Adaptive Sparsity in Gaussian Graphical Models

Eleanor Wong ELESWONG@SCI.UTAH.EDU
Suyash P. Awate SUYASH@SCI.UTAH.EDU
P. Thomas Fletcher FLETCHER@SCI.UTAH.EDU
Scientific Computing and Imaging Institute, 72 S Central Campus Drive, Room 3750, Salt Lake City, UT 84112

Abstract

An effective approach to structure learning and parameter estimation for Gaussian graphical models is to impose a sparsity prior, such as a Laplace prior, on the entries of the precision matrix. Such an approach involves a hyperparameter that must be tuned to control the amount of sparsity. In this paper, we introduce a parameter-free method for estimating a precision matrix with sparsity that adapts to the data automatically. We achieve this by formulating a hierarchical Bayesian model of the precision matrix with a non-informative Jeffreys' hyperprior. We also naturally enforce the symmetry and positive-definiteness constraints on the precision matrix by parameterizing it with the Cholesky decomposition. Experiments on simulated and real (cell signaling) data demonstrate that the proposed approach not only automatically adapts the sparsity of the model, but it also results in improved estimates of the precision matrix compared to the Laplace prior model with sparsity parameter chosen by cross-validation.

1. Introduction

The structure learning problem for Markov random fields is to infer the graph structure of a model, i.e., the conditional dependencies between the random variables, from observed data. In the case of Gaussian graphical models (GGMs) (Rue & Held, 2005), this is equivalent to estimating which entries of the precision matrix are nonzero. These nonzero entries correspond to the edges in the graphical model. For many applications, especially those involving high-dimensional data, it is desirable to prevent overfitting by utilizing priors that favor parsimonious models. Such models should exhibit as few interactions between variables as possible, which in the GGM case corresponds to a sparse precision matrix. A recent approach to simultaneously learn the structure and estimate the parameters in GGMs has been to augment the Gaussian likelihood with a penalty on the L^1 norm of the precision matrix (Banerjee et al., 2008; Duchi et al., 2008; Friedman et al., 2008; Meinshausen & Bühlman, 2006; Rothman et al., 2008; Yuan & Lin, 2007). This can also be thought of as a *maximum a posteriori* (MAP) estimation problem with a Laplace prior on the entries of the precision matrix.

One challenge with these methods is that they require selection of a parameter that controls the amount of sparsity by weighting the penalty term. Friedman et al. (2008) select this parameter through cross-validation by either minimizing prediction error or maximizing likelihood of the left-out data. Foygel et al. (2010) propose an extended Bayesian information criterion (EBIC) for choosing the regularization parameter. Fan and Li (2001) discuss the general setting of coefficient selection and estimation in statistical models using penalized likelihood functions. They point out that one issue with all L^p penalties is that the resulting estimates of large coefficients will be shrunk, i.e., biased towards zero. They introduce a penalty function called the smoothly-clipped absolute deviation (SCAD), which mitigates this bias for large estimates but still requires parameter tuning of the weight on the penalty. Scaled lasso from Sun and Zhang (2012) estimates noise level and regression coefficients to find an appropriate penalty level, but this again faces issues with L^1 penalty's biasedness.

We propose a hierarchical Bayesian model for sparse precision matrix estimation in GGMs that has several attractive features. First, it does not require any parameter tuning—the sparsity of the model is adapted automatically to the data. Sparsity is achieved by

Proceedings of the 30^{th} International Conference on Machine Learning, Atlanta, Georgia, USA, 2013. JMLR: W&CP volume 28.

modeling the precision matrix with a hierarchical scale mixture of Gaussians prior with a parameter-free Jeffreys' hyperprior. Second, we show that this hierarchical prior does not suffer from the bias problem inherent to L^1 penalties. Finally, we naturally enforce the symmetry and positive-definiteness of the precision matrix by utilizing a Cholesky decomposition. Most importantly, we show empirically that our model estimates have improved performance over the widely-used GLASSO (Friedman et al., 2008) method with optimal sparsity parameter chosen by cross-validation. We test performance of the different estimators on simulated data, by measuring error to the correct solution, and on real data from a cell-signaling experiment, where we measure performance based on the ability of the trained model to explain a left-out test set.

2. Adaptively Sparse Precision Matrix Estimation

Let x denote the $n \times d$ matrix of observed data, which we think of as realizations of a multivariate Gaussian random variable $X = (X_1, \ldots, X_d) \sim N(0, \Omega^{-1})$. Our goal is to estimate the precision matrix Ω, while controlling the complexity of the model. First, we review the L^1-penalized likelihood approach for favoring sparse Ω estimates. This is given by the estimation problem

$$\hat{\Omega} = \arg\min_{\Omega} \ln|\Omega| - \operatorname{tr}(\Omega S) - \rho\|\Omega\|_1, \qquad (1)$$

where S is the sample covariance matrix of x, and ρ is a parameter that controls the amount of sparsity. Due to its computational efficiency, the most popular algorithm for solving this estimation problem is the graphical least absolute shrinkage and selection operator (GLASSO) method (Friedman et al., 2008). This is equivalent to a MAP estimate with a Laplacian prior on the entries of Ω. That is, $\Omega_{ij} \sim \text{Laplace}(0, \lambda)$, with $\lambda = 2\rho/n$, which gives the prior density

$$p(\Omega|\lambda) = \left(\frac{\lambda}{2}\right)^d \exp\left(-\lambda\|\Omega\|_1\right).$$

In this section we formulate a Bayesian hierarchical model that is designed to induce sparsity on Ω while removing the need for parameter tuning. We begin with a discussion of a natural parameterization of the Gaussian likelihood using the Cholesky decomposition. Next, we formulate a hierarchical prior on the entries of Ω, which extends the adaptive sparsity prior developed in (Figueiredo, 2003) for sparse regression problems. Finally, we develop a MAP estimation procedure using the Expectation Maximization (EM) algorithm, which we show has closed-form coordinate ascents for the maximization step.

