Published in final edited form as:
Stat Interface. 2013 April 1; 6(2): 243–259.

The Bayesian Covariance Lasso

Zakaria S Khondker,
Department of Biostatistics University of North Carolina Chapel Hill, North Carolina 27599, url:

Hongtu Zhu,
Department of Biostatistics University of North Carolina Chapel Hill, North Carolina 27599

Haitao Chu,
Division of Biostatistics University of Minnesota Minneapolis, Minnesota, U.S.A

Weili Lin, and
Biomedical Research Imaging Center University of North Carolina Chapel Hill, North Carolina 27599

Joseph G. Ibrahim
Department of Biostatistics University of North Carolina Chapel Hill, North Carolina 27599

Abstract

Estimation of sparse covariance matrices and their inverse subject to positive definiteness constraints has drawn a lot of attention in recent years. The abundance of high-dimensional data, where the sample size (*n*) is less than the dimension (*d*), requires shrinkage estimation methods since the maximum likelihood estimator is not positive definite in this case. Furthermore, when *n* is larger than *d* but not sufficiently larger, shrinkage estimation is more stable than maximum likelihood as it reduces the condition number of the precision matrix. Frequentist methods have utilized penalized likelihood methods, whereas Bayesian approaches rely on matrix decompositions or Wishart priors for shrinkage. In this paper we propose a new method, called the Bayesian Covariance Lasso (BCLASSO), for the shrinkage estimation of a precision (covariance) matrix. We consider a class of priors for the precision matrix that leads to the popular frequentist penalties as special cases, develop a Bayes estimator for the precision matrix, and propose an efficient sampling scheme that does not precalculate boundaries for positive definiteness. The proposed method is permutation invariant and performs shrinkage and estimation simultaneously for non-full rank data. Simulations show that the proposed BCLASSO performs similarly as frequentist methods for non-full rank data.

Keywords

Bayesian covariance lasso; non-full rank data; Network exploration; Penalized likelihood; Precision matrix

1. INTRODUCTION

Shrinkage of a high-dimensional covariance matrix or its inverse, known as the precision or concentration matrix, particularly when the dimension of the matrix (*d*) is larger than the sample size (*n*), has drawn a lot of attention in recent years. The abundance of high-dimensional data in structural and functional magnetic resonance imaging where a few dozen subjects are scanned with each scan having thousands of voxels or hundreds of regions of interest [14], spectroscopy, climate studies and many other applications are just a few examples. Another motivation is due to the major use of precision matrices in statistical tools like principal component analysis, linear and quadratic discriminant analysis, inference

on the mean parameters, analysis of independence and conditional independence in graphical models, and so on. They all require estimation of the covariance or precision matrix. The precision matrix has a partial correlation interpretation- off-diagonal elements represent the conditional covariances between the corresponding variables. The results can be summarized in a graph by linking conditionally dependent variables, thereby providing an understanding of how variables, such as the genes or regions of the brain, are related to each other. Hence it is suitable for network exploration.

In-depth theoretical studies of the sample (empirical) covariance matrix S have shown that without regularization, the sample covariance matrix performs poorly in high dimensional settings, hence stimulating research on alternative estimators. When the dimension of the matrix is large, the largest eigenvalue can be very large compared to the smallest eigenvalue, resulting in a large condition number and unstable estimators for the precision matrix S^{-1}. In practice, when n is relatively small compared to the dimension d, the S matrix approaches singularity, therefore leading to unreliable estimates for the precision matrix S^{-1}. In many cases, such a situation may lead to near-zero eigenvalues for S. The problem is even more serious for non-full rank data (when $n < d$). In this case, S has a maximum rank of n which is smaller than its dimension d, and therefore S is singular.

In the frequentist framework, significant work has been done on model selection and precision (covariance) matrix estimation in Gaussian models [1, 12, 11, 8, 31]. The original paper by Dempster [6] introduced the idea of shrinkage estimation which forces some elements of the precision matrix to be zero. In its infancy, the methods for shrinkage estimation involved two steps: (i) identify the “correct” model by determining which elements are zero; (ii) estimate the parameters for the non-zero elements. Edwards [9] has discussed some standard approaches for identifying the model, such as greedy stepwise forward-selection and backward-elimination procedures, achieved through hypothesis testing. Drton and Perlman [8] proposed a conservative simultaneous confidence interval to select a model in a single step as an improvement. [1] and [12] proposed the graphical (covariance) lassso (CLASSO) penalty to force some elements of the precision matrix to zero and simultaneously estimate the rest of the elements. [11] extended the CLASSO approach to the adaptive covariance lasso (ACLASSO) and the smoothly clipped absolute deviation penalty for covariance estimation (CSCAD). [31] and [22] used a penalty on the off-diagonal elements only and called their estimator the sparse permutation invariant covariance estimator (SPICE).

Among Bayesian shrinkage methods, [29] used reference priors on the eigenvalues of the covariance matrix to regularize the eigen structure. Smith and Kohn [27] decomposed the precision matrix Φ as $\Phi = BDB^{\mathrm{T}}$ where B is a lower triangular matrix with 1's on the diagonal and D is a diagonal matrix. They introduced priors on the elements of B and D. [13] extended this idea and used a Cholesky decomposition on the covariance matrix $V = CC^{\mathrm{T}}$, where C is a lower triangular matrix. They also discussed how the ordering of the data can change the zero patterns. These methods use priors on the Cholesky factors and therefore are not permutation invariant; they may not be rational choices when there is no natural ordering of the parameters in the matrix. [2] decomposed the covariance matrix as $\Sigma = \mathrm{diag}(s)R\,\mathrm{diag}(s)$, where s is the $d \times 1$ vector of standard deviations and R is the $d \times d$ matrix of correlation coefficients. They used priors on individual elements of these matrices. [28] extended this idea to the precision matrix Φ and decomposed $\Phi = TCT^{\mathrm{T}}$, where $T = \mathrm{diag}(T_1, \ldots, T_d)$ is a diagonal matrix such that T_k are the inverses of the partial standard deviations and C is a correlation matrix with $C_{kk} = 1$ and $C_{kk'} = -\rho_{kk'}$ for all $k \neq k'$. The graphical methods of [21, 3] rely on hyper-inverse Wishart based sampling. These two-step methods need to exploit decombosibility structures of the graphs, which may not be the case

for unstructured precision matrices. Moreover, they only use Wishart priors and are unable to exploit other types of penalties.

The single-step Bayesian methods primarily rely on priors on the elements arising from some sort of decomposition of the precision (covariance) matrix, which do not readily translate to any recognizable priors on the elements of the precision (covariance) matrix itself. Furthermore, most of these methods are based on sampling the elements of the matrix one at a time which is not efficient. Specifically, these methods pick a single element at a time, find an appropriate boundary that yields a positive definite matrix, and then draw a sample of this element. Drawing one element at a time is inefficient, and coupled with the additional computational complexities in computing boundaries for the elements, these methods are not suitable for high-dimensional matrices. The full posterior distribution of the elements of the precision and covariance matrices under lasso-type penalties have not been explored. The direct L_1 penalties on the elements of the covariance matrix have not been studied in the Bayesian framework. There appears to be a lack of a connection between the popular frequentist penalized approaches and their Bayesian competitors. Our goal is to build a bridge between the frequentist and Bayesian approaches to covariance estimation.

We propose generalized priors which include common frequentist penalties like the adaptive lasso penalty of [11], the lasso (L_1) penalty of [12], and the SPICE penalty of [22] as special cases. Then we introduce a new Bayesian approach for sampling from the posterior distribution of the precision matrix one whole column at a time and rely on multiple tries to achieve the desired acceptance rate. The proposed method is particularly attractive and efficient compared to the existing single-step methods as it updates the matrix one entire column at a time (on the order of d) instead of one element at a time (on the order of d^2). Our sampling scheme rejects any sample that is not a positive definite matrix and is permutation invariant. In addition, the method is based on specifying priors directly on the elements of the precision matrix instead of priors on the elements of a matrix decomposition, and the proposed method performs shrinkage and estimation simultaneously. We also explore the posterior distribution of the elements under the lasso penalty and provide a Bayesian minimax estimator as an alternative to the popular frequentist posterior mode estimators under L_1 penalties.

To illustrate the proposed methodology, we consider data from functional connectivity Magnetic Resonance Imaging (fcMRI) from 90 regions of interest (ROI) of 30 2-year old children. All images were acquired on a 3 Tesla Magnetic Resonance (MR) scanner with a gradient echo-planar imaging sequence. The imaging sequence was repeated 150 times. The images of the first 10-20 time points were typically excluded from the data analysis to ensure that magnetization reaches the steady state. All subjects are healthy normal controls and imaged at sleep without sedation. In this study, the signals were obtained from the remaining 130 time points. Our primary purpose here is to build a network between regions when there is no prior information about the underlying structure of the network or graph.

2. THE GENERAL METHOD

Let $Y_i \sim N_d(0, \Phi^{-1})$ for $i = 1, \ldots, n$ be n independent observations, where $\Phi = (\theta_{kk}) = \Sigma^{-1}$ is a $d \times d$ precision matrix. Then the joint distribution of $Y = (Y_1, \ldots, Y_n)$is given by

$$p(Y|\Theta) \propto (\det \quad \Theta)^{\frac{n}{2}} \quad \exp\left\{-\frac{1}{2}\sum_{i=1}^{n} Y_i^{\mathrm{T}}\Theta Y_i\right\} I(\Theta \succ 0),$$

where $I(\Theta \succ 0)$ is an indicator function of the event that Φ is positive definite. $S=\sum_{i=1}^{n} Y_i Y_i^{\mathrm{T}}/n$ is the maximum likelihood estimator of Σ.

2.1 Proposed Priors

We choose independent exponential priors for the diagonal elements; $\phi_{kk} \sim \mathrm{Exp}(\beta_k)$ and Laplace priors for the off-diagonal elements $\phi_{kk'} \sim \mathrm{Laplace}(0, b_{kk'})$ for $k > k'$ and $k, k' = 1, \ldots, d$. Then, the posterior distribution of Φ, $p(\Phi|Y)$, is given by

$$(\det \quad \Theta)^{\frac{n}{2}} \prod_{k=1}^{d} \exp\left\{-\frac{n}{2} tr\,(S\Theta) - \sum_{k=1}^{d} \beta_k \theta_{kk} - \sum_{k=2}^{d} \sum_{k'=1}^{k-1} b_{kk'} |\theta_{kk'}|\right\},$$

where det(·) denotes the determinant of a matrix. The log-posterior function equals

$$\log \quad p(\Theta|Y) = \frac{n}{2} \log \ \det \quad \Theta - \frac{n}{2} tr\,(S\Theta) - \sum_{k=1}^{d} \beta_k \theta_{kk} - \sum_{k=2}^{d} \sum_{k'=1}^{k-1} b_{kk'} |\theta_{kk'}| + C, \quad (1)$$

where C is a constant independent of Φ. The popular frequentist penalized likelihoods including ACLASSO, CLASSO and SPICE can be derived from (1) as special cases as follows. If we choose $\beta_k = nd\lambda_{kk}/2$ and $b_{kk'} = nd\lambda_{kk'}$ (for $k > k'$), then (1) reduces to

$$\frac{n}{2}\left\{\log \ \det \quad \Theta - tr\,(S\Theta) - \sum_{k=1}^{d} \sum_{k'=1}^{d} d\lambda_{kk'} |\theta_{kk'}|\right\} + C. \quad (2)$$

[11] optimized equation (2) as the objective function in the ACLASSO method, which can be interpreted as the posterior mode under $\mathrm{Exp}(nd\lambda_{kk}/2)$ priors for the diagonal elements and $\mathrm{Laplace}(nd\lambda_{kk'})$ priors for the off-diagonal elements of the precision matrix.

