Bayesian Analysis (2014) **9**, Number 3, pp. 659–684

The Performance of Covariance Selection Methods That Consider Decomposable Models Only

A. Marie Fitch [*] and M. Beatrix Jones [†] Hélène Massam [‡]

Abstract. We consider the behavior of Bayesian procedures that perform model selection for decomposable Gaussian graphical models when the true model is in fact non-decomposable. We examine the asymptotic behavior of the posterior when models are misspecified in this way, and find that the posterior will converge to graphical structures that are minimal triangulations of the true structure. The marginal log likelihood ratio comparing different minimal triangulations is stochastically bounded, and appears to remain data dependent regardless of the sample size. The covariance matrices corresponding to the different minimal triangulations are essentially equivalent, so model averaging is of minimal benefit. Using simulated data sets and a particular high performing Bayesian method for fitting decomposable models, feature inclusion stochastic search, we illustrate that these predictions are borne out in practice. Finally, a comparison is made to penalized likelihood methods for graphical models, which make no decomposability restriction. Despite its inability to fit the true model, feature inclusion stochastic search produces models that are competitive or superior to the penalized likelihood methods, especially at higher dimensions.

Keywords: undirected Gaussian graphical models, covariance selection, feature inclusion stochastic search, decomposable, non-decomposable, graphical lasso, asymptotic behavior

1 Introduction

Gaussian graphical models (Dempster 1972; Lauritzen 1996; Whittaker 2008) are a powerful tool for both exploring the partial independence structure of multivariate data and for regularization of the covariance matrix. Let $\mathcal{G} = (V, E)$ be an undirected graph where $V = \{1, \dots, v\}$ denotes the set of vertices and E the set of edges. A v-dimensional model for $X = (X_j,\ j \in V)$ is said to be Markov with respect to $\mathcal{G}$ if X_i is independent of X_j given $X_{V\setminus\{i,j\}}$ whenever the edge (i, j) does not belong to E. A graphical Gaussian model is a v-dimensional centered Gaussian model that is Markov with respect to $\mathcal{G}$. This implies the inverse covariance matrix Ω will have zero entries corresponding to i, j pairs where the graph has no edge. In high dimensional problems, especially when the sample size n is similar to, or less than, the number of variables v, regularization of the covariance matrix leads to improved estimation. This regularization can be achieved via covariance selection to achieve a sparse inverse

[*] University of Auckland m.fitch@auckland.ac.nz

[†] Massey University m.b.jones@massey.ac.nz

[†] York University massamh@mathstat.yorku.ca

DOI:10.1214/14-BA874

covariance matrix. Whether we are seeking to understand the dependence structure of the data, or to provide a better covariance estimate by reducing the number of parameters, we require models that distinguish relevant edges from irrelevant ones. This introduces computational challenges, particularly in high-dimensions.

The motivation for this article arises from consideration of the implications of restricting model selection to decomposable models in cases where the true model is non-decomposable. Bayesian models frequently restrict consideration to decomposable models for computational convenience (Scott and Carvalho 2008; Armstrong et al. 2009; Jones et al. 2005). Relatively few authors have studied Bayesian methods without restricting the class of models (Dellaportas et al. 2003; Wong et al. 2003; Moghaddam et al. 2009; Dobra et al. 2011), and only Moghaddam et al. (2009) present a method that appears to be scalable to high dimensions.

We begin by examining the asymptotic behaviour of marginal likelihood ratios for graph structures differing by one edge. We demonstrate that models that include all true edges will be favored over those that don't, and that among graphs that contain all the true edges, those that contain fewer superfluous edges will be favored. Among graphs with all true edges and equal numbers of unnecessary edges, the log of the marginal likelihood ratio comparing the models is stochastically bounded. Simulations suggest the preferred minimal triangulation is data dependent, even at very large sample size. Because decomposable approximations that contain all the true edges of a non-decomposable model will necessarily have superfluous edges as well, this final attribute means the behavior of edge inclusion probabilities for the superfluous edges is unpredictable, and they may not be readily identified as superfluous.

Because of this idiosyncratic behavior, we were also interested in how the performance of decomposable Bayesian methods compares to other computationally tractable methods for inverse covariance estimation that do not impose a decomposability restriction. We selected feature-inclusion stochastic search (FINCS, Scott and Carvalho 2008) as a representative decomposable restricted method. We make a comparison with lasso methods recently developed explicitly for the covariance selection problem: graphical lasso (Friedman et al. 2008b) and adaptive graphical lasso (Fan et al. 2009). Our results bear out the conclusion in Fan et al. (2009) that adaptive lasso is an improvement over graphical lasso by most measures. In our comparisons between FINCS and adaptive graphical lasso, the FINCS top model was always sparser and typically had smaller Kullback-Leibler divergence from the true model and better predictive performance.

The rest of this article is organized as follows. In Section 2 we review the properties of decomposable graphs and prove the asymptotic marginal likelihood ratio results. In Section 3, we detail the algorithms used for the two approaches to model selection and their comparison. Section 4 presents simulation studies and a real data example, with the discussion in Section 5.

2 Asymptotic behavior of marginal likelihood ratios

2.1 Preliminaries

Without loss of generality we assume that our data follows a centered multivariate v-dimensional Gaussian distribution denoted $N_v(0, \Sigma)$ where Σ is the covariance matrix. Let $\mathcal{G} = (V, E)$ be an undirected graph with vertex set V and edge set E. Let P_G be the set of positive definite matrices Ω with entries $\Omega_{ij} = 0$ whenever $(i, j) \notin E$. The graphical Gaussian model that is Markov with respect to $\mathcal{G}$ is the set

$$\{N_v(0, \Sigma) : \ \Omega = \Sigma^{-1} \in P_G\}.$$

If the graph $\mathcal{G}$ is complete, then there is no zero restriction on the entries of Ω. Given n sample points from a $N_v(0, \Sigma)$ distribution, the data matrix is the $n \times v$ matrix with i-th row equal to the i-th data point. The matrix $X^T X/n$ is the sample covariance matrix. Of course, the inverse $(X^T X/n)^{-1}$ does not have to have entry (i, j) equal to zero whenever $(i, j) \notin E$. Furthermore if $v > n$, the matrix $X^T X$ is not invertible. The problem of setting some entries of the inverse covariance matrix Ω to zero was first considered by Dempster (1972) as the covariance selection problem.

The reader is referred to Lauritzen (1996) for the basic theory and definitions surrounding undirected Gaussian graphical models. Here we summarize the basic properties necessary for our arguments. If A, B and C are disjoint subsets of V and $A \cup B \cup C = V$ then C *separates* A and B if all paths from A to B must pass through C. The undirected Gaussian graphical models we consider here have the *global Markov property*, that is, $(X_i, i \in A)$ are independent of $(X_j, j \in B)$ conditional on $(X_k, k \in C)$, whenever C separates A and B. We commonly say that A is independent of B given C. If furthermore C is complete then (A, B, C) is called a *decomposition* of $\mathcal{G}$. If we iteratively decompose the graph until no further decompositions can be found, then the subgraphs so found are the set of *prime components*. If the prime components are all complete they are called *cliques* and the graph is a (fully) *decomposable graph*. The existence of a decomposition does not imply that a graph is decomposable: if any of the prime components found by iterative decomposition are not complete and cannot be further decomposed then that component is non-decomposable and therefore so is the whole graph.

Given a graph $\mathcal{G} = (V, E)$, we define a *superset graph* to be one which includes all the edges of $\mathcal{G}$ plus at least one other edge which is not in E. A minimal superset graph of a non-decomposable graph $\mathcal{G}$ includes only the minimal number of extra edges needed to achieve decomposability (a minimal triangulation). In a similar vein we define a *subset graph* to be one which includes no extra edges, and also fails to include at least one of the edges in E.

Given n sample points from the $N_v(0, \Omega^{-1})$ distribution, gathered in an $n \times v$ sample matrix X, the joint density of the rows of X is

$$\frac{|\Omega|^{n/2}}{(2\pi)^{nv/2}} \exp -\frac{1}{2n} \langle \Omega, X^T X/n \rangle$$

where $\langle a, b\rangle$ denotes the trace of $a^T b$. The Diaconis-Ylvisaker conjugate prior is then of the form

$$f(\Omega|\, \mathcal{G}, \delta, \Phi) = h(\mathcal{G}, \delta, \Phi)^{-1} |\Omega|^{(\delta-2)/2} \exp -\frac{1}{2n} \langle \Omega, \Phi\rangle$$

where, following the notation in Roverato (2002), $h(\mathcal{G}, \delta, \Phi)$ denotes the normalizing constant. The distribution with this density is called the hyper-inverse Wishart prior denoted $HIW(\mathcal{G}, \delta, \Phi)$

Using the conjugate $HIW(\mathcal{G}, \delta, \Phi)$ prior, the marginal likelihood of a graph $\mathcal{G}$ is

$$p(X|\mathcal{G}) = (2\pi)^{-nv/2} \frac{h(\mathcal{G}, \delta, \Phi)}{h(\mathcal{G}, \delta*, \Phi^*)} \tag{1}$$

where δ^* and Φ^* are the parameters of the posterior and incorporate, respectively, n and $X^T X$, the sum of squares matrix for the data matrix X. The normalizing constant $h(.)$ can be obtained in closed form for decomposable graphs only; this fact leads to their computational advantages.

