BAYESIAN PENALIZED METHODS FOR HIGH-DIMENSIONAL DATA

Zakaria S Khondker

A dissertation submitted to the faculty of the University of North Carolina at Chapel Hill in partial fulfillment of the requirements for the degree of Doctor of Philosophy in the Department of Biostatistics.

Chapel Hill
2013

Approved by:

Dr. Joseph G. Ibrahim

Dr. Hongtu Zhu

Dr. Wei Sun

Dr. Donglin Zeng

Dr. Kelly S. Giovanello

Dr. Rebecca Santelli

ABSTRACT

ZAKARIA S KHONDKER: BAYESIAN PENALIZED METHODS FOR HIGH-DIMENSIONAL DATA
(Under the direction of Dr. Joseph G. Ibrahim and Dr. Hongtu Zhu)

Big data presents the overwhelming challenge of estimating a large number of parameters, which is much larger than the sample size. Even for a simple linear model, when the number of predictors is larger than or close to the sample size, such model may be unidentifiable and the least squares estimates of regression coefficients can be unstable. To deal with such issue, we systematically investigate three Bayesian regularization methods with applications in imaging genetics. First, we develop a Bayesian lasso estimator for the covariance matrix and propose a metropolis-based sampling scheme. This development is motivated by functional network exploration for the entire brain from magnetic resonance imaging (MRI) data. Second, we propose a Bayesian generalized low rank regression model (GLRR) for the mean parameter estimation and combine this with factor loading method of covariance estimation to capture the spatial correlation among the responses and jointly estimate the mean and covariance parameters. This development is motivated by performing genome-wide searches for associations between genetic variants and brain imaging phenotypes from data collected by Alzheimer's Disease Neuroimaging Initiative (ADNI). Third, we extend GLRR to longitudinal setting and propose a Bayesian longitudinal low rank regression (L2R2) to account for spatiotemporal correlation among the responses as well as estimation of full-rank coefficient matrix for standard prognostic factors. This development is motivated by genome-wide searches for associations between genetic variants and brain imaging phenotypes observed over time with a primary focus on role of aging and the interaction of age with genotype in affecting brain volume.

To my family who made it possible. Specially, to my wife Eema for so many reasons that I cannot list here. For five years, since I quit graduate school, Eema kept pushing me and finally succeeded in getting me back again. She made drastic lifestyle compromises and cutbacks, to live under the poverty line, with two children, on my meager income of a graduate research assistant. She dealt with my frustrations, stemming from having to unnecessarily abandon projects after significant progress and having to unnecessary rework and delays. To my children Bornali and Borneil, who had to compromise their playtime with their father and were understanding of my time constraints.

ACKNOWLEDGMENTS

First, I would like to acknowledge my adviser Dr. Joseph G. Ibrahim for his guidance and support as well as his pushes to keep the ball rolling. I would like to thank my coadvisor Dr. Hongtu Zhu and committee members - Drs. Wei Sun, Donglin Zeng, Kelly Giovanello and Rebecca Santelli for their valuable time. I would also like to thank the departments admissions committee, the faculty, staff, and fellow students who motivated and supported me along the way.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER 1

SHRINKAGE ESTIMATION IN THE LITERATURE

The emergence of high-dimensional data has posed tremendous challenges to the traditional approches to modeling and estimation, in some cases, rendering traditional modeling approaches obsolete. As a remedy the idea of penalized mehods are gaining popularity among the statistical community. Here we discuss the literature of early and latest approches approaches in univariate and multivariate context. Our primary focus is on multivariate approaches since it it most relevant to our problem.

1.1 Shrinkage of Mean Parameters for Univariate Response

The early approaches to deal with high-dimensional problems involved separation approach− variable selection to reduce dimension and then parameter estimation. While subset selection techniques were used in variable selection, estimation was typically done by least squares regression. However, the subset selection approach is unstable due to discontinuity of the process; so is the best single-model variable selection (Breiman, 1996). Bayesian model averaging provides a robust prediction remedy regarding stability; under squared error loss optimal prediction takes the form of Bayes model averaging. Brown et al. (2002) introduced Bayes model averaging incorporating variable selection allowing for fast computation for dimensions up to several hundreds.

