Bayesian Analysis (2019) **14**, Number 4, pp. 1075–1090

Post-Processing Posteriors Over Precision Matrices to Produce Sparse Graph Estimates

Amir Bashir*, Carlos M. Carvalho†, P. Richard Hahn‡, and M. Beatrix Jones§

Abstract. A variety of computationally efficient Bayesian models for the covariance matrix of a multivariate Gaussian distribution are available. However, all produce a relatively dense estimate of the precision matrix, and are therefore unsatisfactory when one wishes to use the precision matrix to consider the conditional independence structure of the data. This paper considers the posterior predictive distribution of model fit for these covariance models. We then undertake post-processing of the Bayes point estimate for the precision matrix to produce a sparse model whose expected fit lies within the upper 95% of the posterior predictive distribution of fit. The impact of the method for selecting the zero elements of the precision matrix is evaluated. Good results were obtained using models that encouraged a sparse posterior (G-Wishart, Bayesian adaptive graphical lasso) and selection using credible intervals. We also find that this approach is easily extended to the problem of finding a sparse set of elements that differ across a set of precision matrices, a natural summary when a common set of variables is observed under multiple conditions. We illustrate our findings with moderate dimensional data examples from finance and metabolomics.

Keywords: covariance selection, decoupling shrinkage and selection, Gaussian graphical models, posterior summary, shrinkage prior.

1 Introduction

Covariance selection modelling (Dempster, 1973) performs model selection on the inverse covariance (precision) matrix. The model selected ought to fit well enough to be of use in predicting the future behaviour of the system. However, a sparse matrix is of particular interest because of its potential interpretability. For multivariate normal data, the precision matrix represents the conditional independence structure of the distribution, with a zero in the i, j position corresponding to independence between variables i and j, conditional on the rest of the variables in the system. Non-zero elements in the matrix therefore correspond to direct relationships that persist after accounting for other variables. This is often conceptualized as a graph, with nodes representing variables, and edges between them corresponding to non-zero elements of the precision matrix. These ideas can be extended to non-normal but continuous data using the nonparanormal approach of Liu et al. (2009).

*Massey University, Albany, Auckland, New Zealand 0745, a.bashir@massey.ac.nz
†McCombs School of Business, University of Texas, Austin, TX 78712, carlos.carvalho@mccombs.utexas.edu
‡Arizona State University, Tempe, AZ 85281, prhahn@asu.edu
§University of Auckland, Auckland, New Zealand 1142, beatrix.jones@auckland.ac.nz

 https://doi.org/10.1214/18-BA1139

Generating a posterior over graphs involves searching a large discrete space and remains computationally challenging, despite recent advances (Mohammadi and Wit, 2015). However, even for a posterior over sparse models, the Bayes estimate with respect to posterior predictive loss (the inverse of the average sampled covariance matrix) is not sparse. When we do not restrict ourselves to sampling from sparse graphs, other families of covariance models are possible: for example, the familiar conjugate inverse Wishart prior, the regularized inverse Wishart (Kundu et al., 2018), factor analysis models (West, 2003), or the Bayesian adaptive graphical lasso (Wang, 2012; Peterson et al., 2013).

In this paper we discuss how sparse estimates of the precision matrix Ω can be produced for any posterior over precision matrices. These procedures consist of three phases. First, a sample is generated from a posterior over precision matrices. Second, a series of progressively sparser estimates is created, typically indexed by some scalar criteria. Finally, a rule is applied to select a final estimate from among those generated. Several previous methods can be placed in this framework. In Kundu et al. (2018), the regularized inverse Wishart is used to generate posterior samples. Progressively stronger shrinkage of the implied regression coefficients is then used to generate increasingly sparse estimates and a local false discovery rate criteria is used to settle on the final estimate. In Wang (2012), the Bayesian adaptive graphical lasso (BAGL) posterior is used. Progressively sparser models are produced by considering the posterior mean of the elements of Ω (ω_{ij}) divided by their mean under an inverse Wishart distribution. Elements with small values are then set to zero. This is dubbed "ratio selection", and 0.5 is suggested as a suitable threshold for selection. Peterson et al. (2013) considers the BAGL posterior with two additional strategies for sparsification: setting to zero elements where the posterior mean has a small absolute partial correlation $|\rho|$, and setting to zero elements where a confidence interval for ω_{ij} includes zero. Suggested selection criteria are $|\rho| = 0.1$ and a 90% confidence interval respectively. We can also view traditional selection of graphical models in this framework, with posterior probability of edge inclusion as the sparsification criteria, and thresholds like 0.5 (the median probability model) as typical selection rules.

Hahn and Carvalho (2015) proposed a novel strategy for the final phase—selecting the sparse estimate—in the context of regression. The key element of the approach, called decoupled shrinkage and selection (DSS), is a selection rule based on the posterior predictive distribution of the fit to future data. This distribution provides the relevant scale on which to consider whether sparsified versions of a model provide adequate fit, and makes explicit the tradeoff between fit and sparsity. Hahn and Carvalho (2015) contains only preliminary suggestions about the use of DSS in the context of precision matrices. In this paper, we explore the procedure in detail, and extend it to the modelling of sparse differences across a set of precision matrices on the same variables, observed under different conditions (for example, case-control differences). The next section provides an initial example to illustrate the procedure. We then evaluate how the DSS selection rule interacts with a range of sparsification strategies, and explore its use with different posteriors over Ω. The extension to the multi-condition case follows. We evaluate the performance of the DSS algorithm for this task and illustrate it with a moderate dimensional example (174 variables).