There are two main prior specification approaches. Jones et al. (2005) use $\delta = 3$ and $\Phi = \tau I$ leading to $\delta^* = \delta + n$, $\Phi^* = \Phi + X^T X$; for our practical examples we use the fractional-Bayes (G-prior) approach of Carvalho and Scott (2009) which specifies the hyper-inverse Wishart scale parameter of the prior in terms of the sums of squares matrix:

$$\delta = gn,\ \Phi = gX^T X; \delta^* = n,\ \Phi^* = X^T X, \tag{2}$$

with g taken to be $1/n$ so that $\delta = 1$.

2.2 Graphs differing by one edge

For the HIW prior, the ratio of prior normalizing constants remains constant as n changes; for the G-prior it converges to

$$\frac{h(\mathcal{G}', 1, \Sigma)}{h(\mathcal{G}, 1, \Sigma)}$$

thus the asymptotic behavior of the marginal likelihood ratio of $\mathcal{G}'$ to $\mathcal{G}$ can be understood by considering:

$$B = \frac{h(\mathcal{G}, \delta^*, \Phi^*)}{h(\mathcal{G}', \delta^*, \Phi^*)}. \tag{3}$$

(Note the marginal likelihood ratio is also the Bayes Factor when graph structures are assumed to be equally likely apriori.)

Suppose that removing the edge (a, b) from decomposable graph $\mathcal{G}$ gives decomposable graph $\mathcal{G}'$. Armstrong et al. (2009) showed that if removing edge (a, b) maintains decomposability, it affects a single clique C. Let $C_1 = C/a$ and $C_2 = C/b$, S=$C/a, b$, and $p \leq v$ be the number of vertices in C. Using results from Lauritzen (1996), they found the following expression for (3):

$$\frac{h(\mathcal{G},\delta^*,\Phi^*)}{h(\mathcal{G}',\delta^*,\Phi^*)} = \frac{\left|\frac{\Phi^*_C}{2}\right|^{\frac{\delta^*+p-1}{2}} \Gamma_p(\frac{\delta^*+p-1}{2})^{-1} \left|\frac{S}{2}\right|^{\frac{\delta^*+p-3}{2}} \Gamma_{p-2}(\frac{\delta^*+p-3}{2})^{-1}}{\left(\left|\frac{\Phi^*_{C1}}{2}\right|\left|\frac{\Phi^*_{C2}}{2}\right|\right)^{\frac{\delta^*+p-2}{2}} \Gamma_{p-1}(\frac{\delta^*+p-2}{2})^{-2}}. \tag{4}$$

This simplification is most obvious when C_1 and C_2 are cliques in $\mathcal{G}'$ with separator S, but Armstrong et al. (2009) bring together results from Lauritzen (1996) that show the equality in general, including the case where $S = \emptyset$.

Let

$$\Psi = \frac{\Phi^*}{n} = \frac{X^t X}{n} + \frac{\Phi}{n}.$$

The asymptotic behavior of Ψ is essentially the behavior of $X^T X/n$. Substituting $n\Psi$ for Φ^* into (4) and simplifying, B, as defined in (3), becomes

$$B = \frac{\Gamma\left(\frac{\delta^*+p-2}{2}\right)}{\Gamma\left(\frac{\delta^*+p-1}{2}\right)} \times \frac{1}{\sqrt{\pi}}\left(\frac{n}{2}\right)^a \times \frac{|\Psi_C|^{\frac{\delta^*+p-1}{2}}|\Psi_S|^{\frac{\delta^*+p-3}{2}}}{|\Psi_{C_1}|^{\frac{\delta^*+p-2}{2}}|\Psi_{C_2}|^{\frac{\delta^*+p-2}{2}}}$$

where

$$a = (p)\frac{\delta^*+p-1}{2} + (p-2)\frac{\delta^*+p-3}{2} - 2(p-1)\frac{\delta^*+p-2}{2} = 1.$$

Now arrange Ψ so that the entries pertaining to S are listed first and those pertaining to $D = \{a, b\}$ are last. The Cholesky decomposition of $\Psi = LL^T$ (see Armstrong et al. 2009) is such that

$$L = \begin{pmatrix} L_S & 0 \\ L_{DS} & L_D \end{pmatrix} = \begin{pmatrix} \Psi_S^{1/2} & 0 \\ \Psi_{DS}\Psi_{D|S}^{-1/2} & \Psi_{D|S}^{1/2} \end{pmatrix}, L_D = \begin{pmatrix} l_{aa} & 0 \\ l_{ba} & l_{bb} \end{pmatrix}.$$

The following identities are applied:

$$|\Psi_C| = \left|\Psi_{D|S}\right||\Psi_S| \quad \text{where} \quad \Psi_{D|S} = \Psi_D - \Psi_{DS}(\Psi_S)^{-1}\Psi_{DS} \tag{5}$$

$$|\Psi_{C1}| = \left|\Psi_{a|S}\right||\Psi_S| \quad \text{where} \quad \Psi_{a|S} = \Psi_a - \Psi_{aS}(\Psi_S)^{-1}\Psi_{aS} \tag{6}$$

$$|\Psi_{C2}| = \left|\Psi_{b|S}\right||\Psi_S| \quad \text{where} \quad \Psi_{b|S} = \Psi_b - \Psi_{bS}(\Psi_S)^{-1}\Psi_{bS} \tag{7}$$

$$\begin{aligned} |\Psi_{D|S}| &= |L_D L_D| = l_{aa}^2 l_{bb}^2 \\ |\Psi_{a|S}| &= (l_{aa})^2 \\ |\Psi_{b|S_q}| &= (l_{ba})^2 + (l_{bb})^2. \end{aligned}$$

The Bayes factor can then be written:

$$B = \frac{n}{2\sqrt{\pi}} \times \frac{\Gamma\left(\frac{\delta^*+p+2}{2}\right)}{\Gamma\left(\frac{\delta^*+p-1}{2}\right)} \times l_{aa}l_{bb} \times \left(\frac{1}{1+\left(\frac{l_{ba}}{l_{bb}}\right)^2}\right)^{\frac{\delta^*+p-2}{2}}.$$

We know that

$$\frac{\Gamma\left(\frac{\delta^*+p+2}{2}\right)}{\sqrt{2}\Gamma\left(\frac{\delta^*+p-1}{2}\right)} \approx \frac{1}{\sqrt{\delta^*+p-2}}.$$

In addition, $X^TX/n = (\sum_{i=1}^n X_i^TX_i)/n$ where X_i is the i-th sample point, and is therefore governed by the strong law of large numbers; the Cholesky factorization is a differentiable function so the mean value therorem implies L_D will be consistent for the analogous function $\Sigma_{D|S}^{1/2}$ of Σ. So l_{aa} and l_{bb} converge to positive limits and need not be considered further. We then have

$$\begin{aligned} B &\approx \frac{n}{\sqrt{n+\delta+p-2}} \times \left(\frac{1}{1+\left(\frac{l_{ba}}{l_{bb}}\right)^2}\right)^{\frac{\delta^*+p-2}{2}} \times W \\ &\approx \sqrt{n}\left(\frac{1}{1+\left(\frac{l_{ba}}{l_{bb}}\right)^2}\right)^{\frac{\delta^*+p-2}{2}} \times W \end{aligned}$$

where $W = l_{aa}l_{bb}\sqrt{2\pi}$.

The quantities l_{ba} and l_{bb} are key to understanding the asymptotic behavior of B. We make two observations: first, that the central limit theorem and delta method imply that these will be asymptotically normal. Second, that $\Psi_{D|S}$ is a consistent estimate of $\Sigma_{D|S}$, which has a statistical as well as an algebraic interpretation: it is the covariance matrix of a, b conditional on the variables in S. Using the global Markov property, we see that the off-diagonal elements of $\Sigma_{D|S}$ (and its Cholesky decomposition, $\Sigma_{D|S}^{1/2}$) will be zero exactly when S separates a and b in the true graph.

In other words, l_{ba} will converge to something non-zero when edge (a, b) is present in the true graph. When edge (a, b) is absent

$$X_n = \sqrt{n}l_{ab} \tag{8}$$

will converge in distribution (to a central Gaussian). In either case, the diagonal elements of $\Sigma_{D|S}^{1/2}$ are positive; let μ be the diagonal entry of $\Sigma_{D|S}^{1/2}$ corresponding to l_{bb}. Then

$$Y_n = \sqrt{n}(l_{bb} - \mu) \tag{9}$$

will also converge in distribution (again to a central Gaussian).