If we set $b_{kk'} = 2\beta_k = n\rho$, the priors for ϕ_{kk} are *i.i.d* $\mathrm{Exp}(n\rho/2)$ and the $\phi_{kk'}$ are *i.i.d* $\mathrm{Laplace}(n\rho)$ for $k > k'$. Then (1) reduces to

$$\log \quad p(\Theta|Y) = \frac{n}{2}\left\{\log \ \det \quad \Theta - tr\,(S\Theta) - \rho \|\Theta\|_{l_1}\right\} + C, \quad (3)$$

where $\|\Theta\|_{l_1} = \sum_{k=1}^{d} \sum_{k'=1}^{p} |\theta_{kk'}|$ is the l_1 norm of Φ. [1] optimized equation (3) in their covariance selection method (ignoring $n/2$), while [12] also optimized equation (3) in their CLASSO method, which is essentially the posterior mode under $\mathrm{Exp}(n\rho/2)$ priors for the diagonal elements and $\mathrm{Laplace}(n\rho)$ priors for the off-diagonal elements of Φ. [1] has shown that (3) is concave in Φ, which yields that the posterior distribution of Φ is uni-modal. Hence, we will use $\mathrm{Exp}(n\rho/2)$ priors for the diagonal elements and $\mathrm{Laplace}(n\rho)$ priors for the off-diagonal elements of Φ so that our log-posterior is the same as the objective function of CLASSO in (3).

If we choose not to penalize the diagonal elements of Φ, then we can let the hyperparameter β_k approach 0 ($\beta_k \to 0$) or equivalently choose improper uniform priors on $(0, \infty)$ for the diagonal elements of Φ. In that case, (3) further reduces to

$$\log \quad p(\Theta|Y) = \frac{n}{2}\left\{\log \ \det \quad \Theta - tr\,(S\Theta) - \rho \|\Theta^{-}\|_{l_1}\right\} + C, \quad (4)$$

where Φ^{-} has the same off-diagonal elements as Φ but all the diagonal elements are zero. [31] and [22] used equation (4) as their objective function (ignoring $n/2$ and C) and calculated the posterior mode in their SPICE method.

2.2 Full Conditionals

For $k = 1, \ldots, d$, we partition and rearrange the columns of Φ and S as follows:

$$\Theta = \begin{pmatrix} \Theta_{-kk} & \theta_k \\ \theta_k^{\mathrm{T}} & \theta_{kk} \end{pmatrix} \quad \text{and} \quad S = \begin{pmatrix} S_{-kk} & \boldsymbol{s}_k \\ \boldsymbol{s}_k^{\mathrm{T}} & s_{kk} \end{pmatrix}, \quad (5)$$

where ϕ_{kk} is the kth diagonal element of Φ, $\phi_k = (\phi_{k1}, \ldots, \phi_{k,k-1}, \phi_{k,k+1}, \ldots, \phi_{kd})^{\mathrm{T}}$ is the vector of all off-diagonal elements of the kth column, and Φ_{-kk} is the $(d-1) \times (d-1)$ matrix of all the remaining elements, i.e., the matrix resulting from deleting the kth row and kth column from Φ. By using the Schur decomposition [25], we have $\det(\Phi) = \det(\Phi_{-kk}) D_k$, where $D_k = (\phi_{kk} - C_k)$ and $C_k = \theta_k^T \Theta_{-kk}^{-1} \theta_k$ are scalar quantities. Similarly, s_{kk} is the kth diagonal element of S, s_k is the vector of all off-diagonal elements of the kth column of S, and S_{-kk} is the matrix of all remaining elements.

Our primary aim is to sample from the posterior distribution of the kth column of Φ for $k = 1, \ldots, d$. It follows from (3) that the conditional densities for θ_{kk} and $\boldsymbol{\theta}_k$ can be written as follows:

$$\begin{aligned} p(\theta_{kk} | y, \theta_k, \Theta_{-kk}, \rho) &\propto D_k^{\frac{n}{2}} \exp\left\{-\tfrac{n}{2}(s_{kk}+\rho)\theta_{kk}\right\}, \\ p(\theta_k | y, \theta_{kk}, \Theta_{-kk}, \rho) &\propto D_k^{\frac{n}{2}} \exp\left\{-\tfrac{n}{2}\left(s_k^{\mathrm{T}}\theta_k + \rho \|\theta_k\|_{l_1}\right)\right\} \times I(D_k > 0), \end{aligned} \quad (6)$$

where $I(A)$ is the indicator function of the event A. The derivation of (6) is given in Appendix I. Under the SPICE penalty, the full conditional distribution for $\boldsymbol{\theta}_k$ is the same while the full conditional distribution for θ_{kk} changes to

$$p(\theta_{kk} | y, \theta_k, \Theta_{-kk}, \rho) \propto D_k^{\frac{n}{2}} \exp\left\{-\tfrac{n}{2} s_{kk} \theta_{kk}\right\}.$$

Note that in (6), we could replace D_k by $\det(\Phi)$ which is computed faster than D_k since $\theta_k^{\mathrm{T}} \Theta_{-kk}^{-1} \theta_k$ requires inverting a $(d-1) \times (d-1)$ matrix and then computing a quadratic form of the same order. However, we will need to compute $\theta_k^{\mathrm{T}} \Theta_{-kk}^{-1} \theta_k$ to sample the diagonal elements θ_{kk} and we will not require any additional computations when sampling the off-diagonals $\boldsymbol{\theta}_k$. We are led to the following theorem whose proof is given in Appendix I.

Theorem 1: Suppose we start with a positive definite current value of Φ and sample from $p(\theta_{kk} | y, \theta_k, \Theta_{-kk}) \propto D_k^{\frac{n}{2}} \exp\left\{-\tfrac{n}{2}(s_{kk}+\rho)\theta_{kk}\right\}$ and

$$p(\theta_k | y, \theta_{kk}, \Theta_{-kk}) \propto D_k^{\frac{n}{2}} \exp\left\{-\frac{n}{2}(\boldsymbol{s}_k + \rho\gamma_k)^{\mathrm{T}} \theta_k\right\} I(D_k > 0),$$

where $\gamma_k = (\gamma_{k1}, \ldots, \gamma_{kd})^{\mathrm{T}}$ and $\gamma_{kj} = \mathrm{sign}(\theta_{kj})$ for $j = 1, \ldots, d$. This sampling process guarantees that we sample positive definite values of Φ at all subsequent steps.

Theorem 1 ensures that the Bayesian covariance lasso (BCLASSO) can achieve positive-definiteness for any nonnegative penalty parameter ρ.

2.3 Proposed Sampling Scheme

Gibbs sampling for the diagonal elements is straightforward since their full conditionals are available in closed form. The full conditionals for the off-diagonals are not available in closed form and therefore we will use the standard Metropolis-Hastings algorithm within Gibbs to sample the off-diagonal elements. In many applications, the off-diagonal elements are nearly symmetric suggesting a normal proposal density as a suitable choice. The mean of the proposal density is chosen to be the current value of Φ and the choice of the variance of the proposal density is determined from the Hessian matrix. We can write

$$\log \quad p\left(\theta_k|y,\theta_{kk},\Theta_{-kk}\right)=0.5n \quad \left\{\log \quad D_k-\left(\boldsymbol{s}_k+\rho\gamma_k\right)^{\mathrm{T}}\theta_k\right\}+C.$$

The first-order derivative of the logarithm of full conditional distribution with respect to $\boldsymbol{\theta}_k$ is $0.5n\left\{D_k^{-1}D_k^{(1)}-\left(\boldsymbol{s}_k+\rho\gamma_k\right)\right\}$, where $D_k^{(1)}=-2\Theta_{-kk}^{-1}\theta_k$ is the first-order derivative of D_k with respect to $\boldsymbol{\theta}_k$. The second-order derivative matrix of the logarithm of the full conditional distribution with respect to $\boldsymbol{\theta}_k$ equals $-0.5n\left\{D_k^{-1}\left(D_k^{-1}D_k^{(1)}D_k^{(1)\mathrm{T}}+D_k^{(2)}\right)\right\}$, where $D_k^{(2)}=-2\Theta_{-kk}^{-1}$ is the second-order derivative of D_k with respect to $\boldsymbol{\theta}_k$. Therefore, the covariance matrix of the proposal density is $V_k=cD_k\left(D_k^{-1}D_k^{(1)}D_k^{(1)\mathrm{T}}-D_k^{(2)}\right)^{-1}|_{\Theta=Q}$, where Q is a suitable estimate of Φ (such as S^{-1}, $(S+aI)^{-1}$, $a>0$, etc.) and $c>0$ is the variance tuning factor discussed below. Note that V_k is positive definite almost surely as long as Q is positive definite. Our proposal density is therefore taken as $q\left(\theta_k\right)\equiv N_{d-1}\left(\theta_k^t,V_k\right)$, where θ_k^t is the current value of the k-th off-diagonal column at iteration t. If x is the proposed value for θ_k^{t+1}, then the Metropolis-Hastings acceptance probability is $\alpha=\min\left\{1,p\left(x|y,\theta_{kk},\Theta_{-kk}\right)/p\left(\theta_k^t|y,\theta_{kk},\Theta_{-kk}\right)\right\}$. Therefore, we set $\theta_k^{t+1}=x$ with probability α and $\theta_k^{t+1}=\theta_k^t$ with probability $1-\alpha$.

There are several possible sampling strategies. We could sample Φ one element at a time, but that will be on the order of d^2, which is less efficient and ignores the possible correlations between the elements in the same column. We could also sample only the lower triangular off-diagonal elements, in which we would sample the $d-1$ vector $(\theta_{12},\ldots,\theta_{1d})$ first, the $d-2$ vector $(\theta_{23},\ldots,\theta_{2d})$ second, and so on. This would update all the elements of Φ by virtue of symmetry, which might be the most efficient way of sampling. However, this sampling procedure still ignores the correlations between the upper triangular elements and the lower triangular elements within the same column. We recommend sampling the whole off-diagonal column all at once, which yields an algorithm on the order of d. Updating the whole off-diagonal column has another advantage in that each $\theta_{kk'}(k\neq k')$ has two chances to get updated. We update $\theta_{kk'}$ when we update column k and again when we update column k' due to $\theta_{kk'}=\theta_{k'k}$. For each cycle, the latter updated value of $\theta_{kk'}$ will replace the first updated value. Thus, this will result in one-step thinning to reduce autocorrelations between samples. Thus the actual replacement rates for the individual elements (θ_{kk}'s) are higher than the acceptance rates of the columns $\boldsymbol{\theta}_k$. Our computations show that the replacement rate is roughly (1 – acceptance rate)^2, implying that the acceptance of column k and column $k'(k\neq k')$ are nearly independent. This implies that, if we target an average replacement rate of 36%, which is enough for an ideal sampling scheme, we will need an average acceptance rate for a column to be around 20%. Therefore, we can use fewer tries and/or a larger variance to obtain an ideal sampling scheme.

Variance tuning will, in most cases, result in shrinkage. We tune the variance in cases where the estimate Q of the parameter Φ leads to an unusually high variance of the proposal density. Such a situation can lead to too many draws of multiple try method, small acceptance rates, and high autocorrelations among sampled elements. This can also happen when we take $Q = S^{-1}$, where S^{-1} is still positive definite but the sample size is small relative to the dimension, leading to an inflated V_k. For high-dimensional cases, when S is singular or close to singular, we can choose $Q = (S + aI)^{-1}$ for a suitable $a > 0$, that is we add a small constant to the diagonals to make Q positive definite. This can also help in making Q more stable when n is not sufficiently large compared to d, since for larger d/n the smaller eigenvalues approach zero to destabilize the inversion.

