–55.

Dempster, A. P. (1972). "Covariance selection." *Biometrics*, 28, 157–75.

Diaconnis, P. and Ylvisaker, D. (1979). "Conjugate priors for exponential families." *Ann. Statist.*, 7, 269–81.

Drton, M. and Perlman, M. D. (2004). "Model selection for Gaussian concentration graphs." *Biometrika*, 91, 591–602.

Edwards, D. M. (2000). *Introduction to Graphical Modelling*. New York: Springer.

Giudici, P. and Green, P. J. (1999). "Decomposable graphical Gaussian model determination." *Biometrika*, 86, 785–801.

Green, P. J. (1995). "Reversible jump Markov chain Monte Carlo computation and Bayesian model determination." *Biometrika*, 82, 711–32.

Harville, D. A. (1997). *Matrix Algebra From a Statistician's Perspective*. Springer Science.

Huang, J., Liu, N., Pourahmadi, M., and Liu, L. (2006). "Covariance matrix selection and estimation via penalized normal likelihood." *Biometrika*, 93, 85–98.

Jones, B., Carvalho, C., Dobra, A., Hans, C., Carter, C., and West, M. (2005). "Experiments in stochastic computation for high-dimensional graphical models." *Statist. Sci.*, 20, 388–400.

Kass, R. and Raftery, A. E. (1995). "Bayes factors." *J. Am. Statist. Assoc.*, 90, 773–95.

Knuiman, M. (1978). "Covariance Selection." *Suppl. Adv. Appl. Prob*, 10, 123–130.

Lauritzen, S. L. (1996). *Graphical Models*. Oxford University Press.

Leonard, T. and Hsu, J. S. J. (1992). "Bayesian inference for a covariance matrix." *Ann. Statist.*, 20, 1669–96.

Letac, G. and Massam, H. (2007). "Wishart distributions for decomposable graphs." *Ann. Statist.*, 35, 1278–323.

Liechty, J. C., Liechty, M. W., and Müller, P. (2004). "Bayesian correlation estimation." *Biometrika*, 91, 1–14.

Madigan, D. and York, J. (1995). "Bayesian Graphical Models for Discrete Data." *International Statistical Review*, 63, 215–232.

Meinshausen, N. and Bühlmann, P. (2006). "High-dimensional graphs with the Lasso." *Ann. Statist.*, 34, 1436–62.

Muirhead, R. J. (2005). *Aspects of Multivariate Statistical Theory*. John Wiley & Sons.

Piccioni, M. (2000). "Independence structure of natural conjugate densities to exponential families and the Gibbs Sampler." *Scand. J. Statist.*, 27, 111–27.

Rajaratnam, B., Massam, H., and Carvalho, C. M. (2008). "Flexible covariance estimation in graphical Gaussian models." *Ann. Statist.*, 36, 2818–2849.

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.

Scott, J. G. and Berger, J. O. (2006). "An exploration of aspects of Bayesian multiple testing." *J. Statist. Plan. Infer.*, 136, 2144–2162.

Scott, J. G. and Carvalho, C. M. (2008). "Feature-inclusion stochastic search for Gaussian graphical models." *J. Comput. Graph. Statist.*, 17, 790–808.

Shen, H. and Huang, J. Z. (2005). "Analysis of call center arrival data using singular value decomposition." *Applied Stochastic Models in Business and Industry*, 21, 251–263.

Smith, M. and Kohn, R. (2002). "Bayesian parsimonious covariance matrix estimation for longitudinal data." *J. Am. Statist. Assoc.*, 87, 1141–53.

Speed, T. P. and Kiiveri, H. T. (1986). "Gaussian Markov distributions over finite graphs." *Ann. Statist.*, 14, 138–150.

Tierney, L. and Kadane, J. (1986). "Accurate Approximations for Posterior Moments and Marginal Densities." *J. Amer. Statist. Assoc.*, 81, 82–86.

Wermuth, N. (1976). "Analogies between multiplicative models in contingency tables and covariance selection." *Biometrics*, 32, 95–108.

Wong, F., Carter, C. K., and Kohn, R. (2003). "Efficient estimation of covariance selection models." *Biometrika*, 90, 809–30.

Yang, R. and Berger, J. O. (1994). "Estimation of a covariance matrix using the reference prior." *Ann. Statist.*, 22, 1195–211.

Yuan, M. and Lin, Y. (2007). "Model selection and estimation in the Gaussian graphical model." *Biometrika*, 94, 19–35.