Theorem 1: When $\Omega_{ab} \neq 0$, that is, when (a, b) is an edge of $\mathcal{G}$, the marginal likelihood ratio B comparing $\mathcal{G}'$, the graph without this edge, to $\mathcal{G}$, will converge to zero.

Proof If $\Sigma_{ab|S} \neq 0$ then $\left(\frac{l_{ba}}{l_{bb}}\right)^2$ tends to a finite positive limit and $1 + \left(\frac{l_{ba}}{l_{bb}}\right)^2 > 1$. The quantity

$$\left(\frac{1}{1 + \left(\frac{l_{ba}}{l_{bb}}\right)^2}\right)^{\frac{\delta^*+p-2}{2}}$$

tends to 0 exponentially. This is much faster than the rate at which $\sqrt{n}$ goes to $+\infty$, so $B \to 0$. □

Lemma 1: Suppose $\Sigma_{ab|S} = 0$ and let X_n, Y_n and μ, be as defined in (8) and (9) above. Then

$$A(n) = \sqrt{n}\frac{l_{ba}}{l_{bb}} = \frac{\sqrt{n}X_n}{Y_n + \mu\sqrt{n}}$$

is bounded in probability.

Proof: see appendix.

Theorem 2: When $\Omega_{ab} = 0$, that is, when (a, b) is not an edge of $\mathcal{G}$, the marginal likelihood ratio B will diverge in favor of $\mathcal{G}'$.

Proof: $\Omega_{ab} = 0$ implies that $\Sigma_{ab|S} = 0$. It follows from Lemma 1 that

$$A(n) = \sqrt{n}\frac{l_{ba}}{l_{bb}}$$

is stochastically bounded, and therefore so is $A^2(n)$. Therefore, $\forall \epsilon > 0$ there exists $A_1(\epsilon)$ and $A_2(\epsilon)$ such that, for all n,

$$A_1 \leq A^2(n) \leq A_2$$

or equivalently

$$\frac{A_1}{n} \leq \frac{A^2(n)}{n} \leq \frac{A_2}{n}$$

with probability $> 1 - \epsilon$. As $n \to +\infty$ this implies

$$e^{-A_2} \leq \left(\frac{1}{1 + \left(\frac{l_{ba}}{l_{bb}}\right)^2}\right)^{\frac{\delta^*+p-2}{2}} \leq e^{-A_1}.$$

The $\sqrt{n}$ term dominates and $B \to +\infty$. □

Theorem 3: Suppose $\mathcal{G}'^1$ and $\mathcal{G}'^2$ are different minimal triangulations of the true graph, i.e. they each contain the same number of edges, including all edges corresponding to the non zero elements of Ω. The log of the marginal likelihood ratio of $\mathcal{G}'^1$ and $\mathcal{G}'^2$ is bounded in probability and therefore the Bayes factor will neither converge to zero, nor diverge.

Proof: The space of decomposable graphs can be traversed by a series of one-edge moves. Moving between $\mathcal{G}'^1$ and $\mathcal{G}'^2$ can be accomplished with some number k of edge additions and an equal number of edge subtractions; the marginal likelihood ratio between $\mathcal{G}'^1$ and $\mathcal{G}'^2$ will be the product of the marginal likelihood ratios for these moves. There are k factors of $\sqrt{n}$ in both numerator and denominator, which cancel each other. Let $R_1(n) \dots R_{2k}(n)$ be the remaining ratio terms. The proof of Theorem 2 implies that the $\log(R_i(n))$ are each stochastically bounded as $n \to \infty$. Therefore, by Lemma 2 in the appendix, $\sum_{i=1}^{2k} \log(R_k)$ is stochastically bounded and $\forall \epsilon > 0$ there exists $B_1(\epsilon), B_2(\epsilon)$ such that for n sufficiently large

$$P\left(B_1 \leq \sum_{i=1}^{2k} \log(R_i(n)) \leq B_2\right) > 1 - \epsilon.$$

Since

$$\exp\left(\sum_{i=1}^{2k} \log(R_i)\right) = \prod_{i=1}^{2k} R_i$$

this implies

$$P\left(e^{B_1} \leq \prod_{i=1}^{2k} R_i \leq e^{B_2}\right) > 1 - \epsilon$$

for finite k. □

Computation of the log-likelihood ratio between the two different triangulations of the four cycle for large simulated data sets (Figure 1) suggests that the convergence result in Theorem 3 is actually the strongest possible: the likelihood ratio is stochastically bounded but is not decreasing in variance, i.e. it is not converging to a constant. We conjecture that l_{ab}/l_{bb} is asymptotically normal, since it is a maximum likelihood estimate (MLE) of $\Sigma_{ab|S}^{1/2}/\Sigma_{bb|S}^{1/2}$. If this conjecture is true we would expect $-\log(R_i)$ to have a scaled chi-squared distribution with one degree of freedom. Evaluations of $-\log(R_1)$ for the simulated data show the scaled chi-squared distribution indeed provides a good fit.

3 Methods

3.1 Feature-inclusion stochastic search

There are many possible algorithms to explore the posterior of graph space. We use feature-inclusion stochastic search (FINCS, Scott and Carvalho 2008) as a representative example. It uses the fractional Bayes formulation in equation (2), and places

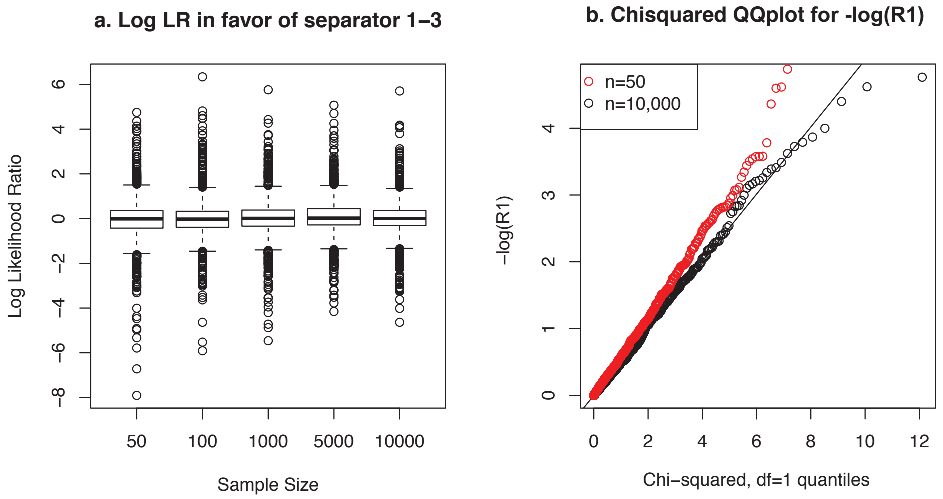

Figure 1: a. Log likelihood ratios between the two possible minimal supersets for the four cycle. Data were simulated from a multivariate normal with Ω corresponding to a four cycle, with diagonal values of 20, and non zero off diagonals set to 9. Each box-and-whisker shows 1000 simulated data sets for each sample size. b. Q-Qplot comparing the χ_1^2 distribution with the observed values of $-\log(R_1)$, The line is fitted to the data with n=10,000; it is constrained to have zero intercept. The fitted slope is 0.5023, which provides an estimate of the scaling factor for the χ_1^2 distribution.

(unnormalized) prior weight of

$$\frac{k!(m-k)!}{(m+1)(m!)}$$

on graphs with k edges out of a possible $m = v(v-1)/2$. This is referred to as a multiplicity correction prior because the more possible graphs there are with k edges, the more strongly graphs of that size are penalized.

FINCS retains a list of the models with highest posterior probability at any given time. We have retained the top 1000 models. This truncated list of models is used to summarize the importance of a particular edge by recording for each model whether the edge is included (1) or not (0). A weighted average of these ones and zeros is then produced, where the weights correspond to the posterior probability of each model, normalized to sum to one over the list of retained models. We refer to these weighted averages as the estimated edge inclusion probabilities, although these are obviously (potentially seriously) biased estimates of the true edge inclusion probabilities. We show these estimate can, nevertheless, be helpful if the goal is recovery of the true graph structure.

The FINCS search combines three types of moves through the space of all possible graphs. Most moves are local moves which exploit the computational advantages of adding or deleting only one edge at a time. Global moves are generated by starting with an empty graph and adding edges in proportion to their current estimated inclusion probability. The graph so formed is usually not decomposable so a randomized median triangulation pair, consisting of a minimal decomposable supergraph ($\mathcal{G}^+$) and a maximal decomposable subgraph ($\mathcal{G}^-$), is found. Posterior probabilities are then calculated for both $\mathcal{G}^+$ and $\mathcal{G}^-$ and the one with the highest posterior probability chosen. Finally, resampling moves revisit graphs in proportion to their posterior probability and thereby ensure that the global moves do not irretrievably direct the search away from 'good' graphs.