The later approaches adopted shrinkage− shrinking the parameters to achieve stability and improve performance. The most popular among them are the L_1 and L_2 priors and their variants. The L_2 priors (penalties) tend to shrink the regression coefficients to achieve stability; it forces the coefficients of highly correlated covariates towards each other by inflating the diagonals of the $X^{\mathrm{T}}X$, where X is the matrix of covariates. The most common example of L_2 priors are normal (Ridge regression) and Cauchy priors. Recently there has been a surge in *black hole* priors− priors that create a singularity at the origin with a black hole around. The prior forces the maximum aposteriori (MAP) estimates of the smaller coefficients to singularity without creating discontinuity to perform simultaneous variable selection and estimation. The most common of them are lasso (Tibshirani, 1996), adaptive lasso (Zou, 2006), smoothly clipped absolute deviation (Fan and Li, 2001), and double Pareto (Armagan et al., 2011). Some hybrids and other variants of these priors include grouped lasso, fused lasso, elastic net, etc. Rothman et al. (2010) proposed simultaneous estimation of sparse coefficient matrix and sparse covariance matrix to improve on estimation error under L_1 penalty. Their approach does not take advantage of the potential correlation among the coefficients.

The penalized methods estimates the coefficients by minimizing the residual sum of squares (RSS) with a constraint, that is, minimizing $||Y - X\beta||^2 + g(\beta)$, where $g(\beta)$ is some penalty function. A popular general choice is $g(\beta) = \lambda \sum_{j=1}^{p} |\beta_j|^\alpha$, $\alpha = 2$ leads to Ridge regression and $\alpha = 1$ leads to lasso regression. Ridge regression typically achieves better prediction performance compared to ordinary least squares (OLS); however, model interpretation is difficult and it never estimates coefficients as exactly zero since the prior is continuous. The lasso, on the other hand, reaches a sparser solution by estimating some coefficients as zero due to discontinuity of the implied prior at zero. Adaptive lasso and SCAD also achieves sparse solutions as lasso.

Bayesian shrinkage regression methods achieve regularization through shrinkage induced by priors. Summarized in tabel 1.1 are the commonly used scaled mixture of normal priors which lead to heavy-tailed priors with a peak around the origin (Carvalho and Scott, 2009a; Armagan et al., 2011; Park and Casella, 2008; Kyung et al., 2010).

1.2 Shrinkage of Covariance Parameters

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 high-dimensional data (when $n < d$) derived from 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, gene arrays where few dozen or hundred samples are arrayed each array containing several hundred to several thousand genes (Davidson and Levin, 2005), spectroscopy, climate studies and many other applications are just a few examples. In this case, S has a maximum rank of n which is smaller than its dimension d, and therefore S is singular.

1.2.1 Frequentist Methods

In the frequentist framework, significant work has been done on model selection and precision (covariance) matrix estimation in Gaussian models (Banerjee et al., 2007;

Table 1.1: Bayesian priors for mean parameter estimation

Model	Prior on β	hyper-prior	hyper-hyper-prior
NJ	$\beta_j \sim N(0, \sigma^2\psi_j)$	$\pi(\psi_j) \propto \frac{1}{\psi_j}$	
NE	$\beta_j \sim N(0, \sigma^2\psi_j)$	$\psi_j \sim \text{Exp}(\frac{\lambda}{2}), \lambda > 0$	
NIGam	$\beta_j\psi_j \sim N(0, \sigma^2\psi_j)$	$\psi_j \sim \text{IGgam}(a_0, a_0 b_0^2)$	
NG	$\beta_j\psi_j \sim N(0, \sigma^2\psi_j)$	$\psi_j \sim \text{Gamma}(a_0, b_0^2)$	
NEG	$\beta_j\psi_j \sim N(0, \sigma^2\psi_j)$	$\psi_j \sim \text{Exp}(\frac{\lambda_j}{2})$	$\lambda_j \sim \text{Gamma}(a_0, b_0)$
LASSO	$\beta_j\psi_j \sim N(0, \sigma^2\psi_j)$	$\psi_j \sim \text{Exp}(\frac{\lambda^2}{2})$	$\lambda^2 \sim \text{Gamma}(a_0, b_0)$
HS	$\beta_j \sim N(0, \psi_j)$	$\psi_j \sim C^+(0, \tau)$	$\tau \sim C^+(0, \sigma)$
gDP	$\beta_j \sim N(0, \psi_j)$	$\psi_j \sim \text{Exp}(\frac{\lambda^2}{2})$	$\lambda \sim \text{Gamma}(a_0, b_0)$

NJ = Normal Jeffry's, NE = Normal Exponential, NIGam = Normal Inverse Gamma, NG = Normal Gamma, NEG = Normal Exponential Gamma, HS = Horseshow, gDP = Genaralized Double Pareto.
$C^+(0, a)$ is a standard half-Cauchy distribution on the positive reals with scale parameter a.