Shrinking the variance too much can lead to a failure in exploring the full range of values for θ_k and also result in high autocorrelations among the elements. Similar problems also arise when there is no shrinkage at all. Thus, in order to optimize the acceptance rates, we shrink the variance moderately and combine that with the multiple try method proposed by [17] with some modifications as discussed below. A combination of shrinkage and multiple tries is necessary since we have the positive definiteness constraint coupled with the high dimension d of Φ. Figures 1 and 2 show the trace plots and autocorrelations for 3 different choices of the proposal density variance. Ideal shrinkage will lead to nice looking trace plots and greatly reduce the autocorrelations among successive values. The use of multiple tries can lead to faster convergence requiring fewer burn-in samples. We can now formally state our algorithm for the k-th off-diagonal column as follows:

1. Draw m independent vectors, $w_1, \ldots, w_m$ from the symmetric proposal density $N_{d-1}\left(\theta_k^t, V_k\right)$, where m is the number of tries; in our simulation we choose $m = 5$.
2. If $I\left(\theta_{kk} - \boldsymbol{w}_j^T \Theta_{-kk}^{-1} \boldsymbol{w}_j > 0\right) = 0$ for all $j = 1, \ldots, m$ then do not replace θ_k and stop; otherwise select w_j from $w_1, \ldots, w_m$ with probability proportional to $p(w_j|\theta_{kk}, \Phi_{-kk})$. Denote the selected vector as w.
3. Draw $\boldsymbol{x}_1^*, \ldots, \boldsymbol{x}_{m-1}^*$ from $N_{d-1}(w, V_k)$, and denote $\boldsymbol{x}_m^* = \theta_k^t$.
4. Replace θ_k^t by w with probability

$$\min\left\{1, \frac{p\left(\boldsymbol{w}_1|\theta_{kk}, \Theta_{-kk}\right) + \cdots + p\left(\boldsymbol{w}_m|\theta_{kk}, \Theta_{-kk}\right)}{p\left(x_m^*|\theta_{kk}, \Theta_{-kk} + \cdots + p\left(x_m^*|\theta_{kk}, \Theta_{-kk}\right)\right.}\right\},$$

where $p\left(\boldsymbol{x}_j^*\right) \propto p\left(\boldsymbol{x}_j^*|\theta_{kk}, \Theta_{-kk}\right)$.

Note that, in the above scheme V_k remains constant for all MCMC samples; $p(w_j|\theta_{kk}, \Phi_{-kk})$ and $p\left(\boldsymbol{x}_j^*|\theta_{kk}, \Theta_{-kk}\right)$ are in the same form as (6) where θ_k is replaced by w_j and $\boldsymbol{x}_j^*$, respectively.

For the BCLASSO method, we have several options for choosing the hyperparameter ρ. First, we can choose a conjugate gamma-type hyperprior for the penalty parameter. If we choose $\rho \sim Gamma(a_0, \beta_0)$, then it could be sampled using the Gibbs sampler. The full conditional of ρ is $\rho|a_0, \beta_0, \Phi, Y \sim Gamma(a_0, \beta_0 + ||\Phi||l_1)$. This choice requires choosing appropriate values of the hyperparameters a_0 and β_0; one could choose noninformative hyperpriors for large sample, however, for small sample the choice is not trivial as it has to be informative to impose penalty. An alternative is to choose the penalty parameter via

cross-validation using the log-likelihood as a maximizer; we chose 5-fold cross-validation for the optimal choice of penalty parameters for each method.

We first compute BCLASSOm, which is the minimax estimator under the L_1-penalty [29]. Since BCLASSOm estimates all of the elements of Φ as non-zero, similar to posterior means, we also compute adhoc BCLASSOs estimators by forcing credible interval-based sparsity. That is, we construct the credible intervals and force an element of BCLASSOm to zero if the interval contains zero. Sparsity can be controlled by either the penalty parameter ρ or the width of the credible interval. A larger ρ or a prior with a larger mean will lead to a more sparse matrix when the width of the credible interval is fixed. A wider credible interval will also lead to a more sparse matrix when the penalty ρ or its prior mean is fixed. We found a credible interval or around 30% to be ideal. Forcing some elements to zero can theoretically result in non-positive definite matrices, however, they are positive definite with high probability given a small credible region is chooses (we suggest below 30%). Our simulation of 600 samples have all resulted in positive definite matrices as evidenced by the ability to compute finite L_1 losses for all cases, since any zero eigenvalue will result in infinite loss and negative eigenvalue would lead to an undefined loss. This credible- interval based thresholding has probabilistic interpretation and deserves further attention in other Bayesian estimation problems in which there is a need for sparsity. The threshholding also allows network exploration since forcing some zeros is the key in such network building.

2.4 Credible Regions

Suppose we have E MCMC samples $\Phi_1, \ldots, \Phi_E$ from the posterior distribution of the d dimensional precision matrix Φ and let $\Psi_e = \log(\Phi_e)$ be the matrix logarithm of the e–th sample and $\Phi_e = \exp(\Psi_e)$ be the matrix exponential of Φ_e. Note that, if $\lambda_1, \ldots, \lambda_d$ are the eigenvalues of Φ and $\gamma_1, \ldots, \gamma_d$ are the eigenvalues of Ψ, then $\gamma_k = \log(\lambda_k)$ for $k = 1, \ldots, d$. Now, let $\bar{\Psi}$ is the posterior arithmetic mean of $\Psi_1, \ldots, \Psi_E$ then $\bar{\Theta}_G = \exp\left(\bar{\Psi}\right)$ is the posterior geometric mean of Φ_e. We define the Euclidean distance between $\Psi_e = (\psi_{e,kk'})$ and the posterior mean $\bar{\Psi} = \left(\bar{\psi}_{kk'}\right)$ given by

$$d_{E,e} = \|\Psi_e - \bar{\Psi}\|_2^2 = \left\{ \sum_{k,k'=1}^{d} \left(\psi_{e,kk'} - \bar{\psi}_{kk'}\right)^{0.5} \right\}^2.$$

Then, we sort the E samples according to the values of $d_{E,e}$ and then use ($d_{E,\alpha/2}$, $d_{E,1-\alpha/2}$) as the $(1 - \alpha)100\%$ credible region for Ψ. Finally, we obtain ($\exp(d_{E,\alpha/2})$, $\exp(d_{E,1-\alpha/2})$) as the $(1 - \alpha)100\%$ geometric confidence region for Φ.

3. SIMULATION STUDY

We used simulations to compare the performance of our BCLASSOm and BCLASSOs estimators with the three frequentist penalized likelihood methods namely, CLASSO [12], ACLASSO [11], and CSCAD [11]. Among the Bayesian methods, the [29] method uses shrinkage on the eigenvalues. This is infeasible in our non-full rank setting as some of the eigenvalues are zero since the dimension of Φ is larger than the sample size (hence the matrix is singular). In [27] and [28], an element-wise sampling was used and does not specify a recognizable prior on the precision (covariance) matrix. We restrict our comparison to permutation invariant methods that work for non-full rank data, use priors and l_1-type penalties directly on the elements of the precision matrix, and perform simultaneous shrinkage and estimation.

For the simulation, we fixed the dimensionality d and considered 3 unstructured and 3 structured matrix types. Among the unstructured types, the sparse matrix has at least 80% zeros on the off-diagonals, the moderately sparse one has at least 40% zeros on the off-diagonals, and the dense matrix has less than 5% zeros on the off-diagonals. The structured matrix types are tri-diagonal, autoregressive order one (i.e., AR(1)), and diagonal. In each case, we first generated a precision matrix. Then we generated 100 datasets for a non-full rank case where the sample size is less than the dimension ($d = 20$, $n = 10$) and compared the performance of each method based on those 100 samples.

We relied on a Cholesky decomposition to generate the 3 unstructured positive definite precision matrices of different sparsity levels. We generated a matrix A such that $A_{kk} = 1$, $A_{kk'} = U[-.5, .5]$ with probability p and $A_{kk'} = 0$ with probability $1 - p$ for $k < k'$, and $A_{kk'} = 0$ for $k > k'$. Then we computed $\Phi = AA^{\mathrm{T}}$ and $\Sigma = \Phi^{-1}$. The degree of sparsity was controlled by p, where a smaller p leads to a more sparse matrix. A tridiagonal precision matrix results in an AR(1) covariance matrix. In this case, the elements of the covariance matrix Σ are $\sigma_{kk'} = \exp(-q|r_k - r_{k'}|)$, where $r_1 < \ldots < r_d$ for some $q > 0$. Here, we chose $r_k - r_{k-1}$ to be $i.i.d$ from $U[0.5, 1]$ for $i = 2, \ldots, d$. An AR(1) precision matrix results in a tridiagonal covariance matrix and we generated the elements $\theta_{kk'} = \exp(-q|r_k - r_{k'}|)$ as above. A diagonal precision matrix results in a diagonal covariance matrix; in this case, we generated the diagonal elements of Σ where σ_{kk} are independently generated from U[1, 1.25] for $k = 1, \ldots, d$. For the BCLASSOs estimators we used thresholding on the elements of BCLASSOm based on 30% credible intervals. This choice of the credible intervals is arbitrary and will depend on the choice of the penalty parameter ρ or the value of the hyperparameters on the prior of ρ.

3.1 Criteria for comparison

There are several loss measures proposed for evaluating the performance in estimation of the precision and covariance matrices as discussed in [29]. Among these, the entropy loss, denoted as L_1, and the quadratic loss, denoted as L_2, are the most commonly used. The L_1 and L_2 loss functions for Φ are defined as

$$\begin{aligned} L_1\left(\Theta, \widehat{\Theta}\right) &= tr\left(\Theta^{-1}\widehat{\Theta}\right) - \log\ \det\left(\Theta^{-1}\widehat{\Theta}\right) - d, \\ L_2\left(\Theta, \widehat{\Theta}\right) &= tr\left(\Theta^{-1}\widehat{\Theta} - I\right)^2. \end{aligned} \tag{7}$$

where $\mathrm{vec}(A) = (a_{11}, \cdots, a_{1d})\ a_{d1}, \cdots, a_{dd})^T$ for any $d \times d$ matrix $A = (a_{ij})$. Similar loss functions for Σ will result in the Bayes estimators $\widehat{\Sigma}_{L_1} = \{E(\Theta|Y)\}^{-1}$ and

$$\mathrm{vec}\left(\widehat{\Sigma}_{L_2}\right) = \{E(\Theta \otimes \Theta|Y)\}^{-1}\mathrm{vec}\{E(\Theta|Y)\}^{-1},$$

respectively. We use $\widehat{\Theta}_{L_1} = \{E(\Sigma|Y)\}^{-1}$ and $\widehat{\Sigma}_{L_1} = \{E(\Theta|Y)\}^{-1}$ in our simulation studies as the BCLASSO estimators for Φ and Σ, respectively. Since $\widehat{\Theta}_{L_2}$ and $\widehat{\Sigma}_{L_2}$ are computationally less efficient, requiring inversion of a $d^2 \times d^2$ matrix at each step of the Monte-Carlo sampling, we do not use them in our simulation. Our estimators $\widehat{\Theta}_{L_1}$ and $\widehat{\Sigma}_{L_1}$ in the simulation are Bayes under the L_1 loss, but not under the L_2 loss. Nevertheless, we were able to achieve reasonable L_2 loss for $\widehat{\Theta}_{L_1}$ and $\widehat{\Sigma}_{L_1}$ in our non-full rank simulation cases. Moreover, using L_1-Bayes estimators is more intuitive since we are using an L_1 penalty. Another measure known as the matrix correlation was defined by [10] as

$R\left(\Theta,\widehat{\Theta}\right)=\mathrm{tr}\left(\Theta\widehat{\Theta}\right)/\left\{\mathrm{tr}\left(\Theta\Theta\right)\mathrm{tr}\left(\widehat{\Theta}\widehat{\Theta}\right)\right\}^{1/2}$. In this measure, the closer the estimator $\widehat{\Theta}$ is to Φ, the higher the value of $R\left(\Theta,\widehat{\Theta}\right)$. We compared our estimates $\widehat{\Theta}_{L_1}$ and $\widehat{\Sigma}_{L_1}$ with the CLASSO, ACLASSO, and CSCAD methods for the L_1 loss, the L_2 loss, and the matrix correlation based on 6 different matrix types of dimension 20. For each of the 6 matrix types, we used 100 Markov chain Monte Carlo (MCMC) samples of size 10 each. For all cases we choose $Q = (S + aI)^{-1}$ with $a = 0.1$, the number of tries as $m = 5$, the value of $c = 0.5$ was chosen to get about 30% acceptance rate. We collected 10, 000 MCMC samples after 5, 000 burn-in, which gave us an average computation time of about 10 minutes for each simulation.