We use the C++ implementation of FINCS described in Scott and Carvalho (2008), with the recommended settings (global moves every 20 iterations, and resampling moves every 10 iterations). For the $v = 4$ cases, where there are only 61 possible decomposable graphs, 100 iterations were used. In all other cases we ran FINCS for 3 million iterations.

There are various ways of extracting estimates and predictions from the list of top models produced by FINCS. The simplest is to use the best model found and to take our estimate of Ω to be $\overline{\Omega}$; the posterior mean conditional on that graph. Note that this is computable only because we are dealing with a decomposable graph. We also consider the ability of inclusion probabilities to point towards the true (non-decomposable) graph by identifying the graph obtained by specifying as edges, those with an inclusion probability of at least 0.8. In this case $\widehat{\Omega}$ is the maximum likelihood estimate conditional on the graph. The R package "glasso" provides a convenient way to compute this by specifying shrinkage penalty `rho = 0`, and zero elements fixed to be those associated with edges with an inclusion probability less than 0.8. Finally, when our main concern is prediction, $\widehat{\Sigma} = \overline{\Omega}^{-1}$ is used to provide a prediction from a particular graph. A model averaged prediction was then obtained by calculating a weighted mean of predictions using each of the top 1000 models (where the weight is the posterior probability normalized over the list of top graphs).

3.2 Graphical lasso and adaptive graphical lasso

In their paper Scott and Carvalho (2008) compare the prediction performance of covariance structures discovered by FINCS to those obtained using lasso regression of each variable on the remaining variables to obtain a sparse graph in the manner of Meinhausen and Bühlmann (2006). The more recent graphical lasso (Friedman et al. 2008b) applies an L_1 penalty directly to the inverse covariance matrix elements with superior performance. Thus the objective function is

$$\log \det \Omega - tr(\Omega S) - \lambda \sum_{i=1}^{p} \sum_{j=1}^{p} |\omega_{ij}| \tag{10}$$

where Ω is a positive definite matrix, S is the sample covariance matrix and $\lambda > 0$ is the penalty.

We use the R-package `glasso`, (R Development Core Team 2009; Friedman et al. 2008a), with the penalty selected by 5-fold cross validation and the sample covariance estimated with an n divisor to obtain our graphical lasso estimate. The graphical lasso algorithm as implemented in `glasso` yields an estimated inverse covariance matrix that is not perfectly symmetric (at 3-4 significant figures). We used an inverse covariance matrix made exactly symmetric by using the average of the i, j^{th} and j, i^{th} elements.

Fan et al. (2009) recommend adaptive graphical lasso, a method that typically obtains sparser graphs than the graphical lasso and ameliorates the lasso's bias towards zero for non-zero elements. The adaptive graphical lasso is implemented using a penalty matrix (ζ) rather than the scalar penalty term of graphical lasso. The elements of ζ are $\zeta_{i,j} = 1/|\tilde{\omega}_{i,j}|^{\gamma}$, where $\tilde{\Omega} = (\tilde{\omega}_{i,j})_{1 \le i,j \le p}$ is any consistent estimate of Ω and $\gamma > 0$. Thus for adaptive graphical lasso the objective function becomes

$$\log \det \Omega - tr(\Omega S) - \lambda \sum_{i=1}^{p} \sum_{j=1}^{p} \zeta_{i,j} |\omega_{ij}|. \tag{11}$$

We implemented adaptive graphical lasso using the symmetrized graphical lasso estimated inverse covariance matrix as $\tilde{\Omega}$, $\gamma = 0.5$ and selecting the penalty by 5-fold cross-validation. We again used the R-package `glasso`, making the estimate exactly symmetric in the same manner as for the graphical lasso estimate.

3.3 Model comparisons

When comparing models, we consider three criteria: the Kullback-Leibler divergence from the true model, the precision and recall of edge selection, and the accuracy of predictions based on the model fit.

The Kullback-Leibler divergence between two density functions f and g is

$$E[\log(f(X)/g(X))]$$

where the expectation is with respect to f; see Whittaker (2008, p168) for the formula. We set f as the true model and g the estimate so that the Kullback-Leibler divergence (KL) is calculated as

$$KL = \frac{1}{2} tr(\Sigma \widehat{\Omega} - I_k) - \frac{1}{2} \log \det(\Sigma \widehat{\Omega}) \tag{12}$$

where I_k is the k by k identity matrix, $\Sigma = \Omega^{-1}$ is the true covariance matrix and $\widehat{\Omega}$

is the estimated Ω matrix . The Kullback-Leibler divergence is used as a measure that will treat Ω matrices as similar if their elements (rather than the pattern of strictly non-zero elements) are similar. Superfluous edges corresponding to very small entries in Ω will have little influence on KL.

For each fitted model we define precision and recall as:
$$\text{precision} = \tfrac{T_E}{T_E+F_E} \quad \text{and} \quad \text{recall} = \tfrac{T_E}{T_E+F_0}$$

where T_E is the number of true edges found, F_E is the number of edges found that are not true edges and F_0 is the number of true edges that were not found. Thus precision is the proportion of edges in the model that are true edges and recall is the proportion of true edges found by the current model. A superset graph as defined in Section 2.1 is thus a model with a recall of one and precision of less than one. A subset graph has a precision of one and a recall of less than one.

Prediction accuracy was considered after estimating $\widehat{\Sigma} = \overline{\Omega}^{-1}$ from a training data set. We then take each test data point and imagine we have observed all variables except the ith, which we wish to predict. The prediction used is the expectation of variable i conditional on the observed values of the other variables, which is a function of $\widehat{\Sigma}$. This is repeated for all i and all test samples, and the sum of squared errors for these predictions is our measure of quality.

4 Simulations

4.1 Large sample behavior of FINCS

To examine the sample size at which "large sample" behavior begins, we consider cycles of size 4, 6, 20, 35, 50 and 70. In each case, the Ω matrix used to simulate the data has all diagonals equal to 20, with all non-zero off diagonals equal to 9, making all partial correlations 0.45. We will refer to this pattern as Ω_{same}. Data was simulated from a multivariate normal distribution using the Cholesky decomposition of Ω_{same}^{-1} and the R function `rnorm` (R development team, 2009).

For each number of variables, a single simulation was performed at various sample sizes. Table 1 shows whether or not a particular dimension and sample size combination resulted in the top graph being a superset graph. This suggests that for moderate to large partial correlations we see superset graphs selected roughly when $n \geq 12v$. It should be noted that, apart from the $v = 4$ case, not all superset graphs are visited by the algorithm. For $v \geq 20$ and $n = 1000$ the top 1000 graphs consist entirely of minimal supersets. (However, when replicate situations were produced for the analyses of section (4.3), some $v = 70$ replicates produced minimal supersets, and some replicates produced subsets.) As n decreases and becomes close to $12v$, in some cases the set of top graphs becomes a mixture of superset and non-superset graphs.

The edge inclusion probabilities for the four and twenty variable cycles are given in supplementary Table 2 for sample size 1000 and supplementary Table 3 for sample size 50. When the sample size is adequate for the top graph to be a superset graph ($n = 50$ or 1000 for $v = 4$ and $n = 1000$ for $v = 20$) there are some large inclusion probabilities for superfluous edges. When $v = 4$, edge (2,4) has posterior probability 0.724, while (3,1) has posterior probability 0.351. Either edge can be added to the true graph to produce a triangulation, but as shown in Figure 1, one can be favored by chance. (Note that for $v = 4$ all possible graphs are visited and thus the true posterior edge probabilities are recaptured–the imbalance cannot be attributed to poor mixing.)

When the top graph is not a superset ($n = 50$, $v = 20$) the inclusion probability for superfluous edges ranges from 0 to 0.3, and the inclusion probability for true edges ranges from 0.5 to 1. This is one of the few examples we have seen where an appropriate threshold (0.5) would recapture the true graph. More typically, thresholding the inclusion probabilities suggests the true graph is non-decomposable, but does not recapture it exactly.

Table 1: Relationship between sample size, number of variables and when the graph with the highest posterior probability is a superset graph. In all cases the true graph was a cycle and all true partial correlations were 0.45. Y indicates superset graph, n indicates not a superset graph. $^*n = v + 1$.

				n		
v	30	50	70	100	240	1000
4	n	Y	Y	Y		Y
6		n	Y	Y		Y
20		n		n	Y	Y
35		n			n	Y
50		n*			n	Y
70			n*			Y

4.2 Effect of partial correlations

We also examined the impact of the size of the partial correlations on the set of top models fitted by FINCS, when the sample size was large ($n = 1000$). For the cycles of size 4 and 20, we changed the off diagonal elements of Ω_{same} so the partial correlations took on the values 0.05, 0.10, 0.15, 0.20, 0.25 and 0.30. We then selected either 10 or 50 top models and observed whether each was a superset graph. Results are shown in Figures 2 and 3.