2 Decoupled shrinkage and selection for graphs

Let Γ be our choice as an estimate of $\Omega = \Sigma^{-1}$. Predictive accuracy, which we will call *fit*, is typically measured by the log likelihood of n^* future observations $\tilde{X}$. The expected value of $\tilde{X}^T\tilde{X}/n^*$ is the posterior mean of the covariance matrix, $\bar{\Sigma}$. The expected fit then becomes:

$$\begin{aligned} \mathrm{E}\left[\mathrm{fit}(\Gamma)\right] &= \mathrm{E}\left[\log\det(\Gamma) - \mathrm{tr}\left(\frac{\tilde{X}^T\tilde{X}\Gamma}{n^*}\right)\right] \\ &= \log\det(\Gamma) - \mathrm{tr}(\bar{\Sigma}\Gamma). \end{aligned}$$

The expected fit is naturally maximised at $\Gamma = \bar{\Sigma}^{-1}$.

Crucial to our approach is recognising that this is an expected fit, and $\mathrm{fit}(\Gamma) = \log\det(\Gamma)) - \mathrm{tr}(\tilde{X}^T\tilde{X}\Gamma/n^*))$ is a random variable governed by the posterior predictive distribution of $\tilde{X}$. We can examine a sample of realisations of this random variable. To avoid considering n^* we imagine it to be very, very large. Samples of $\tilde{X}^T\tilde{X}/n*$ are then equivalent to posterior samples Σ_k of the covariance matrix. Σ_k thus plays the role of future data in the expression below. A sample from the distribution of $\mathrm{fit}(\bar{\Sigma}^{-1})$ can be generated as

$$\mathrm{fit}(\bar{\Sigma}^{-1}|\Sigma_k) = \log(\det(\bar{\Sigma}^{-1})) - \mathrm{tr}(\Sigma_k\bar{\Sigma}^{-1}), \tag{1}$$

for $k \in 1, 2 \ldots, m$. A histogram of these samples provides an estimate of the distribution of $\mathrm{fit}(\bar{\Sigma}^{-1})$. We believe the scale of this distribution is relevant for judging acceptable fit. If the estimate with the best expected fit will have actual fit below F, say, 5% of the time, a sparsified choice for Γ with expected fit F should be considered adequate. We will use the (arbitrary) criteria of 5%, but this remains a decision of the user. Because this is the quantile of a distribution, the choice is relatively intuitive and interpretable. Finding the globally sparsest Γ that satisfies this criteria remains a challenging optimization problem. In the subsequent sections, we examine several heuristic approaches to this task. However, we first work through a simple example to illustrate how the procedure works.

We consider a multivariate dataset where measurements of 174 volatile compounds were obtained using mass spectrometry for fecal samples from 49 subjects (Jayan, 2016). We generate our sample of Σ_k from the BAGL posterior, using the Matlab code of Wang (2012). The name "Bayesian (adaptive) graphical lasso" arises from the fact that posterior mode under a fixed Laplace(λ) prior on all off diagonal elements of Ω corresponds to the graphical lasso estimate. However, in the Bayesian treatment, a hyper prior is placed on λ leading to a generalised Pareto prior. In the adaptive version, independent λ_{ij} from this hyper prior are used for each off-diagonal element. The distribution of the ω_{ij} is then constrained to the space of positive definite matrices. The posterior samples of Ω produced with this algorithm are not sparse. For this example, we will create a sparse model by constraining to zero all ω_{ij} whose P% credible interval includes zero. Increasing P then generates a sequence of increasingly sparse models. Once the set of posterior samples Σ_k (and corresponding $\{\Omega_k = \Sigma_k^{-1}\}$) is in hand, the procedure is as follows:

1. Compute the mean of the Σ_k, the posterior mean of the covariance, denoted $\bar{\Sigma}$. The inverse of this matrix, $\bar{\Sigma}^{-1}$ is the estimate of Ω that maximizes the expected fit to future data, and is not sparse.

2. Now let the sampled Σ_k play the role of future data, and compute the fit of $\bar{\Sigma}^{-1}$ for each future dataset, (fit$(\bar{\Sigma}^{-1}|\Sigma_k)$ from equation (1). Find the 5% quantile of these fits.

3. Compute P% credible intervals for the ω_{ij} using their estimated distribution based on the sampled Ω_k.

4. For elements where this interval includes zero, constrain Γ_P to be zero in this location.

5. Find the Γ_P that maximizes $\log\det(\Gamma_P) - \text{tr}(\bar{\Sigma}\Gamma_P)$.

6. If $\text{fit}(\Gamma|\bar{\Sigma}) = \log\det(\Gamma_P) - \text{tr}(\bar{\Sigma}\Gamma_P)$ is larger than the 5% quantile of $\text{fit}(\bar{\Sigma}^{-1}|\Sigma_k)$, and $P < 100\%$, increase P and return to step 3.

We then use the penultimate Γ generated, the sparsest estimate with adequate fit. The fit and sparsity of the sequence of Γ generated for the fecal volatilome data are shown in Figure 1.

Note that Step (5) is done using the graphical lasso algorithm from Friedman et al. (2008). This was created for optimizing the L1 penalized likelihood:

$$\max_{\Gamma} \left[\log\det\Gamma - tr(\Gamma S) - \lambda \sum_{i \neq j} |\gamma_{ij}| \right]. \tag{2}$$

When S is positive definite, the final stage of this algorithm can be used with $\lambda = 0$ to find the best fitting Γ that follows a specified zero pattern. We will most frequently use the algorithm in this mode, taking $S = \bar{\Sigma}$. The posterior distributions we consider are over positive definite matrices, so $\bar{\Sigma}$ is always positive definite. We also use the graphical lasso in its conventional form, with cross validation to pick the penalty parameter, to generate the starting point for the BAGL Markov chain Monte Carlo algorithm.

3 Comparison of selection strategies

Sparsification based on progressively larger confidence intervals is, of course, not the only possible method of selecting the zero elements of Γ. Any strategy with an ordered scalar criteria that can be "turned up" to increase sparsity can be used to produce a plot like that in Figure 1, and combined with the chosen fit quantile to produce an estimate. Far from being an afterthought to the generation of a posterior distribution, we find that which sparsification strategy is used has a strong influence over the inferred model, and particularly its level of sparsity. For the volatilome data, we compare three additional sparsification strategies to the credible interval strategy employed in Figure 1. These strategies are: 1) 'ratio selection'—taking elements to be zero if their posterior mean is less than K times the value obtained using the conjugate Wishart prior (as suggested in Wang, 2012), 2) thresholding the absolute partial correlations, and 3) choosing non-zero elements based on the (frequentist) adaptive graphical lasso (Fan et al., 2009).