2.1. Gaussian Likelihood Parameterized by the Cholesky Decomposition

Denote the Cholesky decomposition of Ω as $\Omega = LL^T$, where L is lower triangular. Using this decomposition, the multivariate Gaussian model can be naturally formulated in the form of a regression problem (Rue & Held, 2005). Define the coefficients $\beta_{ij} = -L_{ij}/L_{jj}$ and the precision of X_j as $\omega_j = 1/\sigma_j^2 = L_{jj}^2$. Then the lower triangular entries of Ω are given by

$$\Omega_{ij} = \sum_{k=1}^{j} L_{ik}L_{jk} = \sum_{k=1}^{j} \beta_{ik}\beta_{jk}\omega_k, \quad \text{for } i \geq j. \quad (2)$$

Now the multivariate Gaussian model $N(0, \Omega^{-1})$ is equivalent to the set of regression problems,

$$X_j = \mu_j + \sum_{i>j} \beta_{ij}(X_i - \mu_i) + \epsilon_j,$$
$$\epsilon_j \sim N(0, \omega_j^{-1}), \qquad j = 1, \ldots, d. \qquad (3)$$

We use these β_{ij} and ω_j coefficients to parameterize the Gaussian precision matrix in our proposed model. This formulation naturally enforces the symmetry and positive-definiteness of the precision matrix. The matrix $\Omega = LL^T$ is clearly symmetric for any choice of coefficients. Also, the ω_j coefficients are the eigenvalues of Ω, so Ω being positive-definite is equivalent to the ω_j being strictly positive. This is satisfied for any reasonable prior that does not shrink the ω_j to zero, which we demonstrate holds for our model below.

A more practical form of the multivariate Gaussian likelihood utilizes the sufficient statistic, the sample covariance matrix S. Computations involving the sample covariance matrix are more efficient than those using the full data matrix (which is typically larger). The multivariate Gaussian log-likelihood is given by

$$\ell(\Omega|x) \propto n\ln|\Omega| - n\operatorname{tr}(\Omega S).$$

Note that the Gaussian likelihood written this way would technically involve the *biased* sample covariance matrix, i.e., with a factor of $(1/n)$. We rewrite this form of the Gaussian log-likelihood in terms of the β and ω coefficients as

$$\ell(\beta, \omega|x) \propto n\sum_{j=1}^{d} \ln\omega_j - n\sum_{i=1}^{d}\sum_{j=1}^{d} \Omega_{ij}S_{ij} \qquad (4)$$

$$= n\sum_{j=1}^{d} \ln\omega_j - n\sum_{i=1}^{d}\left(2\sum_{j=1}^{i-1}\sum_{k=1}^{j} S_{ij}\beta_{ik}\beta_{jk}\omega_k + \sum_{k=1}^{i} S_{ii}\beta_{ik}^2\omega_k\right) \qquad (5)$$

The two terms inside the parentheses of (5) correspond to the off-diagonal terms in Ω, which arise in pairs, and the diagonal terms, which appear just once.

2.2. Adaptive Sparsity Prior

As first described by Andrews and Mallows (1974), the Laplace distribution is equivalent to the marginal distribution of a scale mixture of Gaussians, with exponential scale distribution. More specifically, let θ be a random variable with the hierarchical distribution

$$\begin{aligned} p(\theta|\tau) &\sim N(0,\tau), \\ p(\tau|\gamma) &\sim \text{Exp}\left(\frac{\gamma}{2}\right). \end{aligned}$$

Then the marginal distribution of θ with respect to γ integrates to

$$\begin{aligned} p(\theta|\gamma) &= \int_0^\infty p(\theta|\tau)p(\tau|\gamma)d\tau \\ &= \frac{\sqrt{\gamma}}{2}\exp\left(-\sqrt{\gamma}|\theta|\right), \end{aligned}$$

giving us the Laplace distribution with $\rho = \sqrt{\gamma}/2$.