Friedman et al., 2008a; Fan et al., 2009; Drton and Perlman, 2004). The original paper by Dempster (1972) 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 (2000) 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 (2004) proposed a conservative simultaneous confidence interval to select a model in a single step as an improvement.

Banerjee et al. (2007) proposed block coordinate descent algorithm which can be interpreted as recursive l_1-norm penalized regression. Suppose $y \sim N(\mu, \Sigma)$, $S = \sum_{i=1}^{n}(y_i - \mu)(y_i - \mu)^{\mathrm{T}}$ and $\Omega = \Sigma^{-1}$ then the estimate takes the form

$$\hat{\Omega} = \arg\max_{\Omega \succ 0} \log\det\Omega - tr(S\Omega) - \lambda||\Omega||_1; \tag{1.1}$$

where $\succ$ stands for positive definite, $||\Omega||_1$ denotes the sum of the absolute values of the elements of the positive definite matrix Ω, and λ is the penalty scalar (proxy for the number of nonzero elements in the matrix). When $S \succ 0$, MLE of Σ can be obtained by setting $\lambda = 0$, however Σ is not invertible for $n < p$.

Let $W = \hat{\Sigma}$ be an estimate of Σ, the dual of their sparse maximum likelihood problem is

$$\hat{\Sigma}^{-1} = \max\{\log\det W : ||W - S||_\infty \leq \lambda\}. \tag{1.2}$$

They choose the penalty parameter as a function of α, the probability of zero element of Σ falsely estimated as non-zero. Their plan is to optimize over one rwo and column of the variable matrix W at a time and repeatedly sweep through all columns until

convergence. In other words, partition W and S as:

$$W = \begin{pmatrix} W_{-kk} & w_k \\ w_k^{\mathrm{T}} & w_{kk} \end{pmatrix} \text{ and } S = \begin{pmatrix} S_{-kk} & s_k \\ s_k^{\mathrm{T}} & s_{kk} \end{pmatrix}, \tag{1.3}$$

where θ_{kk} is the kth diagonal element of Θ, $\theta_k = (\theta_{k1}, \ldots, \theta_{k,k-1}, \theta_{k,k+1}, \ldots, \theta_{kd})^{\mathrm{T}}$ is the vector of all off-diagonal elements of the kth column, and Θ_{-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 Θ. Then the algorithm proceeds as follows:

1. Initialize $W^{(0)} = S + \lambda I$, for $j = 1 \ldots p$ let $W^{(j-1)}$ denote the current iterate. Solve the quadratic program
   $\hat{y} = \arg\min_y \{y^{\mathrm{T}} (W_{-(jj)}^{(j-1)})^{(-1)} y : ||y - s_j||_\infty \leq \lambda\}$.
2. Update rule: $W^{(j)}$ is $W^{(j-1)}$ with w_j replaced by $\hat{y}$.
3. Let $\hat{W}^{(0)} = W^{(p)}$.
4. Check convergence by $tr\{(\hat{W}^{(0)})^{-1} S\} - p + \lambda ||(\hat{W}^{(0)})^{-1}||_1 \leq \epsilon$, where ϵ is the convergence criterion.

Friedman et al. (2008b) used similar partitioning as Banerjee et al. (2007) and showed that minimizing (1.1) is equivalent to minimizing: $\min_\theta ||W_{11}^{1/2}\theta - \frac{1}{2} W_{11}^{-(1/2)} s_{12}||^2 + \rho||\theta||_1$ Where $\hat{\theta}$ is the solution to the above lasso problem (3). That is they use lasso to estimate

$$\hat{\theta} = \arg\min ||W_{11}^{1/2}\theta - \frac{1}{2} W_{11}^{-(1/2)} s_{12}||^2 + \rho||\theta||_1 \tag{1.4}$$

Their covariance LASSO algorithm has the following 3 steps:

1. Start with $W = S + \rho I$. The diagonal of W remains unchanged in what follows.
2. For each $j = 1, 2, \ldots, p$, solve the lasso problem (1), which takes as input the inner

products W_{11} and s_{12}. This gives a $p-1$ vector solution $\hat{\theta}$. Fill in the corresponding row and column of W using $w = 2W_{11}\hat{\theta}$.