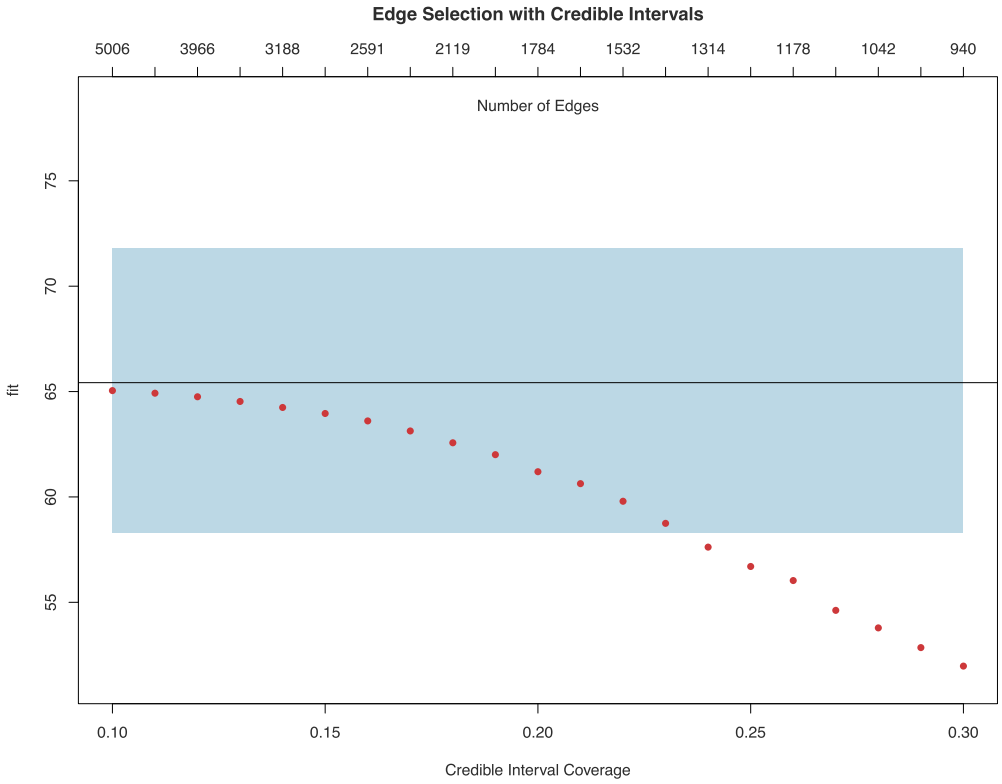

Figure 1: Expected fit for sparse summaries of the precision matrix for 174 volatile compounds measured for 49 individuals. The x axis shows the coverage of the credible interval used to select zero elements for Γ. The top axis shows the resulting number of edges in the graph. The 90% credible interval for the fit of $\bar{\Sigma}^{-1}$ (blue region), and its expected fit (central line) are shown for comparison.

The adaptive graphical lasso modifies the criteria in (2) to find

$$\Gamma_\lambda = \max_{\Gamma} \left[\log\det(\Gamma) - \mathrm{tr}(\Gamma S) - \lambda \sum_{i \neq j} \frac{|\gamma_{ij}|}{\sqrt{\gamma^*_{ij}}} \right], \tag{3}$$

where γ^*_{ij} is an initial estimate of the absolute precision matrix elements. This effectively creates a variable penalty that reduces over-penalization of edges that are clearly non-zero. We again take $S = \bar{\Sigma}$. We use γ^*_{ij} based on the conventional graphical lasso, with the shrinkage parameter chosen by cross validation. In practice, for each element we use the larger of γ^*_{ij} and 0.00001 to avoid numerical instability. The series of increasingly sparse estimates is produced by using increasing values of λ.

We note that for all three strategies only steps (3) and (4) of the algorithm in section 2 are changed. In particular, the penalized estimate produced by the adaptive graphical