The level of sparsity in LASSO regression depends upon the parameter on the L^1 penalty. In this hierarchical-Bayes context, sparsity is controlled by γ. Figueiredo (2003) has proposed to remove γ by replacing the exponential hyperprior on τ by a Jeffreys' noninformative hyperprior, i.e.,

$$p(\tau) \propto \frac{1}{\tau}$$

This is equivalent to a log penalty on θ, which yields sparse solutions and is nearly unbiased for large coefficients. Although this hyperprior is improper, it has the advantages that it is scale-invariant and parameter-free. This modified hierarchical model is no longer equivalent to the Laplace prior, but has been shown in (Figueiredo, 2003) to be effective in regression and classification problems. The same hierarchical prior has also been used by Park and Casella (2008) in a Bayesian formulation of the LASSO model for regression. Adopting this prior for the Gaussian precision matrix estimation problem, we use the following hierarchical model for an adaptive sparsity estimation of Ω:

$$\begin{aligned} X &\sim N(0,\Omega), \\ \Omega_{ij} &\sim N(0,\tau_{ij}), \qquad \text{for } i>j, \\ p(\tau_{ij}) &\propto \frac{1}{\tau_{ij}}. \end{aligned} \tag{6}$$

Notice that we only put a prior on the off-diagonal entries of Ω. In other words, we do not seek to shrink the diagonal elements, although the model could easily be changed to include this. In what follows, we will treat the τ_{ij} as latent variables that are integrated out. This leaves us with a posterior $p(\Omega|x)$ that does not have any parameters that need to be tuned, which is a main advantage over the Laplace prior.

Also, the following simple example demonstrates that the hierarchical model in (6) does not suffer from the bias problem for large coefficients that plagues the Laplace prior. We note that this approximate unbiasedness of the MAP estimate comes at the cost of a non-convex posterior function. As Fan and Li (2001) show, convex penalty functions cannot simultaneously satisfy the properties of sparsity, continuity, and approximate unbiasedness. Consider a 2×2 sample covariance matrix,

$$S = \begin{pmatrix} 1 & s \\ s & 1 \end{pmatrix}^{-1},$$

where s is a scalar parameter satisfying $|s| < 1$ to ensure positive-definiteness. Then the maximum-likelihood estimate (MLE) of Ω is just given by the inverse of S. Now consider the L^1-penalized Gaussian likelihood model in (1), but with the L^1 penalty only on the off-diagonal element, $\Omega_{12} = \Omega_{21}$. Because the diagonal entries of Ω will remain equal, we can parameterize Ω with two unknown coefficients $a = \Omega_{11} = \Omega_{22}$ and $b = \Omega_{12} = \Omega_{21}$. Then the estimate of Ω under the L^1-penalized likelihood is

$$\begin{aligned} \hat{\Omega} &= \begin{pmatrix} \hat{a} & \hat{b} \\ \hat{b} & \hat{a} \end{pmatrix} \\ &= \arg\max_{(a,b)}\ n\ln(a^2-b^2) \\ &\qquad -2n(aS_{11}+bS_{12}) - |b|. \end{aligned} \tag{7}$$

The solution to this problem for the two unique entries a, b of Ω is shown in Figure 1 for varying values of the parameter s in the sample covariance and for a sample size of $n = 10$. This is compared to the MLE solution, which has solution $\hat{a} = 1$ and $\hat{b} = s$, and to the MAP estimate of the posterior under the proposed model (6). The proposed model estimate was computed using the EM algorithm described in the next subsection. There are two important features to notice in these plots. First, the L^1 solution for the off-diagonal b entry gives the familiar "soft-threshold" rule, which forces the estimate to be zero below some threshold, but is biased towards zero by an additive constant for larger coefficient values. Second, the diagonal a entry is nearly perfect (error within 10^{-4}) for the proposed MAP estimate, while the a estimate for the L^1 penalty estimate is biased downwards away from $s = 0$. This is the case even though there is no