We can also define the L_1 and L_2 loss functions for in a similar fashion. The optimal estimators minimize these loss functions. [29] showed that the Bayes (hence minimax) estimators of Φ under L_1 and L_2 are, respectively, given by

$$\widehat{\Theta}_{L_1}=\widehat{\Theta}_{L_1}=\{E\left(\Sigma|Y\right)\}^{-1}$$
$$\mathrm{vec}\left(\widehat{\Theta}_{L_2}\right)=\{E\left(\Sigma\otimes\Sigma|Y\right)\}^{-1}\mathrm{vec}\{E\left(\Sigma|Y\right)\}^{-1}.$$

3.2 Results

Table 1 summarizes the mean L_1 losses and their standard deviations for the six types of precision and covariance matrices. The CSCAD method performs poorly in terms of L_1 loss for small sample non-full rank cases for all types of structures in both the precision and covariance matrices. For both the precision and covariance matrices, CLASSO, SPICE, ACLASSO, BCLASSOm, and BCLASSOs perform similarly. Table 2 summarizes the mean L_2 losses and their standard deviations for these four methods. For all structures, except the diagonal case, CSCAD is worse than CLASSO, SPICE, ACLASSO, BCLASSOm and BCLASSOs, while these five methods perform somewhat similarly for all six structures compared. Only for the diagonal precision matrix does CSCAD perform the best among the 5 methods compared. Table 3 summarizes the mean matrix correlations and their standard deviations. In terms of the matrix correlation measure $R\left(\Theta,\widehat{\Theta}\right)$, CLASSO, BCLASSOm and BCLASSOs perform somewhat similarly in both the precision and covariance matrices. The ACLASSO and SPICE methods perform similarly in the precision matrix, but they are worse than BCLASSO and CLASSO in the covariance matrix. The CSCAD method performs the worst among all the methods in both the precision and covariance matrices for all six types of structures considered. As evident from Tables 1 and 2, although there is minimal or no loss in credible interval based sparsity in the precision matrix, there are substantial gains in matrix loss for the covariance matrix. The SPICE estimator seems to improve on the covariance over CLASSO. Performance of both SPICE and CSCAD improves when sparsity increases. The poor performance of CSCAD is somewhat surprising due to small sample sizes.

4. APPLICATION TO REAL DATA

Example 1: The first dataset is flow cytometry data on $d = 11$ proteins on $n = 7466$ cells from [23]. In [23], a Bayesian network was developed and elucidated most of the signaling relationships reported traditionally and also predicted novel interpathway network causalities, which were verified through experiments. The data was also used by [12] for comparison of the agreements of CLASSO under different values of the penalty parameter. The data was generated from 9 simulations on 11 proteins. We adjusted the data for a random simulation effect as well as fixed effects of simulation and protein. Our purpose was

to build a network between proteins via partial correlations (via Φ). For the maximum likelihood network in Figure 4(b), we used a hard-threshold that gives the same number of connections as those of [23]. The penalties for the CLASSO in Figure 4(d), ACLASSO in Figure 4(e) and CSCAD in Figure 4(f), were obtained through 10-fold cross validation. Since the penalties based on cross validation resulted in a more sparse network for these 3 frequentist methods than that of Sachs, we decided to fix the penalty manually to get the same number of connections. The results are shown in Figures 4 (g), 5(h) and 5(i), respectively. For CSCAD, no matter how small ρ is, the number of connections does not increase after a certain point. Finally, for BCLASSO, we used a gamma prior for the penalty parameter, $\rho \sim Gamma(1, 1)$, and used 80,000 MCMC samples after 20,000 burn-ins to obtain posterior means and credible intervals. We constructed credible intervals of different widths for each element as shown in Table 4; the Bayesian network that has the closest number of connections to that of Sachs is shown on Figure 4. The level of agreement of each of these 4 methods to those of Sachs' results were computed and reported in Table 4. The results indicate similar agreement between the networks of BCLASSO, ACLASSO and CLASSO and [23]'s network when the number of connections are similar.

Example 2: While it is well recognized that the human brain forms large scale networks of distributed and interconnected neuronal populations, the study of different brain networks has been hampered by the lack of non-invasive tools. Recently, the introduction of the resting functional connectivity MRI approach offers, potentially, a potent tool, to specifically alleviate this difficulty, allowing a direct investigation of a wide array of brain networks. Researchers are often interested in exploring the brain networks through partial correlations where the connection between two regions is explored after removing the effect of all other regions.

The data consists of average fcMRI signals from 90 brain regions ($d = 90$) of 30 2-year old children ($N = 30$). All images were acquired on a 3T MR scanner with a gradient echo-planar imaging sequence. The imaging sequence was repeated 150 times. The images of the first 10-20 time points were typically excluded from the data analysis to ensure that magnetization reaches the steady state. All subjects are healthy normal controls and imaged at sleep without sedation. In this study, the signals were obtained from the remaining 130 repeats ($T = 130$, so that $n = NT = 3900$). Our primary purpose here is to build a network between regions after adjusting for subject effects and region specific means. Let Y_{ijk} ($i = 1, \ldots, N$; $j = 1, \ldots, T$; $k = 1, \ldots d$) represent the adjusted average fcMRI signal from subject i at repeat (time) j in region k. Then Y_{ij} is the d-dimensional vector of adjusted responses from subject i on repeat j and $Y_{ij} \sim N_d(0, \Sigma)$. Let $n = NT$, the joint distribution of Y is given by

$$p(Y|\Theta) \propto \{\det(\Theta)\}^{\frac{n}{2}} \quad \exp\left\{-\frac{1}{2}\sum_{i=1}^{N}\sum_{j=1}^{T} Y_{ij}^{\mathrm{T}}\Theta Y_{ij}\right\} I(\Theta \succ 0).$$

The posterior distribution of Φ under the lasso penalty can be written as in (3) and the full conditionals are given in (6). For the penalty parameter ρ, we take $\rho \sim$ Gamma(1, 1). For thresholding, we construct credible intervals of different widths to control sparsity. For CLASSO, ACLASSO, and CSCAD, we used 10-fold cross validation to choose the optimal penalty. We report the resulting precision matrices in Figure 5 and the networks in Figure 6. The summary statistics of the number of connections along with the global efficiencies E_{glob} (a measure of how efficiently the regions communicate in the whole brain) and local efficiencies E_{loc} (a measure of how efficiently the regions in each local area communicate) are reported in Table 5.

The CSCAD method performs poorly compared to the other three methods and shows very few connections across the entire brain, leading to rather low global and local efficiencies. This result contradicts the well formed brain networks of 2 year olds, which has been reported in the literature using both imaging and behavioral approaches [14]. In contrast, CLASSO, ACLASSO, and BCLASSO appear to provide more similar results, demonstrating well connected brain networks. Although there are differences in the regions with the highest number of connections, some consistent patterns are observed from CLASSO, ACLASSO, and BCLASSO. The brain regions that exhibit the highest number of connections with other regions are consistently shown by these three methods in the temporal, frontal and occipital lobes. These results suggest that even at the age of 2 years, children develop well connected networks, particularly in the temporal and frontal areas. More studies are clearly needed to further determine how the proposed approach is capable of better delineating the development of brain networks across different age groups. The top regions picked up by the different methods are listed in Table 5. The BCLASSO results are based on thresholding with a 70% credible interval. This choice was made in order to closely align the total number of connections from BCLASSO to that of CLASSO and ACLASSO.

5. DISCUSSION

We have introduced a general class of priors for the precision matrix which yield the ACLASSO, CLASSO, and SPICE penalties as special cases. We have also developed a sampling scheme for the estimation of the precision and covariance matrices under a special case that corresponds to the lasso penalty, which can facilitate exploration of the full posterior distribution of the matrix under L_1 penalites. Although our proposed priors do not guarantee positive definiteness of Φ, we have developed a fast sampling scheme that guarantees positive definite MCMC samples of the precision matrix at each iteration regardless of the value of the penalty parameter. Our proposed method is the first Bayesian method that uses priors that directly translate into the L_1 penalty, the method works well for non-full rank data, and performs shrinkage and estimation simultaneously. Simulations show that BCLASSO performs similarly to CLASSO, SPICE and ACLASSO for non-full rank data when the sample size is small, while performing better than CSCAD. We will further develop an efficient algorithm to sample from $p(\theta_k|y, \theta_{kk}, \rho)$. The proposed method can be easily extended to more complex models that account for subject-specific variation for building networks in longitudinal data. The priors can be generalized to independent gamma priors for the diagonal elements; $\theta_{kk} \sim$ gamma(a_k, β_k) and independent double gamma priors for the off-diagonal elements $\theta_{kk'} \sim$ double gamma(0, $a_{kk'}$, $b_{kk'}$) for $k > k'$; that is, $p(\theta_{kk'}) \propto |\theta_{kk'}|^{a_{kk'}-1} \exp(-b_{kk'}|\theta_{kk'}|)$, where $a_{kk'} > 0$ and $b_{kk'} > 0$. Then, the posterior distribution of Φ is given by

$$p(\Theta|Y) \propto (\det \Theta)^{\frac{n}{2}} \prod_{k=1}^{d} \theta_{kk}^{\alpha_k - 1} \prod_{k=2}^{d} \prod_{k'=1}^{k-1} |\theta_{kk'}|^{a_{kk'}-1} \exp\left\{ -\tfrac{n}{2} tr(S\Theta) - \sum_{k=1}^{d} \beta_k \theta_{kk} - \sum_{k=2}^{d} \sum_{k'=1}^{k-1} b_{kk'} |\theta_{kk'}| \right\},$$

This is particularly attractive for Bayseian analysis since appropriate choice of shape and scale paramters can lead to an infinite spike of the prior at 0 and heavier tails leading to larger shrinkage of smaller parameters and smaller shrinkage of larger parameters compared to L_1-penalties.