The dimension and partial correlations clearly interact in the way they affect the inferred graphs. With $n = 1000$, the top four-cycle graphs are superset graphs with partial correlations as low as 0.15; this does not happen for 20 node graphs until the partial correlation is 0.3.

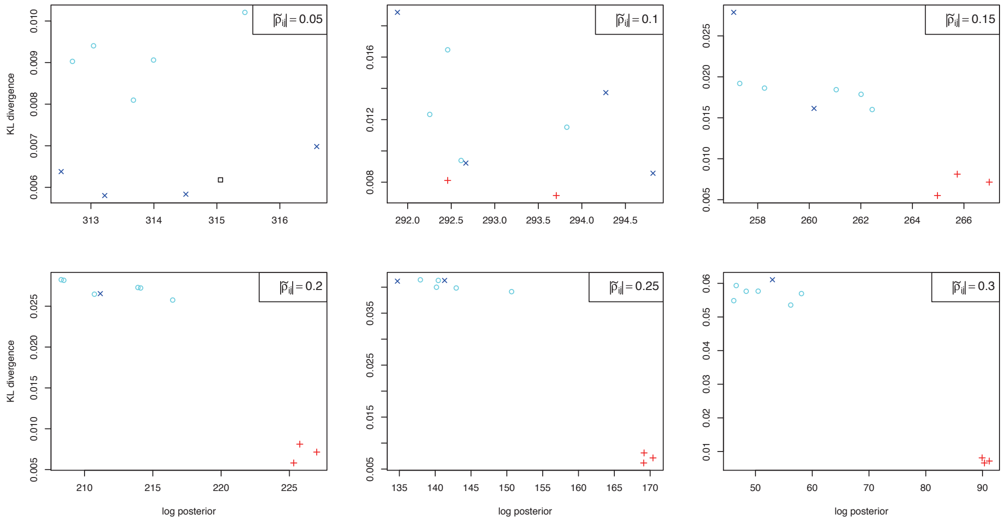

Figure 2: Top 10 graphs found by FINCS for p=4, samples of n=1000 for different $|\tilde{\rho}_{ij}|$. + = superset graphs (red); × = subset graphs (blue); ○=subset plus incorrect edges (light blue); □ = empty graph (black). Note there are only 3 possible superset graphs.

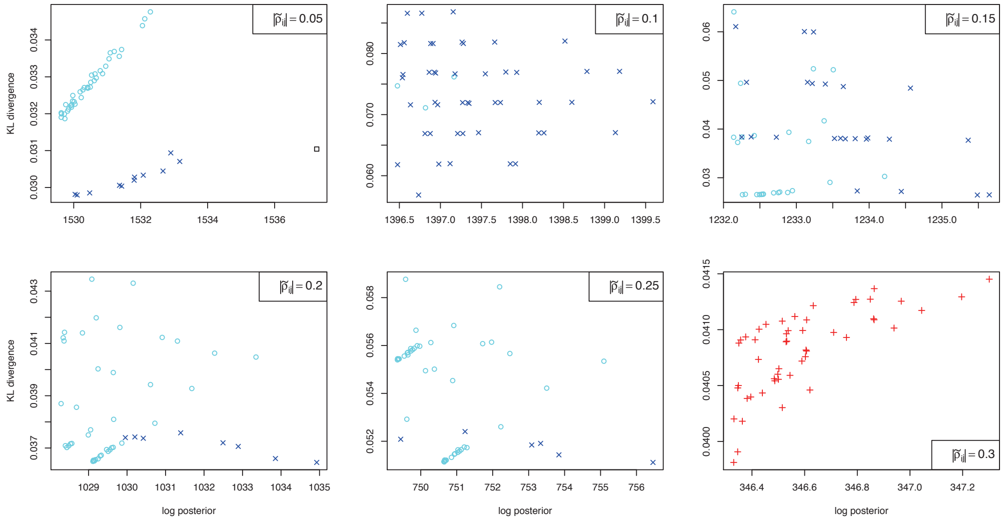

Figure 3: Top 50 graphs found by FINCS for p=20, samples of n=1000 for different $|\tilde{\rho}_{ij}|$. + = superset graphs; × = subset graphs; ○=subset plus incorrect edges; □ = empty graph

4.3 Comparison with penalized likelihood methods

To empirically compare the FINCS based methods (top graph, inclusion probability graph) and the two graphical lasso based methods (graphical lasso and adaptive graphical lasso), we used Ω_{same} with $v = 35, 50, 70$, simulating multivariate normal data as above. Each matrix was used to simulate five different datasets of size $n = \max(50, v+1)$ and five of size $n = 1000$. (The smaller sample size was chosen to allow computation of the MLE for the complete graph, however the results were not instructive and are not shown.) For prediction purposes a test dataset (of size $n = 50$) was also simulated. A model averaged FINCS estimate was added to the comparison when considering prediction accuracy.

The estimation methods were assessed as described in Section 3.3: the precision and recall of edge selection (Figure 4), Kullback-Leibler divergence from the true model (Figure 5) , and the accuracy of predictions based on the model fit (Figure 6).

At large sample size, the recall was usually 1 for all methods, the exception being three replicates for FINCS and $v = 70$. The FINCS based methods had consistently better precision. The precision for a minimal superset for a v node cycle is constrained to be $v/(2v-3)$ and this is where the best FINCS graph sits in most cases. The inclusion probability graphs are able to do better, but the level of improvment decreases with dimension; the top 1000 graphs used to produce these probabilities cover a smaller and smaller proportion of the minimal supersets. The set of minimal supersets retained also becomes less diverse with respect to edges included.

At small sample sizes, there is a precision/recall trade off, with the lasso based techniques having perfect recall but poor precision, and FINCS techniques having very good precision but imperfect recall. The FINCS top graphs have between 1 and 4 edges missing and the recall is further reduced for the inclusion probability based graphs.

Figure 5 shows that in each case the best FINCS graph has divergence from the true model similar to or better than adaptive lasso, while producing a sparser graph. (Graphical lasso is consistently least sparse and most divergent.) The advantage of FINCS appears to increase with dimension, especially at small sample size. For large sample sizes the inclusion probability based FINCS estimates yield slightly lower divergence.

The pattern is similar for prediction errors (Figure 6), although adaptive lasso and lasso produce very good estimates for some cases at large sample size. Model averaged predictions are also shown, but make a small improvement, if any, to the overall accuracy of predictions. All the predictions contributing to the average are very similar, so the estimate of uncertainty is not much increased by model averaging either (data not shown). Again, one issue appears to be the similarity of the top 1000 graphs retained, which are all minimal supersets, and in some cases minimal supersets that also share the same extraneous edges.

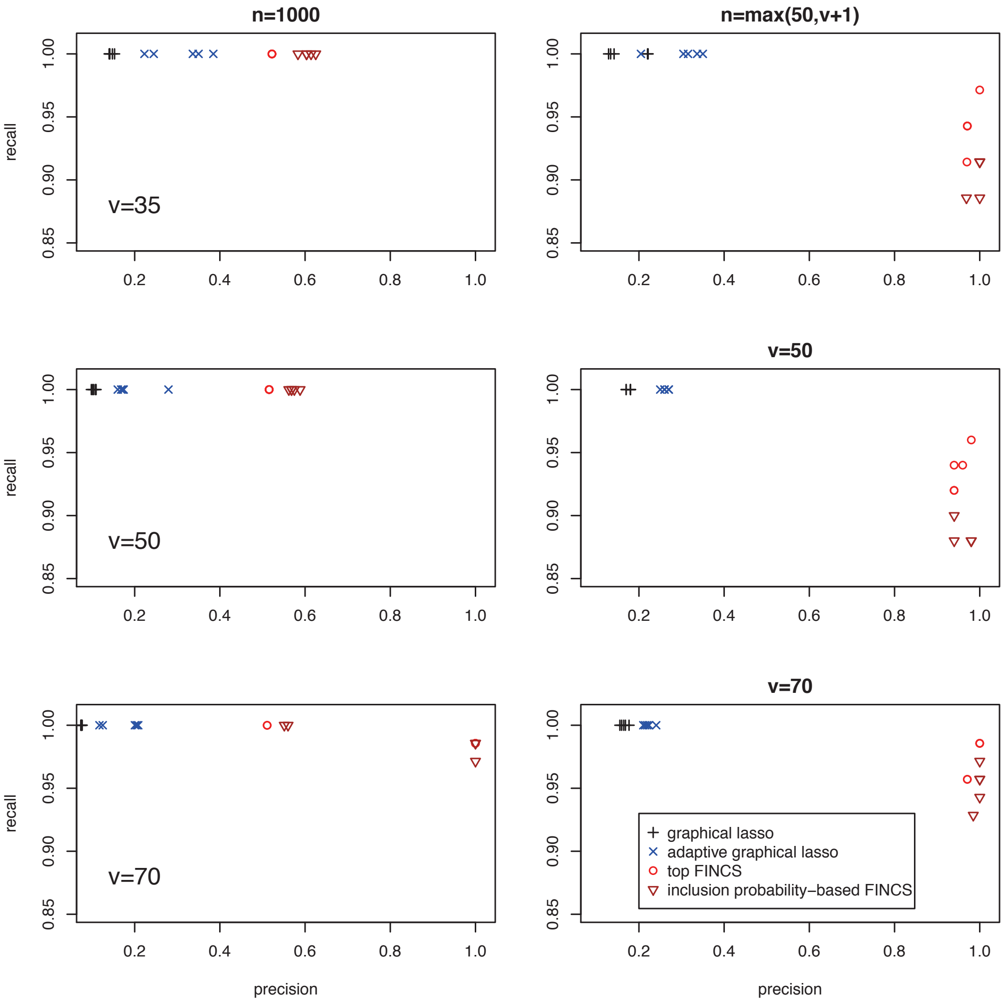

Figure 4: Precision vs. recall for identification of edges from estimates based on data simulated from cycles with $v = 35$, 50, 70 and sample size 1000 or $\max(50,\ v + 1)$.