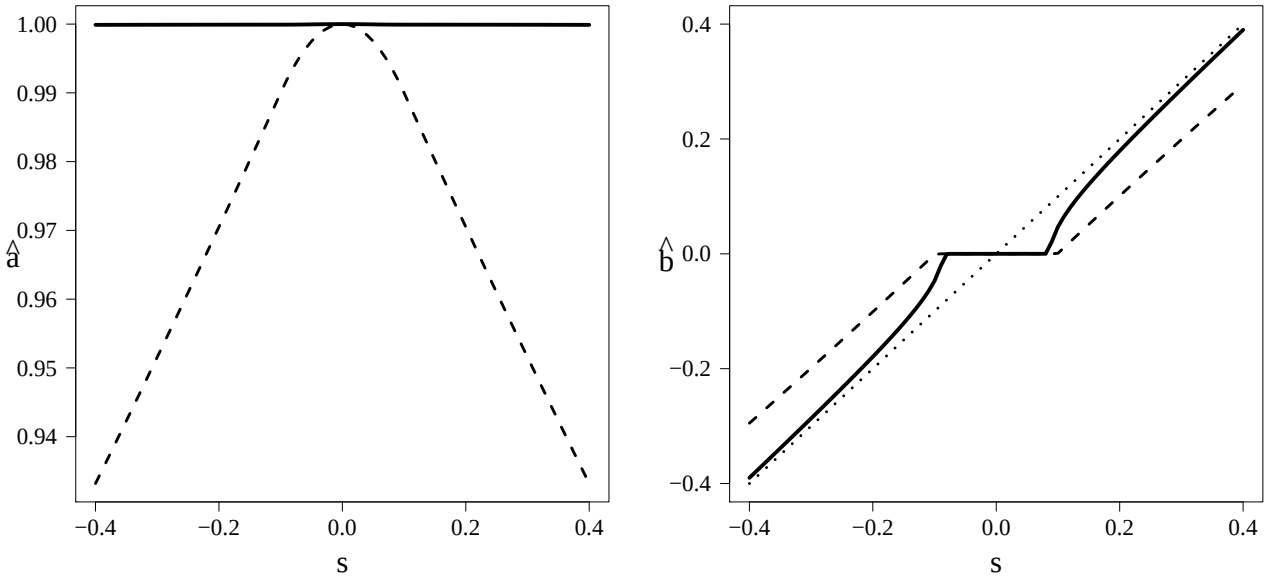

Figure 1. Comparison of estimates corresponding to Gaussian likelihood (dotted line), Laplace prior (dashed line), and proposed hierarchical prior (solid line) for the 2×2 matrix example. Estimates of both the diagonal entry (left plot) and off-diagonal entry (right plot) are shown.

L^1 penalty directly on the a entry, and it arises from the interaction between a and b in the log-determinant term in (7).

2.3. Expectation Maximization Algorithm

We develop an EM algorithm to estimate a sparse Ω as the maximizer of the posterior defined in (6). Although the prior on the precision matrix in (6) is written in terms of the entries Ω_{ij}, we will find it convenient to rewrite this in terms of the β_{ij} and ω_j coefficients discussed in Section 2.1. We can pass back and forth between the two parameterization of Ω using the relationship (2).

E step: The EM algorithm iteratively maximizes the Q function, which is a lower bound on the log-posterior, $\ln p(\beta, \omega|x)$, that we wish to maximize. The Q function is defined as the expectation over the latent variables τ_{ij} of the complete log-posterior, given the current estimate of the parameters at iteration t, which we denote $\hat{\beta}_{(t)}, \hat{\omega}_{(t)}$. Thus, we have

$$\begin{aligned} &Q(\beta, \omega|\hat{\beta}_{(t)}, \hat{\omega}_{(t)}) \\ &= \sum_{i=1}^{d} \sum_{j=1}^{i-1} \int \ln p(\beta, \omega|x, \tau_{ij}) p(\tau_{ij}|x, \hat{\beta}_{(t)}, \hat{\omega}_{(t)}) d\tau_{ij}. \end{aligned}$$

We can split the complete log-posterior inside the integral above into the sum of the log-likelihood, which does not depend on τ, and the log-prior, which does. Then the Q function breaks into two terms,

$$Q(\beta, \omega|\hat{\beta}_{(t)}, \hat{\omega}_{(t)}) = \ell(\beta, \omega|x) + E(\beta, \omega), \qquad (8)$$

where $\ell(\beta, \omega|x)$ is the log-likelihood given by (4), and $E(\beta, \omega)$ corresponds to the integral of the log-prior term. It is given by

$$\begin{aligned} E(\beta, \omega) &= \sum_{i=1}^{d} \sum_{j=1}^{i-1} \int \ln p(\beta, \omega|\tau_{ij}) p(\tau_{ij}|x, \hat{\beta}_{(t)}, \hat{\omega}_{(t)}) d\tau_{ij} \\ &= \sum_{i=1}^{d} \sum_{j=1}^{i-1} \int \ln p(\Omega_{ij}|\tau_{ij}) p(\tau_{ij}|x, \hat{\Omega}_{ij(t)}) d\tau_{ij} \\ &= \sum_{i=1}^{d} \sum_{j=1}^{i-1} \frac{\int \frac{1}{2} \tau_{ij}^{-2} \Omega_{ij}^2 N(\hat{\Omega}_{ij(t)}|0, \tau_{ij}^{-1}) d\tau_{ij}}{\int \tau_{ij}^{-1} N(\hat{\Omega}_{ij(t)}|0, \tau_{ij}^{-1}) d\tau_{ij}} \\ &= \sum_{i=1}^{d} \sum_{j=1}^{i-1} \hat{\Omega}_{ij(t)}^{-2} \Omega_{ij}^2 \\ &= \sum_{i=1}^{d} \sum_{j=1}^{i-1} \hat{\Omega}_{ij(t)}^{-2} \left(\sum_{k=1}^{j} \beta_{ik} \beta_{jk} \omega_k \right)^2. \end{aligned}$$

Here we have used (2) in the last line to write the final expression in terms of the β and ω coefficients. The evaluation of the integral above follows the same derivation as in (Figueiredo, 2003).

M step: The maximization step cannot be done in closed form for the entire set of β and ω coefficients at once. However, we derive a closed-form solution for the maximization of Q along a single coordinate at a time, i.e., a single β_{ij} or ω_j, keeping the other coordinates fixed. We begin by taking the derivate of the Q function (8) with respect to the β_{ab}. First, the derivative of the log-likelihood term is given by

$$\frac{d}{d\beta_{ab}}\ell(\beta,\omega|x) = -2n\left(\sum_{j=1}^{a-1} S_{aj}\beta_{jb}\omega_b + S_{aa}\beta_{ab}\omega_b\right) \quad (9)$$

Next, the derivative of the E term in the Q function is given by

$$\frac{d}{d\beta_{ab}}E(\beta,\omega) = 2\sum_{j=b}^{a-1}\frac{\beta_{jb}\omega_b}{|\hat{\Omega}_{aj(t)}|^2}\sum_{k=1}^{j}\beta_{ak}\beta_{jk}\omega_k + 2\sum_{i=a+1}^{d}\frac{\beta_{ib}\omega_b^2}{|\hat{\Omega}_{ia(t)}|^2}\sum_{k=1}^{a}\beta_{ik}\beta_{ak}\omega_k \quad (10)$$