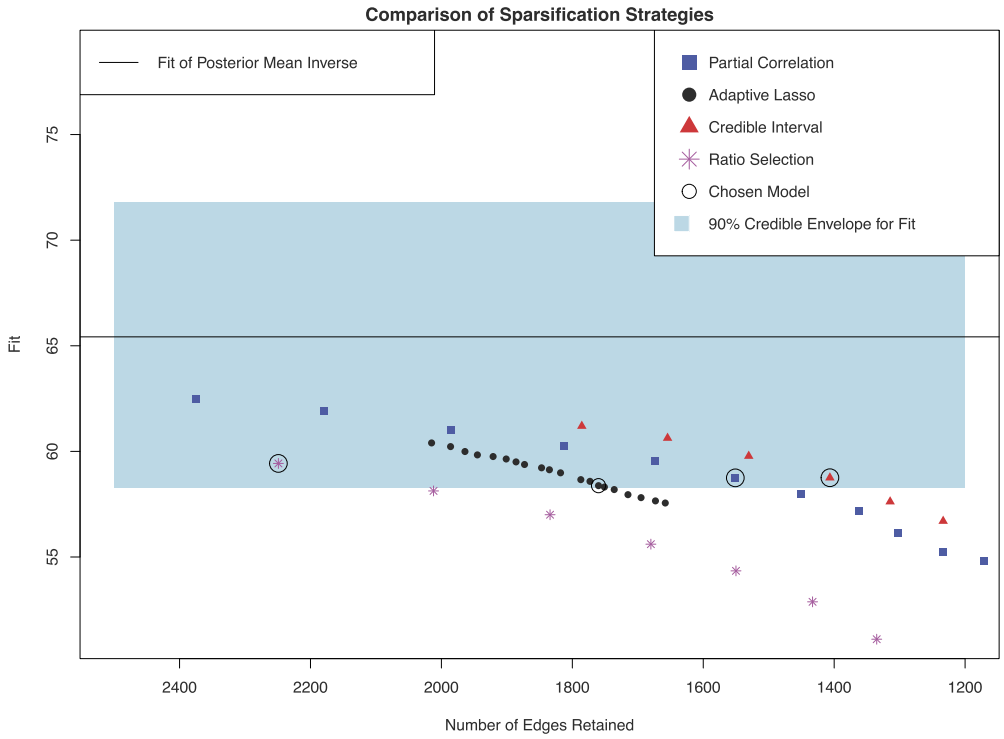

Figure 2: Expected fit for sparse summaries of the precision matrix for 174 volatile compounds measured for 49 individuals, using different selection criteria. The 90% credible interval for the fit of $\bar{\Sigma}^{-1}$ (blue region), and its expected fit (central line) are shown for comparison.

lasso is not used to assess the fit. Instead, Γ is refit using the zero structure of the adaptive graphical lasso estimated matrix, but with no penalization of the non-zero elements.

The sequences of estimates produced can be seen in Figure 2, with further information on the selected models in Table 1. By design, the fit for each of the selected models is approximately the same. The sparsity of these models differs substantially, with the sparsest model produced by the credible interval method. We note that the thresholds chosen by the decoupled shrinkage and selection criteria in some cases differ markedly from the "intuitive" choices for the various selection methods used in Wang (2012) and Peterson et al. (2013): K=0.5 for ratio selection, $|\rho| = 0.1$ for partial correlations, and 90% for the credible interval. To produce values which are in the fit envelope, much less stringent criteria must be used in the partial correlation and credible interval methods.

To get a more comprehensive picture of how the different criteria behave in different situations, we simulate multivariate normal data from two different models with sparse structures used in Wang (2012). The first is a second order autoregressive, or AR(2),

Approach	Criteria to retain	Number of edges	E(fit)
Bayes estimate ($\bar{\Sigma}^{-1}$)	-	15051	65.4
Ratio selection	$K > 0.45$	2212	59.4
Adaptive graphical lasso	$\lambda > 6.8 \times 10^{-7}$	1760	58.4
Partial correlation threshold	$\lvert\rho\rvert > 0.015$	1551	58.7
Credible interval	23% credible interval	1407	58.8

Table 1: Comparison of sparsification strategies for the volatilome data. For each strategy, the criteria corresponding to the sparsest model inside the top 95% of fits is given, as well as the number of edges of that model. By design, the expected fit of each selected model should be approximately the same, and the E(fit) column confirms this. The expected fit of the Bayes estimate is also given for comparison.

process with the following inverse covariance matrix:

$$\begin{pmatrix} 1.00 & 0.50 & 0.25 & 0 & 0 & 0 & \dots & 0 \\ 0.50 & 1.00 & 0.50 & 0.25 & 0 & 0 & \dots & 0 \\ 0.25 & 0.50 & 1.00 & 0.50 & 0.25 & 0 & \dots & 0 \\ 0 & 0.25 & 0.50 & 1.00 & 0.50 & 0.25 & \dots & 0 \\ \vdots & & \ddots & \ddots & \ddots & \ddots & \ddots & \vdots \\ 0 & \dots & 0 & 0.25 & 0.50 & 1 & 0.50 & 0.25 \\ 0 & \dots & 0 & 0 & 0.25 & 0.50 & 1 & 0.50 \\ 0 & \dots & 0 & 0 & 0 & 0.25 & 0.50 & 1 \end{pmatrix}. \tag{4}$$

Second, a star shaped model is used, with a single hub (variable 1) connected to each of the others. The inverse covariance has ones on the diagonal, and 0.1 for the non-zero off diagonal elements. For number of variables $p = 100$, situations with $n = 100, 200$ and 300 are considered. A smaller sample size, $n = 50$, is also considered with $p = 30$. For each combination of sample size and model, 50 replicates are performed. A sample from a BAGL posterior is generated, and the most promising sparsification strategies are used to select models, using sparsest model found with fit above the 5% fit quantile. The strategies compared are adaptive graphical lasso, credible intervals, and thresholding the partial correlation ρ. Results are in Table 2. The credible interval method is generally better for the AR(2) structure, and the adaptive graphical lasso for the star structure, which has smaller off diagonal elements. Thresholding the partial correlations is competitive when considering the AR(2) structure, but clearly worst for the star structure.

4 Effect of the input posterior

The procedure in Section 2 can be applied to any posterior over precision matrices. In this section we illustrate its use with the Wishart prior, the priors associated with Bayesian factor analysis, and the G-Wishart prior.

The Wishart prior is conjugate for the precision matrix. For a p variable system, prior parameters δ and Φ lead to posterior

$$\Omega \sim W(\Phi + X^t X, \delta + n).$$

	AR(2)			Star		
	Sensitivity	Specificity	MCC	Sensitivity	Specificity	MCC
$n = 50$, $p = 30$						
A. graphical lasso	25.6	99.7	45.3	0.0	100.0	0.0
Credible interval	29.0	100.0	51.0	0.0	100.0	1.0
ρ threshold	28.0	100.0	50.0	0.1	100.0	0.3
$n = 100$, $p = 100$						
A. graphical lasso	77.0	98.0	67.2	99.9	97.5	66.6
Credible interval	84.4	98.4	75.3	73.9	94.4	37.6
ρ threshold	87.1	98.5	77.7	47.9	92.8	21.1
$n = 200$, $p = 100$						
A. graphical lasso	95.1	99.3	89.6	100.0	98.9	80.9
Credible Interval	97.9	99.5	93.4	90.0	97.0	57.0
ρ threshold	98.3	99.6	94.6	78.0	94.0	39.0
$n = 300$, $p = 100$						
A. graphical lasso	98.4	99.7	96.0	100.0	99.8	96.1
Credible interval	99.4	99.9	98.4	96.7	98.5	74.0
ρ threshold	99.5	99.9	98.7	92.0	96.5	55.7

Table 2: Multivariate normal data was simulated from two different model structures, AR(2) and Star, for each of the p, n combinations given. BAGL posterior samples were generated, and three different sparsification strategies applied, to select the sparsest model with fit above the 5% fit quantile. Fifty replicates are performed. Average sensitivity, specificity, and Matthews correlation coefficient (MCC) for each scenario are given as percentages.