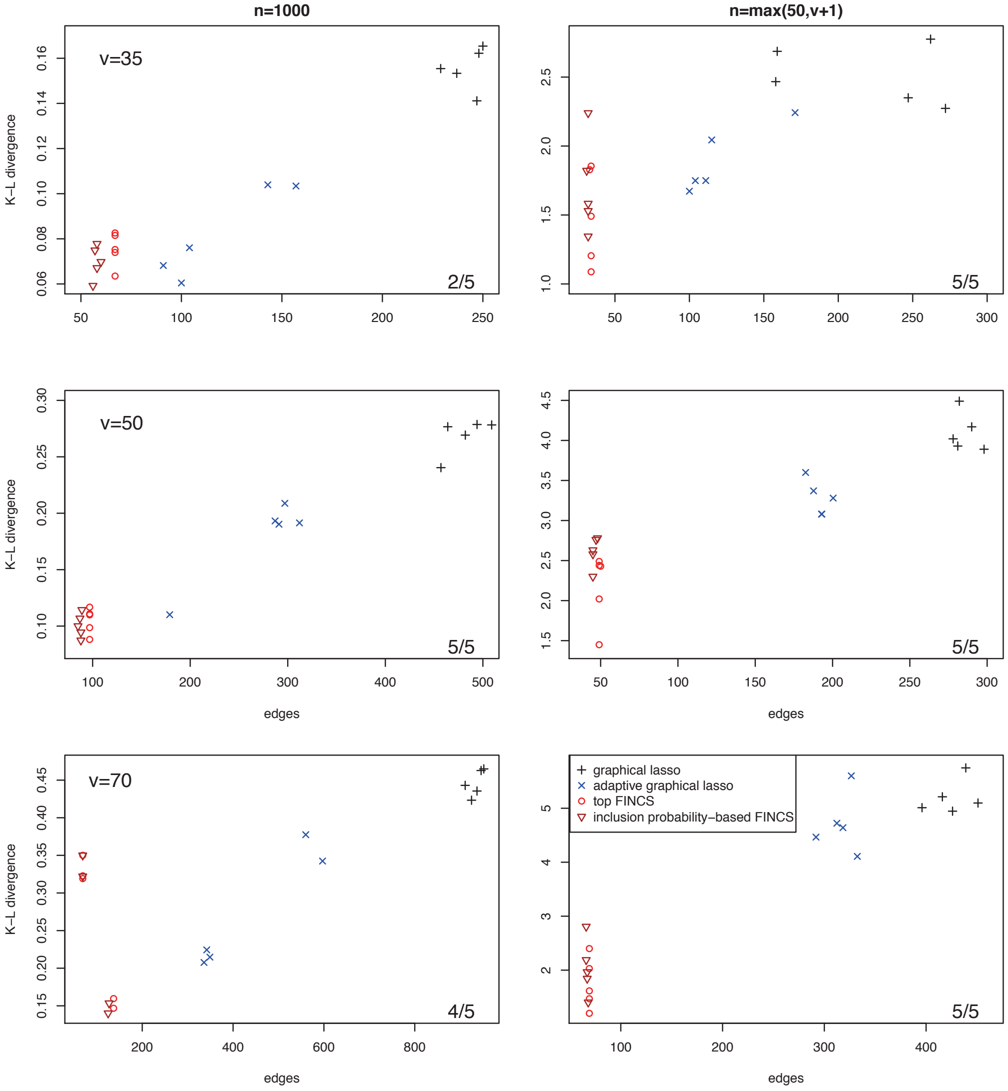

Figure 5: Kullback Leibler divergence vs. number of edges for estimates from data simulated from cycles with $v = 35$, 50, 70 and sample size 1000 or max(50, $v + 1$). Numbers in the lower corner of each panel indicate the number of datasets (out of 5) where the FINCS top graph had lower divergence than the adaptive lasso.

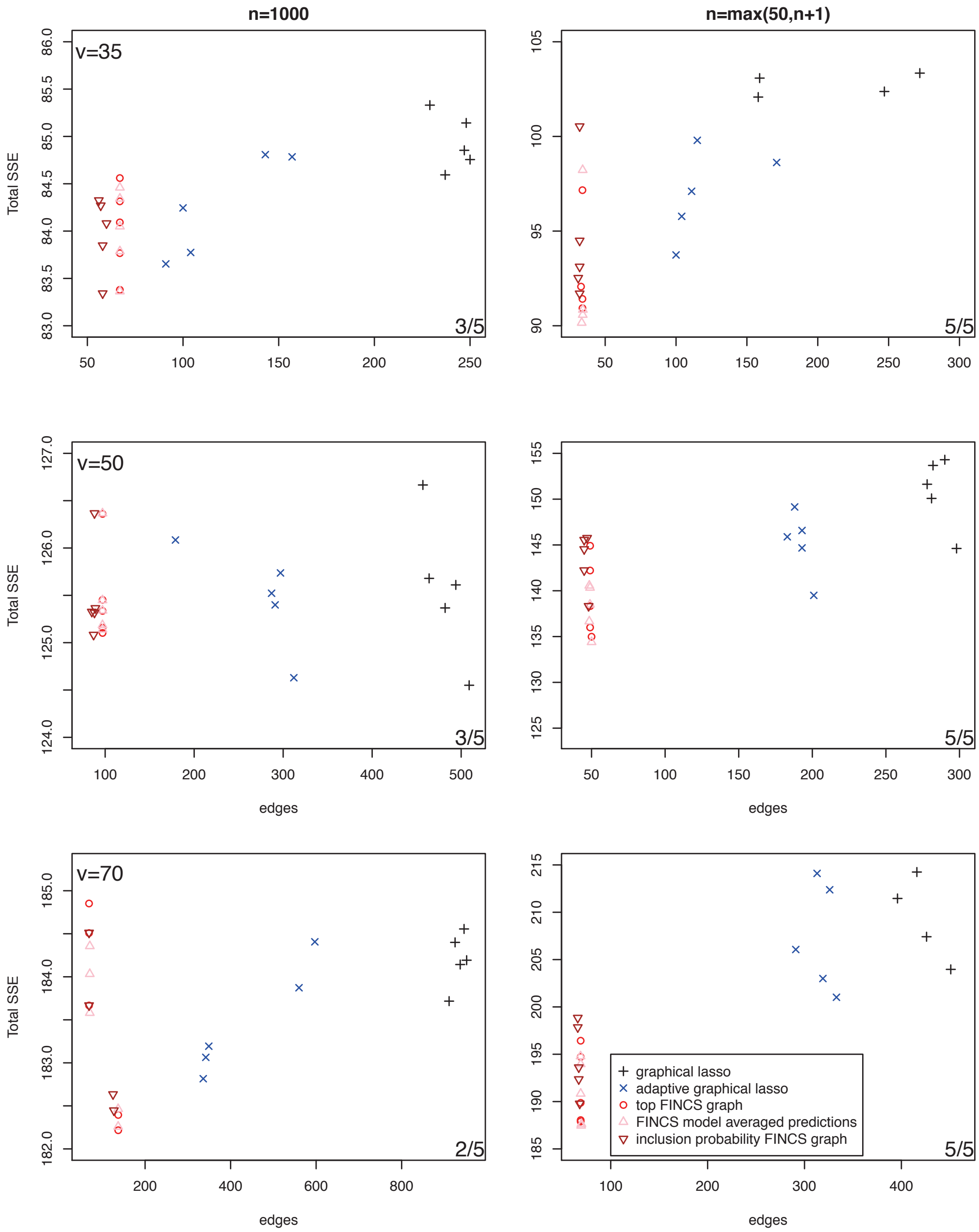

Figure 6: Sum of squared prediction errors on a test set, for estimates from data simulated from cycles with $v = 35$, 50, 70 and sample size 1000 or max(50, $v+1$). Numbers in the lower corner of each panel indicate the number of datasets (out of 5) where the FINCS top graph had lower errors than the adaptive lasso.