Like many Bayesian methods, scalibility to larger dimensions is a challenge for BCLASSO. Nevertheless, the posterior estimators for dimensions up to 50 do well and networks

dimension near 100 works similar to CLASSO and ACLASSO as evidenced by the brain imaging data example. The main advantage of a fully Bayesian approach is the ability to sample the whole posterior distribution instead of just estimating the posterior mode.

Acknowledgments

We thank Professor Robert Tibshirani for the “glasso” package in R and making valuable comments. We also thank Professor Jianqing Fan for sharing his code and Professor Mohsen Pourahmadi for making valuable suggestions.

Appendix I

Proof of Theorem 1: Without loss of generality, we partition and rearrange the columns of current $\Phi^{(t)}$ as

$$\Theta^{(t)}=\begin{pmatrix} \Theta^{(t)}_{-kk} & \theta^{(t)}_{k} \\ \theta^{\mathrm{T}(t)}_{k} & \theta^{(t)}_{kk} \end{pmatrix} \succ 0.$$

Since $\Theta^{(t)}_{-kk} \succ 0$, all of its diagonal elements are positive and all the leading determinants are positive. In particular, $\left(\Theta^{(t)}_{-kk}\right)^{-1} \succ 0$ and so there exists an A_k such that $\left(\Theta^{(t)}_{-kk}\right)^{-1}=A_k A_k^{\mathrm{T}}$ based on the Cholesky decomposition. We only need to show that after updating the first column (θ_{kk} and $\boldsymbol{\theta}_k$), the diagonal element $\theta^{(t+1)}_{kk}$ is positive and the last determinant is positive, i.e., $\det(\Theta)^{(t+1)}=D^{(t+1)}_k \det\left(\Theta^{(t)}_{-kk}\right)>0$; this implies $D^{(t+1)}_k>0$ since $\Theta^{(t)}_{-kk} \succ 0$. When we update the last column (or equivalently row) we have

$$C^{(t)}_k=\left(\theta^{(t)}_k\right)^{\mathrm{T}}\left(\Theta^{(t)}_{-kk}\right)^{-1}\theta^{(t)}_k=\left(A^T_k\theta^{(t)}_k\right)^{\mathrm{T}}A^{\mathrm{T}}_k\theta^{(t)}_k=\eta^{\mathrm{T}}\eta \geq 0.$$

Thus, we have

$$\theta^{(t+1)}_{kk}=C^{(t)}_k+Gamma\left(\frac{n}{2}+1,\frac{n}{2}\left(s_{kk}+\rho\right)\right)>0.$$

Let x be the proposed value for $\theta^{(t+1)}_k$. Then $D_x=\left(\theta^{(t+1)}_k\right)-x^{\mathrm{T}}\left(\Theta^{(t)}_{-kk}\right)^{-1}x$. Thus, $p\left(x|Y,\theta^{(t+1)}_{kk},\Theta^{(t)}_{-kk}\right)$ is proportional to

$$D_x^{\frac{n}{2}} \quad \exp\left\{-\frac{n}{2}\left(\boldsymbol{s}_k+\rho\gamma^t_k\right)^{\mathrm{T}}\theta^t_k\right\} I\left(D_x>0\right).$$

However, $p\left(x|Y,\theta^{(t+1)}_{kk},\Theta^{(t)}_{-kk}\right)=0$ whenever $D_x \leq 0$, so that the Metropolis acceptance probability is

$$\alpha=\min\left\{1,\frac{p\left(x|Y,\theta^{(t+1)}_{kk},\Theta^{(t)}_{-kk}\right)}{p\left(\theta^{(t)}_k|Y,\theta^{(t+1)}_{kk},\Theta^{(t)}_{-kk}\right)}\right\}=0.$$

That is, we can only accept the proposed value when $D_x > 0$. Thus, $\theta_k^{(t+1)}=x \Rightarrow D_x>0$. This is true for any column of Φ that we update, therefore leading to positive definite values of Φ at any stage in the updating process. This completes the proof.

Derivation of (6): Let $\rho \sim$ Gamma(α_0, $0.5n\beta_0$). Then the joint posterior of (Φ, ρ) becomes

$$\begin{aligned} p(\Theta,\rho|Y) &= p(\Theta,\rho|S) \propto \det(\Theta)^{n/2}\rho^{\alpha_0-1} \\ &\quad \exp\left(0.5n\left\{-\mathrm{tr}(S\Theta)-\rho\|\Theta\|_{l_1}-\rho\beta_0\right\}\right), \end{aligned}$$

which yields that

$$p(\rho|\Theta,Y) \propto \rho^{\alpha_0-1} \quad \exp\left\{-0.5n\rho\left(\beta_0+\|\Theta\|_{l_1}\right)\right\},$$

which is the kernel of a Gamma(α_0, $0.5n(\beta_0 + \|\Phi\|l_1)$) distribution. Now, using the partition in (5), we can write $\det(\Phi) = \det(\Phi_{-kk})D_k$, where $D_k=\left(\theta_{kk}-\theta_k^{\mathrm{T}}\Theta_{-kk}^{-1}\theta_k\right)$. Moreover, $\|\Phi\|l_1 = \|\Phi_{-kk}\|l_1 + 2\|\boldsymbol{\theta}_k\|l_1 + \theta_{kk}$ and $\mathrm{tr}(S\Phi)$ is given by

$$\begin{aligned} \mathrm{tr}(S\Theta) &= tr\left\{\begin{pmatrix} S_{-kk} & \boldsymbol{s}_k \\ \boldsymbol{s}_k^{\mathrm{T}} & s_{kk}\end{pmatrix}\begin{pmatrix} \Theta_{-kk} & \theta_k \\ \theta_k^{\mathrm{T}} & \theta_{kk}\end{pmatrix}\right\} \\ &= \mathrm{tr}\begin{pmatrix} S_{-kk}\Theta_{-kk}+\boldsymbol{s}_k\theta_k^{\mathrm{T}}, & S_{-kk}\theta_k+\boldsymbol{s}_k\theta_{kk} \\ \boldsymbol{s}_k^{\mathrm{T}}\Theta_{-kk}+s_{kk}\theta_k^{\mathrm{T}}, & s_k^{\mathrm{T}}\theta_k+s_{kk}\theta_{kk}\end{pmatrix} \\ &= tr(S_{-kk}\Theta_{-kk})+2s_k^{\mathrm{T}}\theta_k+s_{kk}\theta_{kk}. \end{aligned}$$

Therefore, the joint posterior distribution of (Φ, ρ) can be written as

$$\begin{aligned} &= p(\Theta,\rho|Y,\rho) = \det(\Theta)^{n/2} \quad \exp\left(0.5n\left\{-\mathrm{tr}(S\Theta)-\rho\|\Theta\|_{l_1}\right\}\right) \\ &= \det(\Theta_{-kk})^{n/2}D_k^{n/2} \quad \exp\left(0.5n\left\{-\mathrm{tr}(S_{-kk}\Theta_{-kk})-2s_k^{\mathrm{T}}\theta_k-s_{kk}\theta_{kk}-\rho\left(\|\Theta_{-kk}\|_{l_1}+2\|\theta_k\|_{l_1}+\theta_{kk}\right)\right\}\right). \end{aligned}$$

After dropping unnecessary constants, we have

$$\begin{aligned} p(\theta_{kk}|Y,\theta_k,\Theta_{-kk},\rho) &\propto D_k^{\frac{n}{2}} \quad \exp\left\{-\tfrac{n}{2}(s_{kk}+\rho)\theta_{kk}\right\}, \\ p(\theta_k|Y,\theta_{kk},\Theta_{-kk},\rho) &\propto D_k^{\frac{n}{2}} \quad \exp\left\{-\tfrac{n}{2}\left(s_k^{\mathrm{T}}\theta_k+\rho\|\theta_k\|_{l_1}\right)\right\}. \end{aligned}$$

To derive the full conditional distribution of θ_{kk}, we write

$$\begin{aligned} p(\theta_{kk}|Y,\theta_k,\Theta_{-kk},\rho) &\propto D_k^{\frac{n}{2}} \quad \exp\left\{-\tfrac{n}{2}(s_{kk}+\rho)\theta_{kk}\right\} \\ &= (\theta_{kk}-C_k)^{\frac{n}{2}} \quad \exp\left\{-n/2(s_{kk}+\rho)\theta_{kk}\right\} \\ &\propto (\theta_{kk}-C_k)^{\frac{n}{2}}\exp\left\{-\tfrac{n}{2}(s_{kk}+\rho)(\theta_{kk}-C_k)\right\}. \end{aligned}$$

Thus, $\theta_{kk} - C_k|Y, \boldsymbol{\theta}_k, \Phi_{-kk} \sim Gamma(n/2 + 1, n/2(s_{kk} + \rho))$ and this implies that $\theta_{kk}|Y, \boldsymbol{\theta}_k, \Phi_{-kk} \sim C_k + Gamma(n/2 + 1, n/2(s_{kk} + \rho))$.

REFERENCES

1. Banerjee O, Ghaoui LE, d'Aspremont A. Model selection through sparse maximum likelihood estimation. Journal of Machine Learning Research. 2007; 9:485–516.
2. Barnard J, McCulloch R, Meng X. Modeling covariance matrices in terms of standard deviations and correlations, with application to shrinkage. Statistica Sinica. 2000; 10:1281–1311.
3. Carvalho CM, Scott JG. Objective Bayesian model selection in Gaussian graphical models. Biometrika. 2009; 96:497–512.
4. Chen Z, Dunson D. Random effects selection in linear mixed models. Biometrics. 2003; 59:762–769. [PubMed: 14969453]
5. Davidson E, Levin M. Gene regulatory networks. Proc. Natl. Acad. Sci. U.S.A. 2005; 102:4935. [PubMed: 15809445]
6. Dempster AP. Covariance selection. Biometrics. 1972; 28:157–175.
7. Dobra A, Hans C, Jones B, Nevins JR, Yao G, Westb M. Sparse models for exploring gene expression data. Journal of Multivariate Analysis. 2004; 90:196–212.
8. Drton M, Perlman M. Model selection for Gaussian concentration graphs. Biometrika. 2004; 91:591–602.
9. Edwards, DM. Introduction to Graphical Modeling. Springer; New York: 2000.
10. Escoufer Y. Le traitement des variables vectorielles. Biometrics. 1973; 29:751–760.
11. Fan J, Feng Y, Wu Y. Network exploration via the adaptive LASSO and SCAD penalties. Annals of Applied Statistics. 2009; 3:521–541. [PubMed: 21643444]
12. Friedman J, Hastie T, Tibshirani R. Sparse inverse covariance estimation with the graphical lasso. Biostatistics. 2008; 9:432–441. [PubMed: 18079126]
13. Frühwirth-Schnatter S, Tüchler R. Bayesian parsimonious covariance estimation for hierarchical linear mixed models. Stat Comput. 2008; 18:1–13.
14. Gao W, Zhu H, Giovanellom K, Smith J, Shen D, Gilmoref J, Lin W. Evidence on the emergence of the brain default network from 2-week-old to 2-year-old healthy pediatric subjects. PNAS. 2009; 106:6790–6795. [PubMed: 19351894]
15. Haff LR. Minimax estimators for a multinormal precision matrix. Journal of Multivariate Analysis. 1977; 7:374–385.
16. Huang J, Liu N, Pourahmdi M, Liu L. Covariance matrix selection and estimation via penalized normal likelihood. Biometrika. 2006; 93:85–98.
17. Liu JS, Liang F, Wong WH. The multiple-try method and local optimization in metropolis sampling. Journal of the American Statistical Association. 2000; 95:121–134.
18. Park T, Casella G. The Bayesian Lasso. Journal of the American Statistical Association. 2008; 103:681–686.
19. Pourahmadi M. Maximum likelihood estimation of generalized linear models for multivariate normal covariance matrix. Biometrika. 2008; 87:425–435.
20. Paul D. Asymptotics of sample eigenstructure for a large dimensional spiked covariance model. Statistica Sinica. 2007; 17:1617–1642.
21. Rajaratnam M, Carvalho. Flexible covariance estimation in Gaussian graphical models. Annals of Statistics. 2008; 36:2818–2849.
22. Rothman AJ, Bickel PJ, Levina E, Zhu J. Sparse permutation invariant covariance estimation. Electronic Journal of Statistics. 2008; 2:495–515.
23. Sachs K, Perez O, Pe'er D, Lauffenburger D, Nolan G. Causal protein-signaling networks derived from multiparameter single cell data. Science. 2003; 308:523–529. [PubMed: 15845847]
24. Schäfer J, Strimmer K. A shrinkage approach to large-scale covariance matrix estimation and implications for functional genomics. Genetics and Molecular Biology. 2005; 4:32.
25. Schur I. On the characteristic roots of a linear substitution with an application to the theory of integral equations. Mathematische Annalen. 1909; 66:488–510.
26. Scott JG, Carvalho CM. Feature-inclusion stochastic search for Gaussian graphical models. Journal of Computational and Graphical Statistics. 2008; 17:790–808.