The covariance Σ then has mean

$$\frac{\Phi + X^t X}{\delta + n + p - 1}.$$

It is conventional to take Φ as a diagonal matrix with the observed sample variances (or the identity for standardised data), and δ as small as possible. We use this strategy for Φ, but take a slightly different approach for δ, choosing δ so that the posterior mean will match the optimal estimator from Ledoit and Wolf (2004), which also takes the form of a weighted average of the data and a target matrix.

Bayesian factor analysis (BFA) assumes the data arises from a set of independent, Gaussian distributed factors:

$$X_i \sim \Lambda\eta_i + \epsilon_i,\ \eta \sim N(0, I),\ \epsilon \sim N(0, T),\ T \text{ diagonal}.$$

Many choices are possible for priors on Λ. We choose a mixture of normal and point mass priors on the λ_{ij}, with mixing proportion beta(1/3, 1/3) (the default in the R package `bfa`, Murray, 2016). We set the number of factors equal to $p/2$.

Finally, we use the G-Wishart approach implemented in the `BDgraph` package (Mohammadi and Wit, 2015). This packages uses a birth-death (BD) algorithm to sample

Method	Edges	Performance on hold-out data % variance unexplained	log likelihood
Wishart	1711	17.6	82
Wishart-selection	486	17.5	81
BFA	1711	18.1	78
BFA-selection	384	19.5	74
BAGL	1711	16.1	98
BAGL-selection	184	17.1	96
G-Wishart(0.5)	1711	16.1	107
G-Wishart(0.5)-selection	315	15.7	105
G-Wishart($2/p$)	1711	16.6	101
G-Wishart($2/p$)-selection	92	20.7	89

Table 3: Models fit to the first 60 months of the mutual fund data. Performance criterial were evaluated on a hold out set consisting of the subsequent 26 months of data. For each different prior, both $\Gamma = \bar{\Sigma}^{-1}$ and Γ produced by sparsifying $\bar{\Sigma}^{-1}$ are considered. In each case, the average proportion of unexplained variation and log likelihood evaluated for the hold-out set are given.

a posterior over G, but we make use of the additional package functions to then sample $\Omega|G$. Conditional on G, Ω has a G-Wishart distribution (Roverato, 2002; Letac and Massam, 2007). This distribution is a Wishart conditioned on elements not corresponding to edges in G being equal to zero. We choose the prior Wishart parameters $\Phi = I$ and $\delta = 3$. In the `BDgraph` framework, we must also specify a prior over G. In this prior, each edge is included with prior probability π. We use two different settings, $\pi = 0.5$, which favors graphs with roughly half the possible edges, and $\pi = 2/p$, which favors very sparse graphs.

We wish to judge the performance of the different posterior distributions, and do this using a set of test data, after a set of training data has been used to fit the models. We evaluate both the log likelihood of the test data, and the average proportion of unexplained variability (error sum of squares over total sum of squares), where for each test observation we predict each individual variable, assuming the other variables in that observation have already been observed. As an example we consider the mutual fund data from Scott and Carvalho (2008) and Fitch et al. (2014). The data consist of 59 variables, with $n = 60$ for the training set and $n = 26$ for the test set. We generate posteriors over the precision matrix using BAGL, the Wishart, BFA, and the G-Wishart with graph prior parameters 0.5 and $2/p$. In each case we compute the log likelihood and average proportion of unexplained variability based on $\bar{\Sigma}^{-1}$, and on its sparsified version. Sparsification is done using the credible interval method. Results are given in Table 3.

The two criteria, log likelihood and percent unexplained variability, produce a somewhat different picture. There is very little difference in the percent unexplained variability across most of the models, while the log likelihood varies more dramatically. This is because the unexplained variability for each variable is computed assuming all other variables have been observed. An observation that is far from the multivariate mean relative to the inferred variability will have low log likelihood, but potentially each uni-

variate component of that observation is consistent with its conditional distribution. In other words, each variable has most of its deviation from the mean explained by the other variables. This will occur when the deviations from the mean across variables are consistent with the inferred covariance matrix. Thus the Wishart method, the fastest of the methods considered, is competitive in the percent of unexplained variability, and about 72% of the edges in the model can be removed with minimal loss of fit. Sparsified BAGL results in the sparsest estimate that manages reasonable performance in percent variance unexplained. We again note that the BAGL posterior samples contain no exact zeros. The G-Wishart(0.5) posterior mean produced the best results for the log likelihood of the test data. Sparsification of this estimate had little impact on the log likelihood, and slightly improved the percent of variability explained. We note that the sparsified graph is in fact sparser than the MAP graph from the sample generated, which had 752 edges. While the G-Wishart($2/p$) mean performs well, the selected graph performs poorly in prediction (but is very sparse). This suggests that when using the G-Wishart, selecting a prior probability for edge inclusion "large enough" followed by post processing to induce sparsity may be a more straightforward strategy than tuning the prior probability of edge inclusion.