Both equations (9) and (10) are linear in β_{ab}. So, maximization of the Q function with respect to β_{ab} is equivalent to solving a linear equation $k_1\beta_{ab} + k_0 = 0$, that is, setting $\beta_{ab} = -k_0/k_1$, with

$$\begin{aligned} k_0 = &-2n\sum_{j=1}^{a-1} S_{aj}\beta_{jb}\omega_b \\ &+2\sum_{j=b}^{a-1}\frac{\beta_{jb}\omega_b}{|\hat{\Omega}_{aj(t)}|^2}\sum_{\substack{k=1\\k\neq b}}^{j}\beta_{ak}\beta_{jk}\omega_k \\ &+2\sum_{i=a+1}^{d}\frac{\beta_{ib}\omega_b}{|\hat{\Omega}_{ia(t)}|^2}\sum_{\substack{k=1\\k\neq b}}^{a}\beta_{ik}\beta_{ak}\omega_k, \end{aligned} \quad (11)$$

$$\begin{aligned} k_1 = &-2nS_{aa}\omega_b \\ &+2\omega_b^2\left(\sum_{j=b}^{a-1}\frac{\beta_{jb}^2}{|\hat{\Omega}_{aj(t)}|^2} + \sum_{i=a+1}^{d}\frac{\beta_{ib}^2}{|\hat{\Omega}_{ia(t)}|^2}\right). \end{aligned} \quad (12)$$

Next, the derivative of Q with respect to ω_m is

$$\frac{dQ}{d\omega_m} = \frac{n}{\omega_m} - n\sum_{i=1}^{d}\left(2\sum_{j=1}^{i-1}S_{ij}\beta_{im}\beta_{jm} - S_{ii}\beta_{im}^2\right) - \sum_{j=m}^{d-1}\sum_{i=j+1}^{d}\frac{2\beta_{im}\beta_{jm}}{|\hat{\Omega}_{ij(t)}|^2}\left(\sum_{k=1}^{j}\beta_{ik}\beta_{jk}\omega_k\right).$$

Setting to zero and multiplying through by ω_m, this becomes a quadratic in ω_m. So, the maximum of Q with respect to ω_m is a solution to the equation $c_2\omega_m^2 + c_1\omega_m + c_0 = 0$, with

$$c_0 = n, \quad (13)$$

$$c_1 = -n\sum_{i=1}^{d}\left(2\sum_{j=1}^{i-1}S_{ij}\beta_{im}\beta_{jm} - S_{ii}\beta_{im}^2\right) - \sum_{j=m}^{d-1}\sum_{i=j+1}^{d}\frac{2\beta_{im}\beta_{jm}}{|\hat{\Omega}_{ij(t)}|^2}\left(\sum_{\substack{k=1\\k\neq m}}^{j}\beta_{ik}\beta_{jk}\omega_k\right), \quad (14)$$

$$c_2 = -\sum_{j=m}^{d-1}\sum_{i=j+1}^{d}\frac{2\beta_{im}^2\beta_{jm}^2}{|\hat{\Omega}_{ij(t)}|^2}. \quad (15)$$

Notice that c_2 is always negative, and thus the discriminant of the quadratic formula, $c_1^2 - 4c_0c_2$, is always positive, and both solutions are real. Also, the fact that c_0 is always nonzero ensures that $\omega_m = 0$ is never a solution, and therefore the estimate $\hat{\Omega}$ is guaranteed to remain strictly positive-definite. An important implementation detail is that many of the entries of $\hat{\Omega}_{(t)}$ will be driven to zero, but we need to divide by them in the above calculations. A similar issue arises in quadratic approximations of penalized likelihoods. Following the approach in (Hunter & Li, 2005), we add a small epsilon to the denominator to ensure numerical stability. This gives the following fast gradient ascent algorithm, which iterates over each coordinate and updates them one at a time using the above equations (linear in β_{ab} and quadratic in ω_m).

Algorithm 1 Expectation Maximization for Sparse Gaussian Estimation

Input: Sample covariance matrix S

1. Initialize $\hat{\beta}, \hat{\omega}$ to MLE, given by solutions to the d regression problems in (3)
2. Use (2) to initialize $\hat{\Omega}_{(0)}$
3. Until convergence, i.e., until gradient is zero:
   a. Set each $\hat{\beta}_{ab} = -k_0/k_1$, using (11), (12).
   b. Set each $\hat{\omega}_m = (c_2 + \sqrt{c_1^2 - 4c_2c_0})/2c_2$, using (13), (14), (15).
   c. Update $\hat{\Omega}_{(t)}$, using (2)

3. Experiments

3.1. Simulated Data

We generate data from a range of precision matrices of varying size and sparsity levels and compare our results with that of ordinary least squares (OLS), i.e., the Gaussian MLE, and GLASSO. The sizes of the precision matrices are $d^2 = 10^2$, 20^2, and 40^2. We test on

Table 1. Results on simulated data.