3. Continue until convergence.

Fan et al. (2009) solved the following equation for sparse matrix under the penalized likelihood framework.

$$\max_{\Omega \in S_p} \log\det\Omega - tr(\hat{\Sigma}\Omega) - \sum_{i=1}^{p}\sum_{j=1}^{p} p_{\lambda_{ij}}(|\omega_{ij}|), \tag{1.5}$$

where ω_{ij} is the $(i,j)th$ element of Ω, λ_{ij} is the tuning parameter and $p(.)$ is the generic penalty function on each element.

Schäfer and Strimmer (2005) minimized the MSE compromising between bias and variance. If $\hat{\Theta}$ is the unrestricted estimate and $\tilde{\Theta}$ the restricted estimate from a reduced model then the optimal estimate is $\hat{\Theta}^* = \lambda\tilde{\Theta} + (1-\lambda)\hat{\Theta}$ for a suitable shrinkage intensity $\lambda \in [0,1]$. The value of λ is determined by minimizing the risk $R(\lambda) = E(L(\lambda)) = E(\sum_{i=1}^{p}(\hat{\theta}_i^* - \theta_i)^2$. They minimized this analytically to obtain the optimal value λ^* as a function of variances of $\hat{\Theta}$ and $\tilde{\Theta}$, their covariances and bias respectively, which is unique and always exist. For practical purpose they replace those variances, covariances and biases by their unbiased sample counterparts to obtain $\hat{\lambda}^*$. For finite samples the value might be negative or exceed unity, in which case they truncate it to zero or one.

1.2.2 Bayesian Non-Graph Theory Methods

Bayesian covariance estimation followed two major paths. The non-graph theory methods disregard the underlying graphical structures and perform shrinkage via priors on the elements, eigenvalues, and decompositions of the matrix. The graph theory methods rely assuming particular graphical structure and hyper-inverse Wishart priors conditional on the graph. Among Bayesian shrinkage methods, Yang and Berger (2007)

used reference priors on the eigenvalues of the covariance matrix to regularize the eigen structure.

Smith and Kohn (2002) decomposed $\Sigma^{-1} = \Omega = BDB^{\mathrm{T}}$ where B is a lower triangular matrix with 1's on the diagonal and D is a diagonal matrix. They introduced an indicator matrix γ where $b_{ij} = 0$ iff $\gamma_{ij} = 0$ $b_{ij} \neq 0$ iff $\gamma_{ij} = 1$, which ensures that the lower triangular elements of B can be 0 with positive probability. For a given γ, some often the lower-triangular elements of B will be zero. For the unconstrained elements of B, denoted B_γ, they used fractional prior as $p(B_\gamma|\gamma, D) \propto p(e|B, D, \gamma)^{1/n}$. The elements of γ are taken independent a priori, with $p(\gamma_{ij} = 1|\omega) = \omega$, which implies that there will $p(p-1)\omega/2$ nonzero elements in B. They assumed uniform [0,1] prior for ω.

Frühwirth-Schnatter and Tüchler (2008) used Cholesky decomposition on hierarchical linear mixed models to identify zeros on the covariance matrix $\Sigma = CC^{\mathrm{T}}$, where C is a lower triangular matrix (Smith and Kohn (2002) decomposed Σ^{-1}). This approach allows to shrink random effects twards fixed ones. They also discussed how the ordering of the data can change the zero pattern; although the rank of Σ, rank of C and the number of 0 columns are unaffected by ordering. Noncentral parameterization along with Cholesky decomposition reduces the problem of variance-covariance selection to the more common problem of Bayesian variable selection in multiple regression. Prior for γ, the indicator matrix that determines which elements of C are zeros, is selected such that $P(\gamma_{ij} = 1|\tau) = \tau$, where the hyperparameter $\tau \sim U[0, 1]$. So the number of non-sero element in C follows binomial distribution $B(\frac{p(p+1)}{2}, \tau)$. For the Cholesky factor C conditionally fractional prior was chosen that depends on random effects. They developed the MCMC scheme for simulation.

Barnard et al. (2000) used *separation strategy* as $\Sigma = \text{diag}(S) \quad R \quad \text{diag}(S)$ where S is vector of the standard deviations and R is the correlation matrix. There is a practical motivation for this separation since most practitioners think in terms of standard deviations and correlations. They assume $S \sim N(\xi, \Lambda)$, alternative would be to choose independent scaled inverted chi-squared distributions for each of the variances. They assume R independent of S, $\{r_{ij}, i \neq j\}$ are *a priori* exchangeable, and priors are diffuse to reflect week prior knowledge about R. They explored two extreme cases- (1) marginally uniform, which can be obtained from the commonly used inverse-Wishart distribution for Σ, and (2) jointly uniform prior for r_{ij}.