27. Smith M, Kohn R. Bayesian parsimonious covariance matrix estimation for longitudinal data. J. Am. Statist. Assoc. 2002; 87:1141–1153.

28. Wong F, Carter CK, Kohn R. Efficient estimation of covariance selection models. Biometrika. 2003; 90:809–830.

29. Yang R, Berger IO. Estimation of covariance matrix using the reference prior. The Annals of Statistics. 1994; 22:3, 1195–1211.

30. Yuan M. Efficient computation of l_1 regularized estimates in Gaussian graphical models. Journal of Computational & Graphical Statistics. 2007; 17:809–826.

31. Yuan M, Lin Y. Model selection and estimation in the Gaussian graphical model. Biometrika. 2007; 94:19–35.

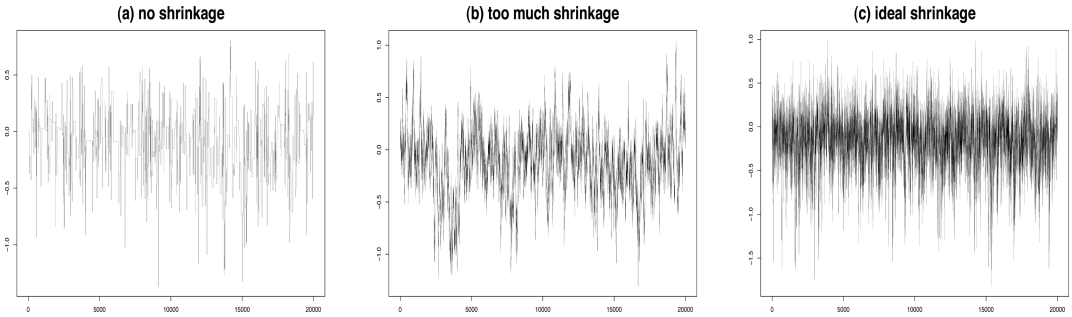

Figure 1.
Trace plots of θ_{12} for d = 5 and n = 10 showing the impact of variance tuning of the proposal density.

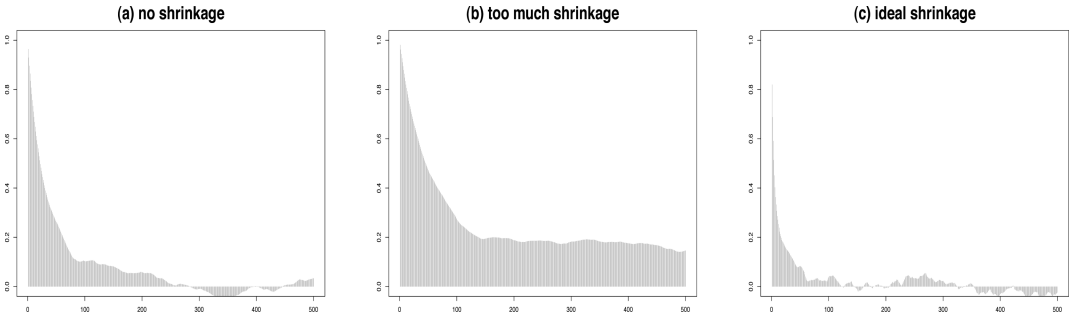

Figure 2.
Autocorrelation plots for θ_{12} for d = 5 and n = 10 showing the impact of variance tuning of the proposal density.

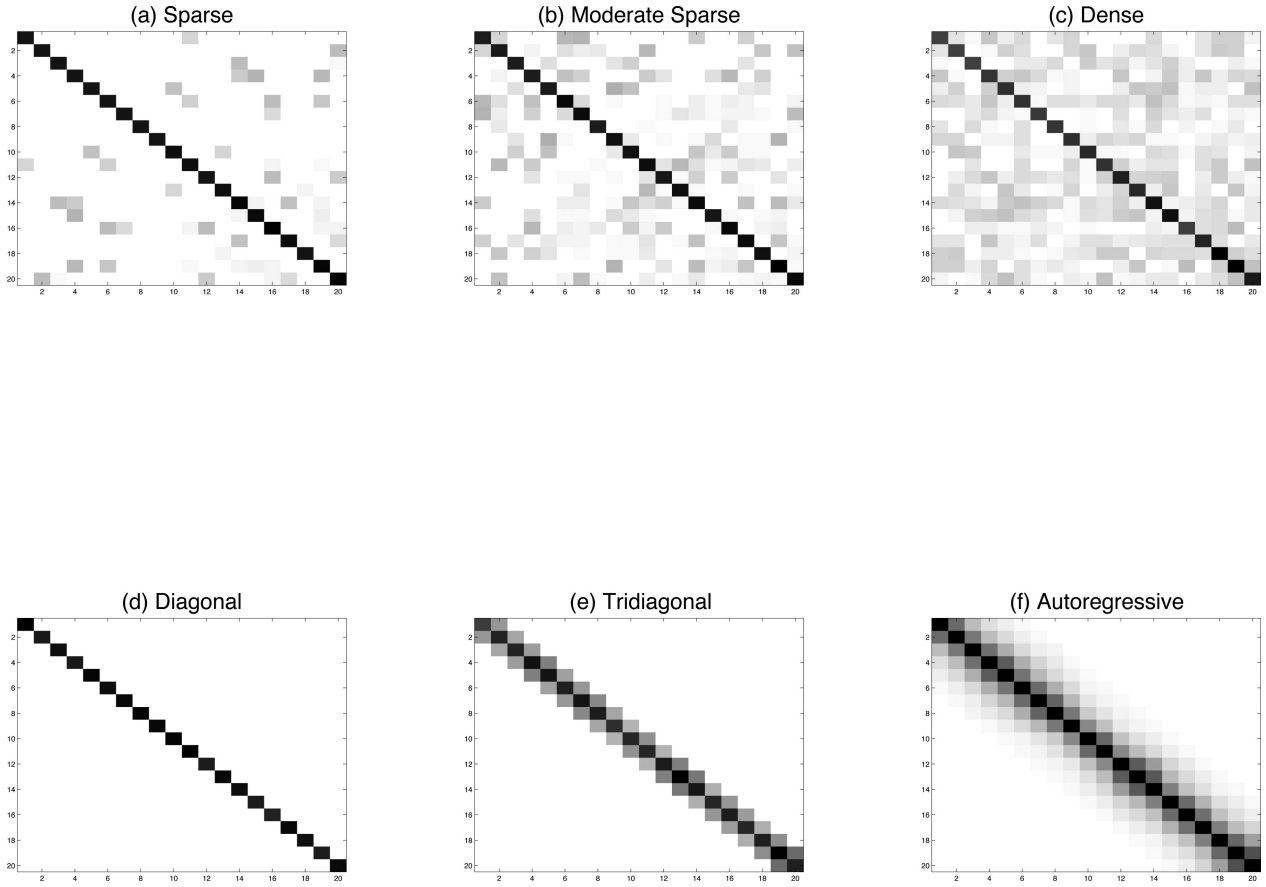

Figure 3.
Image plots of the six types of precision matrices (Φ) considered in the simulation study. The top 3 are unstructured and the bottom 3 are structured.

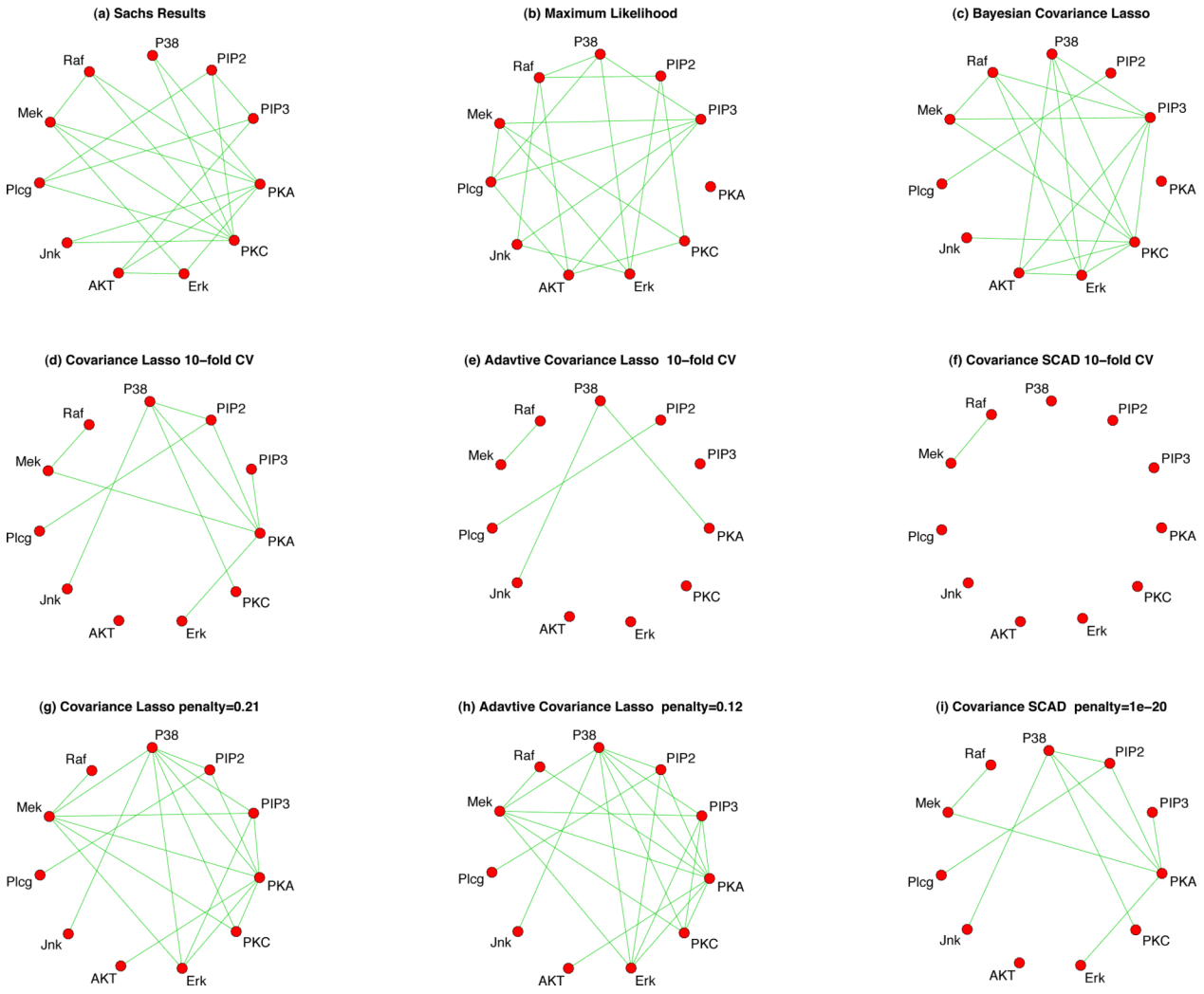

Figure 4.
Networks for 11 proteins from Sachs et al. (2003)