FROBENIUS NORM OF ERRORS

Matrix Size	10x10			20x20			40x40		
% L nonzero	5%	10%	20%	5%	10%	20%	5%	10%	20%
OLS	1.008	1.502	2.569	1.387	2.031	2.788	1.335	3.063	4.367
GLASSO	0.689	2.176	2.501	1.370	1.978	2.702	1.308	2.901	4.084
Proposed	**0.420**	**0.942**	**1.614**	**0.529**	**1.010**	**1.628**	**0.457**	**1.538**	**3.395**

NONZERO ERROR

Matrix Size	10x10			20x20			40x40		
% L nonzero	5%	10%	20%	5%	10%	20%	5%	10%	20%
OLS	0.142	0.211	0.341	0.109	0.173	0.178	0.054	0.097	0.129
GLASSO	0.163	0.418	0.330	0.106	0.168	0.172	0.052	0.091	0.121
Proposed	**0.112**	**0.184**	**0.242**	**0.084**	**0.116**	**0.121**	**0.040**	**0.071**	**0.117**

ZERO ERROR

Matrix Size	10x10			20x20			40x40		
% L nonzero	5%	10%	20%	5%	10%	20%	5%	10%	20%
OLS	0.092	0.117	0.158	0.063	0.076	0.096	0.031	0.067	0.082
GLASSO	0.031	0.044	0.156	0.063	0.075	0.094	0.030	0.063	0.076
Proposed	**0.010**	**0.010**	**0.010**	**0.002**	**0.012**	**0.015**	**0.000**	**0.005**	**0.014**

% NONZERO PREDICTION ACCURACY

Matrix Size	10x10			20x20			40x40		
% L nonzero	5%	10%	20%	5%	10%	20%	5%	10%	20%
OLS	**100.0**	**100.0**	**100.0**	**100.0**	**100.0**	**100.0**	**100.0**	**100.0**	**99.99**
GLASSO	**100.0**	99.62	**100.0**	**100.0**	**100.0**	99.94	**100.0**	99.78	99.98
Proposed	**100.0**	79.23	83.18	89.74	91.94	78.20	93.69	85.46	87.41

% ZERO PREDICTION ACCURACY

Matrix Size	10x10			20x20			40x40		
% L nonzero	5%	10%	20%	5%	10%	20%	5%	10%	20%
OLS	0.12	0.14	0.00	0.25	0.18	0.14	0.35	0.12	0.17
GLASSO	57.56	44.86	0.27	0.46	0.43	0.43	0.85	1.18	1.33
Proposed	**100.0**	**99.59**	**98.75**	**99.59**	**96.77**	**95.27**	**99.99**	**96.82**	**94.29**

precision matrices composed from lower triangle matrices with levels of 5, 10, and 20% non-zero entries in the off-diagonals. The diagonal entries are Gaussian distributed with $\mu_{\text{diag}} = 1$ and $\sigma_{\text{diag}} = 0.1$, and the non-zero off-diagonal entries are Gaussian distributed with $\mu_{\setminus\text{diag}} = 0$ and $\sigma_{\setminus\text{diag}} = 1$. This ensures that the precision matrices we work with are positive definite. For each of the nine scenarios, we test our method on twenty different sets of data, each having $1.5d^2$ sample points. We run a first pass of our algorithm initializing from the OLS solution with an epsilon value of 1e-3 and a stopping criterion of when the maximum change in the estimated entries is less than 1e-7. We then repeat the algorithm initializing from the output of the previous step with epsilon values of 1e-5 and 1e-7. This three-step refinement to lower epsilons while maintaining numerical stability serves to avoid local minima and confirms that the choice of epsilon does not affect the solution. The MAP estimate should correctly set sparse entries exactly to zero, such as is shown in Figure 1 (right plot). However, because we solve the MAP with a gradient descent, some zero entries may not be numerically exactly zero. To differentiate between truly nonzero entries and zero entries that numerically off, we maximize the posterior with respect to a numerical zero threshold value (i.e., a value for which all entries numerically smaller get set to zero). We stress that this is simply part of the optimization process, and that the MAP estimates do truly produce sparse solutions. In all experiments initializing from the OLS solution, the EM algorithm converges to a local maximum that is better than the GLASSO and OLS solutions. This improvement in the solutions says that we converge to a reasonable answer, even if the EM algorithm does not guarantee we find the global maximum. All computations are done in R.

The GLASSO models used for comparison are selected through tenfold cross-validation for the regularization parameter ρ that maximizes likelihood. Friedman et al. (2008) use the list of ten ρ values suggested by glassopath() function in the GLASSO package for their cross-validation. Because we have found that the optimal ρ sometimes lies below this range, we expand the range of glassopath() values with ten more equally

spaced ρ values starting from 1e-4 to the minimum glassopath() value to make sure that the optimal ρ is within range.

The first metric we use to rate the performance of our proposed method is the Frobenius norm of the difference of the estimations from the true Ω, i.e., the root mean squared error, $\|\Omega - \hat{\Omega}\|_F$. We also split this error to see how much of it results from what should be zero versus nonzero terms in the true Ω. With the threshold that maximizes the MAP estimate, we then calculate the percentage of zero entries in the true solution our method identifies correctly. The results have been averaged over the twenty different trials from each scenario and are summarized in Table 1.