4.4 Mutual fund data

We explored the behaviour of each prediction method with real data on the 59-node mutual-funds dataset used in Scott and Carvalho (2008). We split the 86-month sample into a 60 month training set (the first 60 months) and a 26 month prediction set (the remaining 26 months) which enabled us to compare predictions using FINCS derived estimates of the covariance matrix with predictions using lasso derived estimates of the covariance matrix.

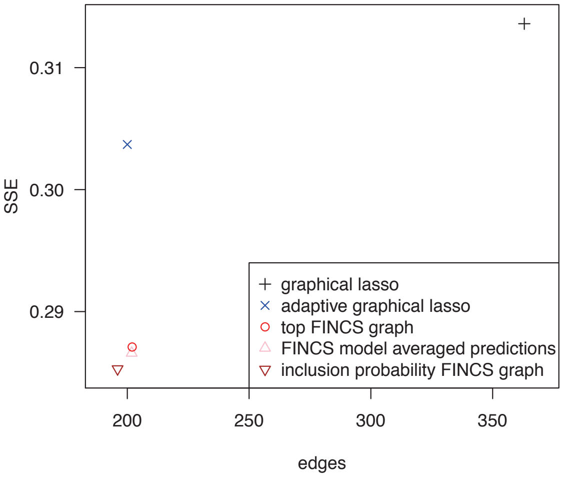

Figure 7: Sum of squared errors vs. number of edges for predictions using the mutual-funds data.

For the mutual funds data, the three FINCS based methods selected similar edges resulting in a similar sum of squared errors (see Figure 7). In contrast to the simulated data, the adaptive graphical lasso estimate is similar in sparsity to the models discovered by FINCS. This could be explained by the (unknown) true structure of the mutual funds data, which is potentially quite different from a large cycle. However, FINCS has sum of squared error 5-8% less than the lasso based models, similar to what was seen for the $v = 50, n = 51$ simulations. We initially suspected the relatively poor SSE for the adaptive lasso was due to shrinkage in the elements of $\widehat{\Omega}$ relative to the FINCS estimates. While this may be a partial explanantion a comparison of the actual edges found by the two methods reveals that although the number of edges is similar, the actual edges found vary considerably, with only 30% of the edges found being common to both models.

5 Discussion

Our examination of the asymptotic behavior of the likelihood ratios suggests that, with sufficient data, Bayesian methods fitting decomposable graphical models should converge to the set of minimal supersets of the true graph. The log likelihood ratio between minimal supersets is bounded in probability; simulation suggests the relative posterior will depend on both the true value of Σ and the data, regardless of the sample size. We are not aware of other examples of this phenomenon, but it may be worth looking for in other contexts where the choice of model space forces overparametrization.

The behavior of a particular posterior exploration algorithm, feature inclusion stochastic search, largely reflected these theoretical results. When $n = 1000$, partial correlations were large (0.45), and dimension was moderate ($20 \leq v \leq 50$), the top graph (and in fact the top 1000 graphs) were always minimal superset graphs.

The size of the partial correlations and number of variables interacted when determining what partial correlations were adequate to produce superset graphs; with $n = 1000$ and $v = 4$, superset graphs were obtained with partial correlation 0.15; this increased to 0.3 when $v = 20$. These rules of thumb are dependent on the particular prior choices made: the use of the G-prior for the covariance matrix, and the multiplicity correction prior over graphs. They are also likely to be dependent on the fact we have chosen to study large cycles. These are a worst case scenario in terms of the number of edges that must be added to create a decomposable graph, but a best case when considering removing edges to create such a model. Behavior of (e. g.) the lattice models considered in Dobra et al. 2012 may be quite different.

One might hope that, even when restricting the search to decomposable models, the edge inclusion probabilities (even the biased edge inclusion probabilities produced by FINCS) would point to the true non-decomposable model. In our experiments, the graph based on thresholding the edge-inclusion probabilites is typically non-decomposable, but this approach rarely recaptures the true graph exactly. The theoretical potential for some minimal supersets to be heavily favored over others in posterior probability offers one possible explanation for why some of the 'extra' edges have high inclusion probabilities. This is the only plausible explanation for the $v = 4$ results, where the small dimension allows us to be confident we have the true posterior. At higher dimension, other possible causes are the particulars of the FINCS algorithm: some superfluous edges may have high inferred posterior probability because we have failed to explore the alternative triangulations, or because there are so many minimal triangulations that retaining only the top 1000 graphs will necessarily exclude some. The 'global move' in FINCS is undoubtedly an aid to mixing, but because it generates proposals based on inclusion probabilities it will not move away from a situation where a superfluous edge has already attained a high inclusion probability.

The large number of possible triangulations and the limited number of graphs retained also contribute to the limited benefit of model averaging. In cases where FINCS circulates among superset models, top FINCS graphs all essentially represent the same model. Requiring retained graphs, or graphs selected by the global move, to exceed a minimum Kullback-Leibler divergence from already-retained graphs, could ensure that truly different models (graphs) are available for model averaging purposes. We leave exploration and implementation of these ideas to future work.

Despite this, the results for FINCS compared to adaptive lasso were very good. Even though the FINCS top graph was constrained to include 'extra' edges to make the model decomposable, it had better precision in identifying edges than the adaptive graphical lasso. Results for KL divergence and SSE were competitive with adaptive lasso at large sample sizes, with no method universally preferred across the data sets considered; at low sample sizes, the top FINCS graph did the best, and its advantage increased with the dimension of the problem considered. Thus, despite the idiosyncracies caused by restricting to decomposable models, this approach should not be discounted when something faster than fitting an unrestricted Gaussian graphical model is needed.

References

Armstrong, H., Carter, C. K., Wang, K. F. K., and Kohn, R. (2009). "Bayesian covariance matrix estimation using a mixture of decomposable graphical models." *Statistical Computing*, 19: 303–316. 660

Carvalho, C. M. and Scott, J. G. (2009). "Objective Bayesian model selection in Gaussian graphical models." *Biometrika*.
URL http://biomet.oxfordjournals.org/content/early/2009/05/04/biomet.asp017.abstract 662

Dellaportas, P., Giudici, P., and Roberts, G. (2003). "Bayesian inference for nondecomposable graphical Gaussian models." *Sankhyā*, 65: 43–55. 660

Dempster, A. (1972). "Covariance Selection." *Biometrics*, 28: 157–175. 659, 661

Dobra, A., Lenkoski, A., and Rodriguez, A. (2011). "Bayesian inference for general Gaussian graphical models with application to multivariate lattice data." *Journal of the American Statistical Association*, 106: 1418–1433. 660

Fan, J., Feng, Y., and Wu, Y. (2009). "Network exploration via the adaptive LASSO and SCAD penalties." *The Annals of Applied Statistics*, 3(2): 521–541. 660, 669

Friedman, J., Hastie, T., and Tibshirani, R. (2008a). *glasso: Graphical lasso- estimation of Gaussian graphical models*. R package version 1.2.
URL http://www-stat.stanford.edu/~tibs/glasso 669

— (2008b). "Sparse inverse covariance estimation with the graphical lasso." *Biostatistics*, 9(3): 432–441.
URL http://biostatistics.oxfordjournals.org/cgi/content/abstract/kxm045v1 660, 668

Jones, B., Carvalho, C., Dobra, A., Hans, C., Carter, C., and West, M. (2005). "Experiments in Stochastic Computation for High-Dimensional Graphical Models." *Statistical Science*, 20(4): 388–400. 660, 662

Lauritzen, S. L. (1996). *Graphical Models*. Oxford: Clarendon Press. 659, 661, 663

Meinhausen, N. and Bühlmann, P. (2006). "High-dimensional Graphs and Variable Selection with the Lasso." *The Annals of Statistics*, 34(3): 1436–1462. 668

Moghaddam, B., Marlin, B. M., Khan, M. E., and Murphy, K. P. (2009). "Accelerating Bayesian Structural Inference for Non-Decomposable Gaussian Graphical Models." *Proceedings of the 23rd Neural Information Processing Systems Conference*, 1285–1293. 660

R Development Core Team (2009). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0.
URL `http://www.R-project.org` 669

Roverato, A. (2002). "Hyper inverse Wishart distribution for non-decomposable graphs and its application to Bayesian inference for Gaussian Graphical Models." *Scandinavian Journal of Statistics*, 29: 391–411. 662

Scott, J. G. and Carvalho, C. M. (2008). "Feature-Inclusion Stochastic Search for Gaussian Graphical Models." *Journal of Computational and Graphical Statistics*, 17(4): 790–808. 660, 666, 668, 677

Whittaker, J. (2008). *Graphical Models in Applied Multivariate Statistics*. UK: John Wiley and Sons. 659, 669

Wong, F., Carter, C. K., and Kohn, R. (2003). "Efficient estimation of covariance selection models." *Biometrika*, 90(4): 809–830. 660

Acknowledgments

Tha authors wish to thank Thomas Lumley for helpful discussion, and to thank the referees and the associate editor for suggestions which have greatly improved the paper.