5 Identifying differences in Ω across conditions

Considerable attention has been paid to identifying changes in the precision matrix under C different conditions (Danaher et al., 2013; Cai et al., 2016; Guo et al., 2011; Zhao et al., 2014; Tian et al., 2016), including Bayesian approaches (Mitra et al., 2016; Peterson et al., 2015). The concepts introduced here can also address this situation. We emphasise estimating common elements as well as common structures: in other words, elements of the condition specific precision matrices Ω_c and $\Omega_{c'}$ where $\omega_{cij} = \omega_{c'ij} \neq 0$. Rather than directly modelling any shared characteristics between the matrices, we generate independent posteriors using the data from each condition. A sample from the joint posterior is formed simply by combining the first posterior sample for each group, second samples, and so on. The fit of a particular estimate is the sum of the fits, weighted by the sample size of each group. We then subject the $\bar{\Sigma}_c^{-1}$ to modification that makes some elements identical. The fit of these modified estimates, relative to the Bayes point estimates, is then judged as in the preceding sections.

To modify the $\bar{\Sigma}_c^{-1}$, we use a variation of the fused Joint Graphical Lasso (JGL), described in Danaher et al. (2013). The original algorithm uses a penalized likelihood of the form:

$$\max_{\Gamma_C} \left[\sum_{c=1}^{C} n_c \left(\log\det \Gamma_c - \mathrm{tr}(S_c \Gamma_c) \right) + \lambda_1 \sum_{c=1}^{C} \sum_{i \neq j} |\gamma_{cij}| + \lambda_2 \sum_{c<c'} \sum_{i,j} |\gamma_{cij} - \gamma_{c'ij}| \right],$$

where the λ_1 governs the sparsity of the estimates, and λ_2 governs their similarity. This function is optimized using the alternating direction method of multipliers (ADMM) algorithm as outlined in Danaher et al. (2013). The L1 penalties can produce exact equality between off diagonal elements of the C precision matrices, but this happens

only when penalization is quite strong. In fact, when starting with matrices that were already identical at many positions, and moderate penalty parameters, the optimization reduces the number of identical elements. Danaher et al. (2013) note that many conventional criteria for choosing penalty parameters (Akaike or Bayesian information criteria, cross validation) produce overly dense models, and suggest penalty selection be guided by "practical considerations".

As well as using $\bar{\Sigma}_c$ rather than S, we change the penalization strategy. If sparse matrices are desired, an adaptive penalty analogous to (3) is used. For penalization of element differences across conditions, we consider only the choice of elements to be identical across conditions. Suppose the selected elements are in a set $\mathcal{H}$ (the strategy for choosing this set is discussed below). Our objective function becomes:

$$\max_{\Gamma_C} \left[\sum_{c=1}^{C} n_c \left(\log\det \Gamma_c - \mathrm{tr}(\bar{\Sigma}_c \Gamma_c) \right) + \lambda_2 \sum_{c<c'} \sum_{i,j \in \mathcal{H}} |\gamma_{cij} - \gamma_{c'ij}| \right],$$

with λ_2 taken large enough that $|\gamma_{cij} - \gamma_{c'ij}|$ is forced to zero for $(i,j) \in \mathcal{H}$. Optimization is again via an ADMM algorithm. We use the posteriors of the Ω_c to select the elements of $\mathcal{H}$. Specifically, if a P% credible interval for $\gamma_{cij} - \gamma_{c'ij}$ includes zero, $(i,j)_{c,c'}$ is included in $\mathcal{H}_{\mathcal{P}}$. P is increased to generate increasingly sparse sets of differences. No effort is made to sparsify the original graphs. Instead, the entire fit budget is spent increasing the number of common elements across conditions. Alternate strategies are possible if both sparse matrices and sparse matrix differences are desired.

We also note in our case the n_c are tuning parameters. Smaller values of n_c increase the flexibility of Γ_c to vary from $\bar{\Sigma}_c^{-1}$ relative to other conditions. We use n_c proportional to the sample sizes from each group. This should be a reasonable choice as sample sizes typically will indicate the concentration of the posterior, but other strategies would be possible.

To evaluate the performance of this procedure, we consider detecting differences between pairs of Ω matrices, where the true set of differences is sparse. In each case, $p = 100$, $n = 200$, and a random selection of 50 edges differs between the pair, with $|\gamma_{cij} - \gamma_{c'ij}| = 0.1$. We note that, based on Table 2 of Danaher et al. (2013), this is a challenging sample size for detection of differential edges with JGL. Two different scenarios are considered: a case where the matrices are sparse, and one where they are dense (but still with a sparse set of differences). In the sparse case, one matrix has the AR(2) structure specified in (4), with the second matrix differing at 50 random elements. In the dense case, the initial inverse covariance matrix has ones on the diagonal and 0.05 at all off diagonal elements. For each scenario, 50 replicates are performed. The posterior distributions are generated with BAGL.

The results are compared to results from JGL in Table 4. To provide a fair comparison of accuracy, JGL is tuned to produce approximately the same number of differences as our algorithm. The JGL solution is found for a fine grid of λ values, and the one with number of selected differences closest to the one found with our method is chosen. Both λ_1 and λ_2 are varied in the sparse scenario, while λ_1 is set to zero in the dense scenario

	Sparse Case		Dense Case	
	Posterior Summary	JGL	Posterior Summary	JGL
Sensitivity	58.6	42.6	8.9	8.1
Specificity	97.9	97.8	98.9	98.9
MCC	35.5	25.4	7.1	6.4

Table 4: Sensitivity, specificity, and Matthews correlation coefficient for our procedure for detecting precision matrix differences, averaged over 50 replicates of simulated data with $p = 100$, $n = 200$. In each case there are 50 off-diagonal differences between the matrices being compared, with magnitude 0.1. The matrix before changes is either an AR(2) structure (sparse case), or one with diagonal 1 and all off-diagonals 0.05. We compare to the JGL inferred differences, with λ selected to match the number of edges detected.

and only λ_2 varied. The difference between JGL and the posterior summary method in number of differences selected was at most two across the individual pairs of matrices. The average number of differences is identical between the two methods.