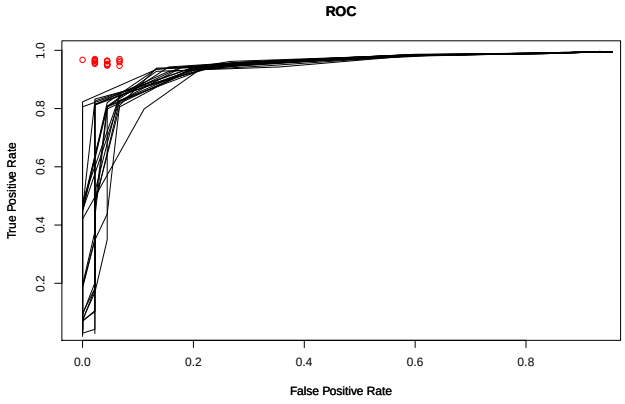

Figure 2. Comparison of specificity/sensitivity of our method (red dots) to the ROC for GLASSO (black curves) on the 40x40, 5% nonsparse, synthetic data (20 random data sets). Correctly detected sparse entries are considered true positives.

From the results, we can see that our method has the best performance in terms of errors and zero prediction in all scenarios, whereas GLASSO tends to fail especially in denser and larger matrices. The bias from the Laplace prior in GLASSO shrinks entries that should be nonzero as described in Section 2.2, whereas our method mitigates this bias for large entries. This can be seen in the comparison of the nonzero errors, where in some cases of the experiment, GLASSO has a higher nonzero error than the OLS estimate. For the error and prediction accuracy of zero entries, our method performs the best by far. It is known that cross-validation parameter selection for ρ in GLASSO is overly-conservative, i.e., it does not strongly induce sparsity, which can be seen in the poor performance for GLASSO on zero prediction and zero entry error. Another reason for GLASSO's poor performance is that for even only moderately dense matrices, cross-validation yields ρ values very close to 0, and thus gives results near that of the OLS, as can be observed in the similar zero error rates between OLS and GLASSO for the denser matrices. Because of our hierarchical prior that adaptively induces sparsity, our method is able to consistently predict with mid-90% accuracy the zero entries in Ω across the various sparsity scenarios. In the nonzero accuracy numbers, GLASSO is consistently close to 100%, but this is because its overly-conservative sparsity finds far too many entries to be nonzero. We note that a large majority of the entries in the matrices are zero, so the zero accuracy contributes more to the overall accuracy. Figure 2 shows ROC curves for GLASSO generated by varying the sparsity parameter ρ. Our method achieves higher specificity/sensitivity than GLASSO, over all parameter values.

3.2. Cell Signaling Data

To show how our method works on real-world data, we run our algorithm on the protein signaling dataset from (Sachs et al., 2005). This is the same dataset analyzed in (Friedman et al., 2008). The $d = 11$ protein dataset has $n = 7466$ samples. In both GLASSO and our proposed method, the likelihood term dominates the prior because of the large sample size. We find that the optimal ρ for running GLASSO on all of the data is 0, which is what Friedman et al. (2008) have reported. Therefore, OLS, GLASSO, and our proposed method produce almost identical results.

The estimated graph presented in (Sachs et al., 2005) is conventionally accepted by biology experts. A more interesting experiment is to train on smaller data samples and test whether GLASSO and our proposed method would yield estimates similar to what biologists expect, keeping in mind that Sachs' directed graph with only binary values may not be an exact ground truth for direct comparison. Sachs' graph contains 43 edges and 78 pairs of nodes without edges.

We train both GLASSO and our method on samples of 200 points. Averaging over 100 runs, we calculate the percentage of zero and nonzero entries in agreement between the Sachs' model and both GLASSO and our proposed model. For our method, we repeat the same three-step epsilon refinement procedure that we do for synthetic data experiments with epsilon down to 1e-9 because the entries in the precision matrix of the cell signaling data are much smaller by a factor of 1/1000. For GLASSO, we do leave-one-out cross-validation (LOOCV) on the subsample over the range of ρ from glassopath augmented with ten evenly spaced values from 1e-4 to the minimum glassopath value to find the optimal ρ. Also, as we do with the

Table 2. Results on zero and nonzero entry prediction of cell signaling data.

	% Zero Prediction Accuracy	% Nonzero Prediction Accuracy	% Overall Prediction Accuracy
GLASSO	34.576%	77.418%	49.800%
Proposed	71.512%	68.744%	**70.527%**

Table 3. Likelihood test from cell signaling data.

	Likelihood	Frobenius Norm of Error
OLS	-4.47319e5	5.60782e-3
GLASSO	-4.48852e5	5.12098e-3
Proposed	**-4.45147e5**	**5.06741e-3**

simulated data experiment, to account for numerical error we apply a threshold of delta according to the maximum posterior. We also try this for GLASSO, but GLASSO's optimization attains its maximum posterior with no thresholding necessary. We note that the results from GLASSO are not symmetric positive definite. The average relative difference between corresponding off-diagonal entries is 19%, calculated by $(\hat{\Omega}_{ij} - \hat{\Omega}_{ij}^T)/(\hat{\Omega}_{ij} + \hat{\Omega}_{ij}^T)$.