Appendix

Proof of lemma 1: Since $|Y_n + \mu\sqrt{n}| \geq ||Y_n| - \mu\sqrt{n}|$,

$$\frac{\sqrt{n}|X_n|}{||Y_n| - \mu\sqrt{n}|} \geq \frac{\sqrt{n}|X_n|}{|Y_n + \mu\sqrt{n}|}.$$

Thus, for any given $K > 0$,

$$P\left(\frac{\sqrt{n}|X_n|}{|Y_n + \mu\sqrt{n}|} > K\right) \leq P\left(\frac{\sqrt{n}|X_n|}{||Y_n| - \mu\sqrt{n}|} > K\right).$$

Y_n and X_n converge in distribution and are therefore bounded in probability, i.e., $\forall\epsilon > 0$

$$\begin{array}{lll} \exists C(\epsilon) > 0 & \text{such that} & P(|Y_n| \geq C(\epsilon)) \leq \epsilon \text{ and} \\ \exists B(\epsilon) > 0 & \text{such that} & P(|X_n| \geq B(\epsilon)) \leq \epsilon. \end{array}$$

So, for all n:

$$\begin{aligned} P\left(\frac{\sqrt{n}|X_n|}{||Y_n| - \mu\sqrt{n}|} > K\right) &= P\left(\frac{\sqrt{n}|X_n|}{||Y_n| - \mu\sqrt{n}|} > K, |Y_n| < C\right) \\ &\quad + P\left(\frac{\sqrt{n}|X_n|}{||Y_n| - \mu\sqrt{n}|} > K, |Y_n| \geq C\right) \\ &\leq P\left(\frac{\sqrt{n}|X_n|}{||Y_n| - \mu\sqrt{n}|} > K, |Y_n| < C\right) + \epsilon. \end{aligned}$$

Note that $|Y_n| < C$ implies that $||Y_n| - \mu\sqrt{n}| \leq |C + \mu\sqrt{n}|$. This allows us to address the case where $|Y_n| < C$. For all n we have:

$$\begin{aligned} P\left(\frac{\sqrt{n}|X_n|}{||Y_n| - \mu\sqrt{n}|} > K, |Y_n| < C\right) &\leq P\left(\frac{\sqrt{n}|X_n|}{|C + \mu\sqrt{n}|} > K, |Y_n| < C\right) \\ &\leq P\left(\frac{\sqrt{n}|X_n|}{|C + \mu\sqrt{n}|} > K\right). \end{aligned}$$

We note that if we take $K = B(\epsilon)/\mu$ we get

$$\begin{aligned} P\left(\frac{\sqrt{n}|X_n|}{|C + \mu\sqrt{n}|} > K\right) &= P\left(\frac{|X_n|}{\left|\frac{C}{\sqrt{n}} + \mu\right|} > K\right) \\ &= P\left(|X_n| > B\left(\frac{C}{\mu\sqrt{n}} + 1\right)\right) \\ &\leq P(|X_n| > B) \leq \epsilon. \end{aligned}$$

Combining all these results, we obtain that for all n we have

$$P\left(\frac{\sqrt{n}|X_n|}{|Y_n + \mu\sqrt{n}|} > K\right) \leq P\left(\frac{\sqrt{n}|X_n|}{||Y_n| - \mu\sqrt{n}|} > K\right) \leq 2\epsilon.\square$$

Lemma 2 If $U_n,\ n = 1, \ldots, +\infty$ and $V_n,\ n = 1, \ldots, +\infty$ are two sequences of random variables that are stochastically bounded, then their sum is also stochastically bounded.

Proof Since U_n and V_n are stochastically bounded, for any ϵ there exists $A(\epsilon)$ such that for all n, $P(|X_n| \geq A(\epsilon)) \leq \epsilon$ and $P(|Y_n| \geq A(\epsilon)) \leq \epsilon$. We have the following sequence of event inclusions:

$$\{2A(\epsilon) \leq |U_n + V_n|\} \subset \{2A(\epsilon) \leq |U_n| + |V_n|\} \subset \{A(\epsilon) \leq |U_n|\} \cup \{A(\epsilon) \leq |V_n|\}.$$

We therefore have

$$P(|U_n + V_n| \geq 2A(\epsilon)) \leq P(|U_n| \geq A(\epsilon)) + P(|V_n| \geq A(\epsilon)) \leq 2\epsilon,$$

which proves the lemma.

Supplementary tables

Table 2: Inclusion probability matrix for 4-node and 20-node cycle when n=1000 and $\tilde{\rho}_{ij}$=0.45. Inclusion probabilities associated with true edges are in **bold**; inclusion probabilities that are 1.000 and are associated with other edges are in *italics*; inclusion probabilities shown as 1 are 1.000, those shown as 0 are 0.000.

*	**1.000**	0.351	**1.000**
*	*	**1.000**	0.724
*	*	*	**1.000**

*	**1**	0.6	0.0	0.5	0	0	0	0.0	0	0.3	1.0	0.2	0.0	0.0	0	0	0	0.4	**1**
*	*	**1**	0	0.2	0	0	0	0	0	0.0	0.2	0	0	0	0	0	0	0	0
*	*	*	**1**	*1*	0	0	0	0	0	0.1	0.5	0	0	0	0	0	0	0	0.0
*	*	*	*	**1**	0	0	0	0	0	0	0.0	0	0	0	0	0	0	0	0
*	*	*	*	*	**1**	0.5	0.0	0.4	0.0	1.0	0.6	0.0	0	0	0	0	0	0.0	0.0
*	*	*	*	*	*	**1**	0.0	0.3	0.0	0.3	0.0	0	0	0	0	0	0	0	0
*	*	*	*	*	*	*	**1**	0.9	0.1	0.4	0.00	0	0	0	0	0	0	0	0
*	*	*	*	*	*	*	*	**1**	0.0	0.0	0	0	0	0	0	0	0	0	0
*	*	*	*	*	*	*	*	*	**1**	0.9	0.0	0	0	0	0	0	0	0	0
*	*	*	*	*	*	*	*	*	*	**1**	0	0	0	0	0	0	0	0	0
*	*	*	*	*	*	*	*	*	*	*	**1**	0.0	0.0	0.0	0	0	0	0.0	0.0
*	*	*	*	*	*	*	*	*	*	*	*	**1**	0.2	0.0	0	0	0	0.6	0.5
*	*	*	*	*	*	*	*	*	*	*	*	*	**1**	0.0	0	0	0	0.7	0.3
*	*	*	*	*	*	*	*	*	*	*	*	*	*	**1**	0.0	0	0.0	1.0	0.1
*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	**1**	0.3	0.4	1.0	0.0
*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	**1**	0.5	0.5	0
*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	**1**	0.3	0
*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	**1**	0
*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	**1**
*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*

Table 3: Inclusion probability matrix for 4-node and 20-node cycle when n=1000 and $\tilde{\rho}_{ij}$=0.45. Inclusion probabilities associated with true edges are in **bold**; inclusion probabilities that are 1.000 and are associated with other edges are in *italics*; Inclusion probabilities shown as 1 are 1.000, those shown as 0 are 0.000.

*	**1.000**	*0.890*	**0.768**
*	*	**0.902**	0.597
*	*	*	**0.993**
*	*	*	*

*	**1.0**	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0	0.0	0	0.0	0	0	0.0	0.0	**1**
*	*	**0.5**	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0	0	0	0	0	0.0	0.0	0.0	0.0
*	*	*	**0.8**	0.3	0.0	0.0	0	0	0	0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
*	*	*	*	**1.0**	0.0	0.0	0	0	0	0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
*	*	*	*	*	**0.9**	0.2	0.0	0	0	0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
*	*	*	*	*	*	**0.9**	0.0	0.0	0.0	0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
*	*	*	*	*	*	*	**1**	0.0	0	0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
*	*	*	*	*	*	*	*	**1.0**	0.0	0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
*	*	*	*	*	*	*	*	*	**1**	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
*	*	*	*	*	*	*	*	*	*	**1**	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
*	*	*	*	*	*	*	*	*	*	*	**0.1**	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
*	*	*	*	*	*	*	*	*	*	*	*	**1.0**	0.0	0	0.0	0	0	0.0	0
*	*	*	*	*	*	*	*	*	*	*	*	*	**1.0**	0.0	0	0	0	0	0
*	*	*	*	*	*	*	*	*	*	*	*	*	*	**1.0**	0.0	0.0	0	0	0
*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	**0.7**	0.3	0.0	0.0	0.0
*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	**1**	0.0	0	0.0
*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	**1**	0.0	0.0
*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	**1**	0.0
*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	**0.9**
*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*