The posterior summary method gives improved sensitivity in the sparse case, consistently across replicates (the standard deviation of the 16% of improvement in sensitivity was 6%). The dense case was more challenging, and performance of the two algorithms was comparable. However, we note JGL as described in Danaher et al. (2013) lacks a criteria for selecting the relevant λ.

As a higher dimensional example we compare the precision of the 49 fecal volatilome measurements from Section 2, a set of control measurements, to the precision for a set of 42 cases. (In this example, the control individuals are 8 year old children born at term, and the cases are 8 year old children that were born pre-term.) We base our inferences on posteriors independently generated from the Bayesian adaptive graphical lasso. The graph of differential edges is shown in Figure 3, and represents 499 differences. Most vertices have differences for 1–10 of their incident edges. The figure highlights 14 vertices that have differences at more than 10 of their incident edges. There are also two vertices that are not involved in any altered edges.

6 Discussion

This paper introduces an extremely flexible method for producing sparse summaries of a posterior over precision matrices, as well as an extension of the method to a case where sparse differences across a set of precision matrices is desired. Starting from a posterior distribution over the precision is key to the method in several ways. The posterior provides the scale upon which degradation of the fit is judged. The convergence properties of the algorithms used to produce an estimate with a particular pattern of sparsity rely on the input of a positive definite covariance matrix, which the posterior mean will satisfy. Finally, the strategy for sparsification can also use information from the posterior beyond the posterior mean. Using posterior credible intervals for each element to choose the pattern of sparsity is a successful strategy in

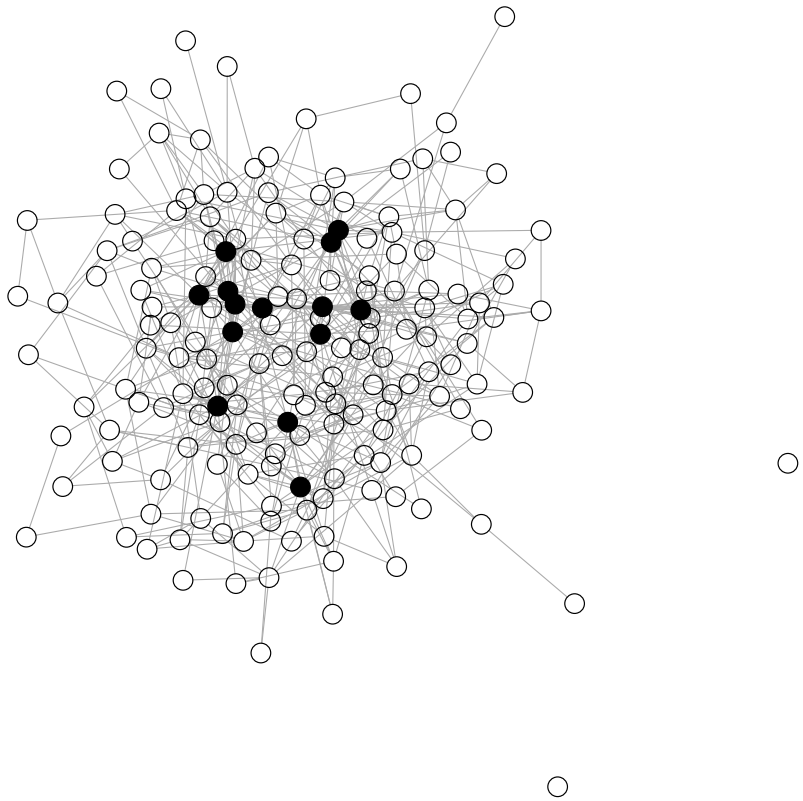

Figure 3: Graph with edges corresponding to our inferred differential precision matrix elements for 174 volatile compounds measured for 42 cases and 49 controls. Vertices with more than 10 altered matrix elements are shown in black.

our simulation study, although the structure suggested by application of the adaptive graphical lasso to $\bar{\Sigma}$ may be preferred when sensitivity to small non-zero ω_{ij} is a priority.

The starting posterior can be generated from a range of models. In particular, the posterior need not be over sparse matrices. When the true precision is sparse incorporating this information in the prior is clearly useful. However, this might be in the form of a prior that encourages off diagonal elements to be close to zero rather than exactly zero (for example BAGL). One possibility for computationally tractable sparse models that we have not explored is decomposable models (Green and Thomas, 2013). In a data rich situation these algorithms typically add edges to the model to make it decomposable (Fitch et al., 2014), so we might expect similar results to the G-Wishart(0.5) where there is no pressure for the posterior sample to visit very sparse models, and the sparsified model actually outperforms $\bar{\Sigma}^{-1}$ on some criteria. In our example the "non-sparse" priors (Wishart or BFA) produce models with more edges, but a great deal of sparsity can still be introduced. These priors are appealing because they are tractable even in

relatively high dimensions. Other computationally tractable priors in Huang and Wand (2013) and Kundu et al. (2018) are also potentially useful.

Finally, we demonstrate how our approach can be used to understand differences in precision matrices across conditions. There is not a Bayesian method that addresses the problem of inferring which elements are identical, but non-zero, across such a set of precision matrices (although there are certainly methods that aim to "share strength" across conditions). A consequence of this is that our method is suitable for detecting sparse differences between precision matrices where the individual matrices are dense, although our simulation study indicates this is a challenging problem at moderate samples sizes. While frequentist approaches readily produce a series of increasingly sparse differences, they do not provide a clear criteria for selection. Our method for comparing precision matrices has the further advantage that any available innovation for Bayesian modelling of a single precision matrix, whether a new prior or a computational improvement, is immediately extendable to the multi-matrix case, rather than requiring a separate implementation. We envision that even if detailed modelling of similar elements, as in Peterson et al. (2015), will eventually be carried out, our method will be a useful preliminary to investigate modelling choices.