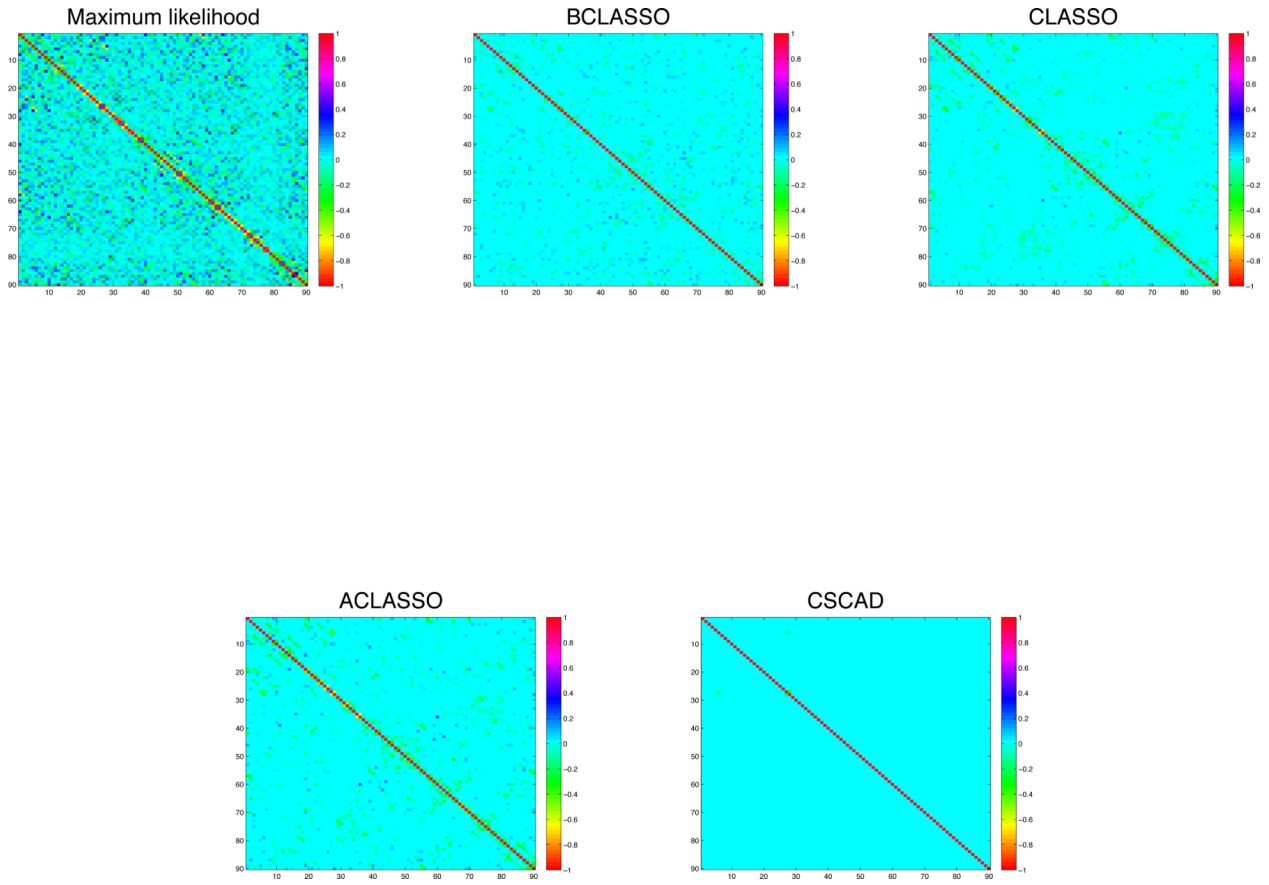

Figure 5.
Image plots of the partial correlation matrices for 90 regions of 2-year old children' brains using the five different methods

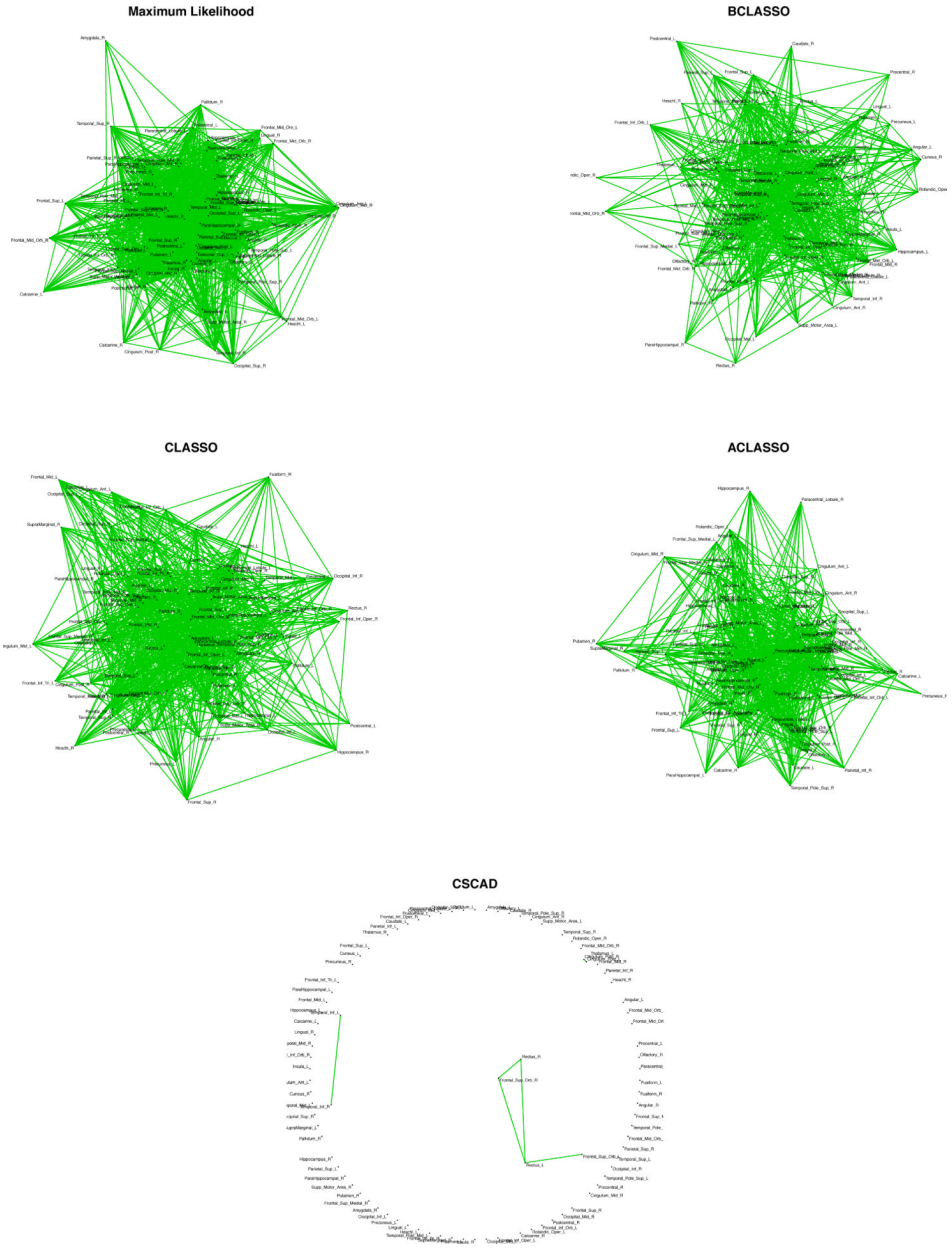

Figure 6.
Networks for 90 regions of 2-year old children' brains using the different methods

Table 1

Mean L_1 losses (and standard deviations) for the different methods

Type		CLASSO	SPICE	ACLASSO	CSCAD	BCLASSOm	BCLASSOs
Sparse	Θ	2.38(0.43)	4.18(1.91)	5.71(2.62)	14.27(12.18)	4.82 (1.11)	4.52 (0.87)
	Σ	4.03(1.03)	2.99(0.76)	5.65(1.42)	19.02(7.92)	13.65(1.70)	3.56 (0.67)
Moderately	Θ	3.29(0.52)	5.09(2.69)	6.07(2.74)	13.44(8.80)	6.09 (1.29)	5.79 (1.03)
Sparse	Σ	5.76(1.42)	4.17(0.77)	6.41(1.25)	22.84(9.61)	12.93(1.89)	5.10 (0.87)
Dense	Θ	4.90(0.53)	7.08(1.83)	7.65(2.46)	17.22(11.49)	6.87 (1.17)	6.39 (0.87)
	Σ	9.53(2.24)	6.35(1.04)	9.95(1.31)	31.91(27.49)	12.82(1.80)	6.04 (0.69)
AR(1)	Θ	5.44(0.57)	7.60(2.16)	8.12(2.14)	13.11(7.14)	7.02 (1.15)	7.00 (0.87)
	Σ	9.53(2.24)	7.48(1.29)	7.95(1.31)	31.91(27.49)	12.42(5.97)	5.97 (0.59)
Tridiagonal	Θ	5.70(0.57)	7.99(1.83)	7.80(2.32)	19.45(9.45)	10.43(1.42)	9.27 (0.95)
	Σ	11.37(2.65)	9.37(1.79)	9.19(1.48)	24.50(9.10)	12.79(1.82)	12.18(1.16)
Diagonal	Θ	2.41(2.75)	3.69(2.11)	7.00(3.79)	10.49(5.87)	4.31 (1.46)	3.98 (0.98)
	Σ	4.23(4.74)	2.43(0.79)	7.27(5.44)	18.47(9.72)	13.81(1.67)	4.47 (0.99)

Note: CLASSO = covariance lasso; ACLASSO = adaptive covariance lasso; SPICE = sparse permutation invariant covariance estimator; ACLASSO = adaptive covariance lasso; CSCAD = smoothly clipped absolute deviation for covariance; BCLASSOm = Bayesian covariance; lasso with L_1 minimax estimator; BCLASSOs = Bayesian covariance lasso with sparsity forced;

Stat Interface. Author manuscript; available in PMC 2014 February 16.