For calculating the prediction accuracies relative to Sachs' model, we convert the MAP estimates for GLASSO and our method into binary matrices where all nonzero entries are coded into ones. The prediction accuracies of the zero and nonzero entries for both methods are displayed in Table 2, along with the combined prediction accuracies for all entries of the precision matrix. The results are consistent with the accuracy rates from synthetic experiments. Our proposed method performs much better than GLASSO with the zero prediction but slightly worse with the nonzeros. We observe that for GLASSO the cross-validation can at times choose a model with a ρ very close to zero, resulting in a precision matrix that is not sparse. Overall, our proposed method gives estimations with higher similarity to the conventionally-accepted graph from (Sachs et al., 2005).

Because of the large sample size in the full set of data, the OLS precision matrix based on the entire dataset is a good approximation to the true precision matrix. The fit of the models to the left-out test data is measured by the Gaussian log-likelihood, $\ell(\hat{\Omega}|x_{\text{test}})$. Another experiment we do is to calculate as a measure of the error rate the Frobenius norms of the differences between the OLS precision matrix based on the entire dataset and the estimated models trained from subsampling via GLASSO and our method. Again, we use the three-step epsilons up to 1e-9 for our method and do LOOCV for GLASSO over the same range of ρ as the previous cell data experiment. Results are averaged over twenty trials. Table 3 shows these results. Compared to OLS and GLASSO, our method has a higher likelihood and a lower norm difference from the OLS precision matrix from all of the dataset.

4. Conclusions

In this paper, we have introduced a hierarchical Bayesian model for the estimation of a sparse precision matrix in a Gaussian graphical model. The primary advantage of our approach, in comparison to the commonly used L^1 penalty methods, is the ability of our model to adapt to different levels of sparsity without the need for parameter tuning. In fact, this adaptive sparsity proved to give much improved structure learning (zero identification) over GLASSO with optimal sparsity parameter chosen by cross-validation in experiments on simulated data. In addition, we showed that the estimated coefficients from the proposed model do not suffer from the same bias problems that L^1 penalties display.

There are several avenues for future work. First, it would be nice to understand what theoretical guarantees can be derived from the proposed hierarchical model, e.g., asymptotic consistency. To the best of our knowledge, this has not been worked out for such priors even in the regression case. Such guarantees do exist for L^1-penalized likelihoods. Finally, we expect that the true benefits of our proposed model will be evident in applications where the dimension is much greater than the sample size, for example, in genetics analysis and in identifying functional brain networks.

Acknowledgments

This work was supported by NIH/NIMH Grant Number R01MH084795.

References

Andrews, D. F. and Mallows, C. L. Scale mixtures of normal distributions. *Journal of the Royal Statistical Society*, 36(1):99–102, 1974.

Banerjee, Onureena, Ghaoui, Laurent El, and d'Aspremont, Alexandre. Model selection through sparse maximum likelihood estimation for multivariate Gaussian or binary data. *Journal of Machine Learning Research*, 9:485–516, 2008.

Duchi, John, Gould, Stephen, and Koller, Daphne. Projected subgradient methods for learning sparse Gaussians. In *Proceedings of the Conference on Uncertainty in Artificial Intelligence*, 2008.

Fan, Jianqing and Li, Runze. Variable selection via nonconcave penalized likelihood and its oracle properties. *Journal of the American Statistical Association*, 96(456):1348–1360, 2001.

Figueiredo, Mário A. T. Adaptive sparseness for supervised learning. *IEEE Transactions on Pattern Analysis and Machine Learning*, 25(9):1150–1159, 2003.

Foygel, Rina and Drton, Mathias. Extended Bayesian information criteria for Gaussian graphical models. In *Proceedings of Neural Information Processing Systems*, 2010.

Friedman, Jerome, Hastie, Trevor, and Tibshirani, Robert. Sparse inverse covariance estimation with the graphical lasso. *Biostatistics*, 9(3):432–441, 2008.

Hunter, David R. and Li, Runze. Variable selection using MM algorithms. *The Annals of Statistics*, 33 (4), 2005.

Meinshausen, Nicolai and Bühlman, Peter. High-dimensional graphs and variable selection with the lasso. *The Annals of Statistics*, 34(3):1436–1462, 2006.

Park, Trevor and Casella, George. The Bayesian lasso. *Journal of the American Statistical Society*, 103(482):681–686, 2008.

Rothman, Adam J., Bickel, Peter J., Levina, Elizaveta, and Zhu, Ji. Sparse permutation invariant covariance estimation. *Electronic Journal of Statistics*, 2:494–515, 2008.

Rue, Håvard and Held, Leonhard. *Gaussian Markov Random Fields: Theory and Applications*. Chapman and Hall/CRC, 2005.

Sachs, Karen, Perez, Omar, Pe'er, Dana, Lauffenburger, Douglas A., and Nolan, Gary P. Causal protein-signaling networks derived from multiparameter single-cell data. *Science*, 308:523–529, 2005.

Sun, Tingni and Zhang, Cun-Hui. Sparse matrix inversion with scaled lasso. arXiv:1202.2723 [math.ST].

Yuan, Ming and Lin, Yi. Model selection and estimation in the Gaussian graphical model. *Biometrika*, 94(1):19–35, 2007.