References

Cai, T. T., Li, H., Liu, W., and Xie, J. (2016). "Joint estimation of multiple high dimensional precision matrices." *Statistica Sinica*, 26(2): 445–464. MR3497754. 1084

Danaher, P., Wang, P., and Witten, D. M. (2013). "The joint graphical lasso for inverse covariance estimation across multiple classes." *Journal of the Royal Statistical Society Series B–Statistical Methodology*, 76(2): 373–397. MR3164871. doi: https://doi.org/10.1111/rssb.12033. 1084, 1085, 1086

Dempster, A. P. (1973). "Covariance selection." *Biometrics*, 21(1): 157–175. 1075

Fan, J., Feng, Y., and Wu, Y. (2009). "Network exploration via the adaptive LASSO and SCAD penalties." *Annals of Applied Statistics*, 3(2): 521–541. MR2750671. doi: https://doi.org/10.1214/08-AOAS215. 1078

Fitch, A. M., Jones, M. B., and Massam, H. (2014). "The performance of covariance selection methods that consider decomposable models only." *Bayesian Analysis*, 9(3): 659–684. MR3256059. doi: https://doi.org/10.1214/14-BA874. 1083, 1087

Friedman, J., Hastie, T., and Tibshirani, R. (2008). "Sparse inverse covariance estimation with the graphical lasso." *Biostatistics*, 9: 432–441. 1078

Green, P. and Thomas, A. (2013). "Sampling decomposable graphs using a Markov chain on junction trees." *Biometrika*, 100(1): 91–110. MR3034326. doi: https://doi.org/10.1093/biomet/ass052. 1087

Guo, J., Levina, E., Michailidis, G., and Zhu, J. (2011). "Joint estimation of multiple graphical models." *Biometrika*, 98(1): 1–15. MR2804206. doi: https://doi.org/10.1093/biomet/asq060. 1084

Hahn, P. R. and Carvalho, C. M. (2015). "Decoupling shrinkage and selection in Bayesian linear models: a posterior summary perspective." *Journal of the American Statistical Association*, 110: 435–448. MR3338514. doi: https://doi.org/10.1080/01621459.2014.993077. 1076

Huang, A. and Wand, M. P. (2013). "Simple marginally noninformative prior distributions for covariance matrices." *Bayesian Analysis*, 8(2): 439–452. MR3066948. doi: https://doi.org/10.1214/13-BA815. 1088

Jayan, S. (2016). "Analysis of the metabolic profiles from faeces and plasma of children born very preterm." Master's thesis, University of Auckland. 1077

Kundu, S., Mallick, B. K., and Baladandayuthapani, V. (2018). "Efficient Bayesian regularization for graphical model selection." *Bayesian Analysis*. Advance publication. 1076, 1088

Ledoit, O. and Wolf, M. (2004). "A well-conditioned estimator for large-dimensional covariance matrices." *Journal of Multivariate Analysis*, 88(2): 365–411. MR2026339. doi: https://doi.org/10.1016/S0047-259X(03)00096-4. 1082

Letac, G. and Massam, H. (2007). "Wishart distributions for decomposable graphs." *Annals of Statistics*, 35(3): 1278–1323. MR2341706. doi: https://doi.org/10.1214/009053606000001235. 1083

Liu, H., Lafferty, J., and Wasserman, L. (2009). "The nonparanormal: semiparametric estimation of high dimensional undirected graphs." *Journal of Machine Learning Research*, 10: 2295–2328. MR2563983. 1075

Mitra, R., Müller, P., and Ji, Y. (2016). "Bayesian graphical models for differential pathways." *Bayesian Analysis*, 11(1): 99–124. MR3447093. doi: https://doi.org/10.1214/14-BA931. 1084

Mohammadi, A. and Wit, E. (2015). "Bayesian structure learning in sparse Gaussian graphical models." *Bayesian Analysis*, 10(1): 109–138. MR3420899. doi: https://doi.org/10.1214/14-BA889. 1076, 1082

Murray, J. (2016). "bfa: Bayesian factor analysis." R package version 0.4. 1082

Peterson, C. B., Stingo, F. C., and Vannucci, M. (2015). "Bayesian inference of multiple Gaussian graphical models." *Journal of the American Statistical Association*, 110(509): 159–174. MR3338494. doi: https://doi.org/10.1080/01621459.2014.896806. 1084, 1088

Peterson, C. B., Vannucci, M., Karakas, C., Choi, W., Ma, L., and Maletic-Savatic, M. (2013). "Inferring metabolic networks using the Bayesian adaptive graphical lasso with informative priors." *Statistics and its Interface*, 6(4): 437–558. MR3164658. doi: https://doi.org/10.4310/SII.2013.v6.n4.a12. 1076, 1080

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(3): 391–411. MR1925566. doi: https://doi.org/10.1111/1467-9469.00297. 1083

Scott, J. and Carvalho, C. (2008). "Feature inclusion stochastic search for Gaussian graphical models." *Journal of Computational and Graphical Statistics*, 17(4): 790–808. MR2649067. doi: https://doi.org/10.1198/106186008X382683. 1083

Tian, D., Gu, Q., and Ma, J. (2016). "Identifying gene regulatory network rewiring using latent differential graphical models." *Nucleic Acids Research*, 44(17): e140. 1084

Wang, H. (2012). "Bayesian graphical lasso models and efficient posterior computation." *Bayesian Analysis*, 7: 867–886. MR3000017. doi: https://doi.org/10.1214/12-BA729. 1076, 1077, 1078, 1080

West, M. (2003). "Bayesian factor regression models in the large p, small n paradigm." In Bernardo, J., Bayarri, M., Berger, J., Dawid, A., Heckerman, D., Smith, A. F. M., and West, M. (eds.), *Bayesian Statistics.*, volume 7. Oxford: Oxford University Press. MR2003537. 1076

Zhao, S. D., Cai, T. T., and Li, H. (2014). "Direct estimation of differential networks." *Biometrika*, 101(2): 253–268. MR3215346. doi: https://doi.org/10.1093/biomet/asu009. 1084