Table 2

Mean L_2 losses (and standard deviations) for the different methods

Type		CLASSO	SPICE	ACLASSO	CSCAD	BCLASSOm	BCLASSOs
Sparse	Θ	15.26(13.94)	51.45(55.79)	96.16(111.92)	497.49(1,542.28)	50.97(19.84)	51.70(19.96)
	Σ	101.15(65.48)	52.58(55.77)	19.04(22.16)	1,027.38(966.93)	215.78(17.12)	50.63(19.98)
Moderately	Θ	18.11(18.66)	58.23(137.26)	87.90(144.86)	278.80(980.90)	62.54(22.77)	63.94(23.15)
Sparse	Σ	167.56(104.94)	62.20(142.07)	44.09(55.24)	1,457.29(1,194.27)	199.12(20.64)	59.43(22.78)
Dense	Θ	11.38(12.00)	60.61(66.64)	91.98(111.98)	549.42(1,293.01)	33.09(18.97)	36.60(20.44)
	Σ	244.69(146.12)	78.22(78.46)	66.76(63.69)	1,364.45(1,702.53)	177.33(23.33)	30.09(19.33)
AR(1)	Θ	13.85(16.27)	57.19(71.12)	86.76(92.51)	208.60(664.13)	50.10(21.85)	48.95(21.75)
	Σ	318.06(189.03)	810.42(429.57)	95.49(88.67)	3,152.38(4,734.21)	30.46(18.40)	10.45(13.93)
Tridiagonal	Θ	11.80(13.44)	55.39(65.98)	76.50(100.36)	580.22(1,058.43)	153.65(24.70)	24.61(34.17)
	Σ	394.00(236.01)	17.06(17.48)	109.96(100.93)	1,433.68(1,053.86)	159.86(25.95)	88.63(27.74)
Diagonal	Θ	4.16(52.40)	53.25(63.51)	101.06(11.52)	1.02(4.34)	51.20(20.13)	51.41(20.13)
	Σ	75.39(217.11)	13.54(26.94)	17.84(34.07)	898.15(312.30)	223.66(15.74)	84.89(22.54)

Table 3

Mean matrix correlations (and standard deviations) for the different methods

Type		CLASSO	SPICE	ACLASSO	CSCAD	BCLASSOm	BCLASSOs
Sparse	Θ	0.92(0.01)	0.84(0.06)	0.87(0.02)	0.69(0.09)	0.85(0.02)	0.88(0.02)
	Σ	0.89(0.04)	0.84(0.06)	0.76(0.06)	0.80(0.09)	0.92(0.02)	0.91(0.02)
Moderately	Θ	0.90(0.01)	0.82(0.05)	0.86(0.02)	0.66(0.09)	0.83(0.02)	0.85(0.03)
Sparse	Σ	0.85(0.03)	0.80(0.05)	0.74(0.06)	0.79(0.07)	0.89(0.01)	0.86(0.02)
Dense	Θ	0.86(0.01)	0.80(0.05)	0.83(0.02)	0.64(0.10)	0.79(0.04)	0.81(0.02)
	Σ	0.79(0.04)	0.67(0.05)	0.72(0.05)	0.70(0.05)	0.80(0.02)	0.73(0.02)
AR(1)	Θ	0.79(0.02)	0.71(0.06)	0.78(0.02)	0.59(0.07)	0.80(0.02)	0.80(0.02)
	Σ	0.78(0.03)	0.66(0.05)	0.72(0.04)	0.69(0.05)	0.75(0.02)	0.73(0.02)
Tridiagonal	Θ	0.87(0.02)	0.80(0.05)	0.85(0.02)	0.66(0.09)	0.75(0.01)	0.78(0.03)
	Σ	0.81(0.03)	0.74(0.05)	0.72(0.04)	0.77(0.05)	0.85(0.01)	0.79(0.03)
Diagonal	Θ	0.92(0.95)	0.87(0.07)	0.85(0.89)	0.66(0.77)	0.88(0.02)	0.90(0.02)
	Σ	0.86(0.95)	0.86(0.06)	0.72(0.78)	0.85(0.86)	0.95(0.01)	0.94(0.02)

Table 4

Agreement of Methods with the Results from Sachs et al. (2003)

Method	No	Se	Sp	PPV	NPV
Sachs	19	1.00	1.00	1.00	1.00
Maximum Likelihood	20	0.37	0.64	0.35	0.66
BCLASSO 10%	30	0.58	0.47	0.37	0.68
BCLASSO 20%	21	0.47	0.67	0.43	0.71
BCLASSO 25%	18	0.42	0.72	0.44	0.70
BCLASSO 30%	13	0.32	0.81	0.46	0.69
BCLASSO 35%	13	0.32	0.83	0.46	0.72
BCLASSO 40%	8	0.26	0.92	0.63	0.70
BCLASSO 50%	8	0.26	0.92	0.63	0.70
CLASSO 10-fold CV	10	0.32	0.89	0.60	0.71
ACLASSO 10-fold CV	4	0.16	0.97	0.75	0.69
SCAD 10-fold CV	1	0.05	1.00	1.00	0.67
LASSO $\rho = 0.21$	19	0.47	0.72	0.47	0.72
ACLASSO $\rho = 0.12$	19	0.47	0.72	0.47	0.72
SCAD $\rho = 10^{-3}$	10	0.32	0.89	0.60	0.71
SCAD $\rho = 10^{-20}$	10	0.32	0.89	0.60	0.71

Note: CI = credible interval: No = Number of connections; Se = Sensitivity; Sp = Specificity; PPV = Positive predictive value; NPV = Negative predictive value.

Table 5

Regions With the Highest Number of Connections Picked by the Four Methods

Maximum Likelihood		Covariance Lasso		Adaptive Covariance Lasso		Bayesian Covariance Lasso	
Temporal Pole Mid L	35	Rectus L	33	Temporal Inf L	27	Occipital Inf R	32
Frontal Mid L	31	Temporal Inf L	30	Temporal Inf R	24	Temporal Sup L	31
Precentral R	29	Temporal Inf R	28	Rectus L	23	Frontal Inf Oper R	30
Occipital Sup R	29	Cingulum Post L	28	Supp Motor Area L	23	Angular R	29
Fusiform L	28	Frontal Sup Orb R	27	Cingulum Post L	22	Temporal Mid R	29
Temporal Inf L	28	Heschl L	26	Heschl L	20	Amygdala L	28
Temporal Inf R	28	Supp Motor Area L	25	Frontal Sup Orb R	19	Frontal Mid Orb L	26
Temporal Pole Sup R	27	Frontal Mid Orb R	24	Paracentral Lobule L	19	Frontal Inf Tri L	26
Temporal Pole Mid R	27	Paracentral Lobule L	24	Precentral R	19	Cingulum Mid L	25
Precentral L	26	Olfactory L	24	Olfactory L	18	Parietal Inf L	25
Frontal Sup L	26	Parietal Sup R	23	Occipital Mid R	18	Occipital Sup L	25
Frontal Inf Orb R	26	Frontal Mid Orb R	22	Temporal Pole Mid L	18	Frontal Mid Orb L	24
Temporal Mid L	26	Amygdala L	22	Parietal Sup L	18	Occipital Sup R	24
Angular R	25	Pallidum R	22	Frontal Sup Orb L	18	Occipital Inf L	22
Frontal Sup Orb R	24	Precentral R	21	Frontal Mid Orb R	17	Calcarine L	22
Frontal Mid Orb R	24	Frontal Mid R	21	Parietal Sup R	17	Hippocampus L	22
Occipital Inf L	24	Occipital Mid R	21	Frontal Mid Orb R	17	ParaHippocampal L	21
Parietal Inf R	24	Frontal Mid Orb L	21	Amygdala L	17	Temporal Mid L	21
Frontal Sup Orb L	23	Caudate L	21	Pallidum R	17	Heschl L	20
Frontal Inf Tri R	23	Heschl R	21	Frontal Mid R	17	Caudate L	19
Rectus L	23	Insula L	21	Frontal Mid Orb L	17	Thalamus L	19
Postcentral L	23	Putamen L	21	Calcarine R	17	Precuneus R	19
Parietal Sup R	23	Temporal Pole Mid L	20	Temporal Pole Sup R	17	Olfactory L	18
Frontal Sup R	22	Occipital Inf R	20	Occipital Inf R	16	Frontal Sup L	18
Frontal Inf Orb L	22	Parietal Sup L	20	Cingulum Post R	16	Lingual R	18
Rolandic Oper L	22	Rolandic Oper R	20	Frontal Mid L	16	Temporal Inf L	17
Frontal Sup Medial R	22	Cingulum Post R	20	Frontal Sup R	16	Temporal Pole Mid L	17
Occipital Inf R	22	Calcarine R	20	ParaHippocampal L	16	Pallidum L	17
Frontal Inf Oper R	21	Caudate R	20	Angular R	16	Angular L	17

Maximum Likelihood		Covariance Lasso		Adaptive Covariance Lasso		Bayesian Covariance Lasso	
Occipital Sup L	21	Temporal Pole Sup R	19	Occipital Inf L	16	SupraMarginal L	17
SupraMarginal L	21	Frontal Sup Orb L	19	Frontal Mid Orb L	16	Frontal Mid R	17
Temporal Sup R	21	Cingulum Ant R	19	Olfactory R	15	Calcarine R	16
Frontal Mid R	20	Cingulum Mid R	19	Cingulum Mid L	15	Thalamus R	16
Supp Motor Area L	20	Putamen R	19	Putamen L	14	Fusiform L	16
Supp Motor Area R	20	Thalamus L	19	Caudate R	14	Frontal Inf Orb L	16
Parietal Inf L	20	Frontal Mid L	18	Frontal Sup L	14	Frontal Inf Orb R	16
Angular L	20	Frontal Sup L	18	Supp Motor Area R	14	Temporal Pole Sup L	16
Precuneus L	20	Frontal Sup R	18	Hippocampus R	14	Parietal Sup R	16
Frontal Inf Oper L	19	Supp Motor Area R	18	Amygdala R	14	Olfactory R	15
Frontal Inf Tri L	19	ParaHippocampal L	18	Fusiform L	14	Paracentral Lobule R	15
Cingulum Post L	19	Frontal Sup Medial L	18	Precuneus L	14	Insula R	15
Calcarine L	19	Lingual L	18	Caudate L	13	Temporal Sup R	15
Occipital Mid R	19	Hippocampus R	18	Heschl R	13	Cuneus L	15
Fusiform R	19	Amygdala R	18	Lingual L	13	Occipital Mid R	15
Parietal Sup L	19	Thalamus R	18	Precentral L	13	Frontal Mid L	14
SupraMarginal R	19	Precentral L	17	Parietal Inf R	13	Temporal Pole Mid R	14
Temporal Pole Sup L	19	Angular R	17	Rolandic Oper L	13	Occipital Mid L	14
ParaHippocampal L	18	Parietal Inf R	17	Temporal Pole Mid R	13	Postcentral R	14
Precuneus R	18	Rolandic Oper L	17	Calcarine L	13	Parietal Sup L	14
Temporal Sup L	18	Parietal Inf L	17	Insula R	13	Temporal Pole Sup R	13
Frontal Mid Orb L	17	Cingulum Ant L	17	Temporal Sup R	13	Cingulum Ant L	13
Rolandic OperR	17	Cuneus R	17	Cingulum Ant R	12	Frontal Inf Oper L	13
ParaHippocampal R	17	Hippocampus L	17	Cingulum Mid R	12	Temporal Inf R	13
Lingual R	17	Olfactory R	17	Putamen R	12	Cingulum Post R	12
Occipital Mid L	17	Cingulum Mid L	17	Thalamus R	12	Amygdala R	12
Temporal Mid R	17	Occipital Sup R	16	Cuneus R	12	Precentral L	12
Frontal Mid Orb R	16	Fusiform L	16	Frontal Inf Oper L	12	Cuneus R	12
Rectus R	16	Occipital Inf L	16	Fusiform R	12	Parietal Inf R	12
Cingulum Ant R	16	Precuneus L	16	ParaHippocampal R	12	Fusiform R	12
Cingulum Mid R	16	Frontal Inf Oper L	16	Frontal Inf Orb L	12	SupraMarginal R	12