Wong et al. (2003) decomposed $\Sigma^{-1} = \Omega = TCT^{\text{T}}$ where T is a diagonal matrix such that T_i is the inverse of the partial standard deviation of y_{it} and C is a correlation matrix with $C_{ii} = 1$ and $C_{ij} = -\rho_{ij}$ the partial correlation coefficients. They put noninformative gamma priors on $\{T_i, i = 1, ..., p\}$ assuming T_i i.i.d. and independent of the elements of C, $p(T_i) \propto p(\Omega_{ii}) \frac{d\Omega_{ii}}{dT_i} \propto T_i^{2\alpha-1} e^{-\beta T_i^2}$. Their sampling scheme used MCMC based on the following Metropolis-Hastings algorithm: $q(T_i|Y, T_{(-i)}, C)$ and $q(dC_{ij}|Y, T, C_{(-ii)})$. The T_i, are generated one at a time using a Gaussian proposal. The C_{ij} are generated one at a time using a Metropolis-Hastings proposal that allows C_{ij} to be identically zero, and that uses a Gaussian proposal for the continuous part of the conditional density.

Chen and Dunson (2003) used modified Cholesky decomposition $\Sigma = LL^{\text{T}} = \Lambda\Gamma\Gamma^{\text{T}}\Lambda$ where $\Gamma = diag(\gamma_1, ..., \gamma_p)$ then the random effects model becomes $y_i = X_i\alpha + Z_iLb_i + \epsilon_i$. In their first paper they applied the method to linear mixed model and in the second paper they applied to logistic regression.

Huang et al. (2006) proposed nonparametric method for identifying parsimony in estimating covariance matrix using modified Cholesky decomposition. If $cov(y) = \Sigma$ and

$\epsilon = Ty$ with $cov(\epsilon) = diag(\sigma_1^2, ..., \sigma_n^2) = D$ then $\Sigma^{-1} = T^{\mathrm{T}} D^{-1} T$ and the likelihood becomes $-2l(\Sigma, y) = \sum_{t=1}^{n} log(\sigma_t^2) + \sum_{t=1}^{n} \frac{\epsilon_t}{\sigma_t^2}$. Thus the modified Cholesky decomposition provides a parameterization of the covariance matrix with unconstrained parameters and tramsfers the difficult task of modeling a covariance matrix to that of variable selection in the sequence of regression $y_t = \sum_{j=1}^{t-1} \phi_{tj} y_j + \epsilon_t$.

Their penalized likelihood estimator was derived as the Bayes posterior mode under independent diffuse priors. The algorithm amounts to applying a similar regression algorithm repeatedly to the rows of the Cholesky factor T. The authors claimed this method to be better than smoothing when T is sparse instead of being smooth.

1.2.3 Bayesian Graph Theory Methods

Bayesian graph theory methods exploit decomposability and use hyper-inverse Wishart priors to sample from the marginal.

Dawid and Lauritzen (1993) introduced the notion of a probability distribution on a multivariate space called hyper Markov law and concentrated on the set of Markov probabilities over some decomposable graph. They discussed sampling distributions of maximum likelihood estimators in decomposable graphical models, and showed that hyper Markov laws form natural conjugate prior distributions for a Bayesian analysis of these models. They also constructed a range of specific hyper Markov priors, including the hyper multinomial, hyper Dirichlet, hyper Wishart, and hyper-inverse Wishart laws. Their work has led many to exploit the hyper-inverse Wishart (HIW) priors for Gaussian graphical model.

Guidici and Green (1999) used HIW priors on the precision matrix conditional on decomposable graphs for Bayesian model determination in Gaussian graphical models. They introduced hierarchical Bayesian Gaussian graphical models and designed reversible jump Markov chain Monte Carlo (MCMC) algorithm for structural and quantitative learning using local computations.

Let $G = (V, E)$ be an undirected *graph* with *vertex* set V of v elements and $edge-set$ E. If there is an edge $(a, b) \in E$ then the v*ertices* a and b are called *neighbors* in G and if all vertices are connected the graph is *complete*. A complete subgraph which is not contained in another complete subgraph is a *clique*. Subgraphs (A, B, C) of G forms a decomposition of G if any path from A to B goes through C. In other words if $V = A \cup B$, $C = A \cap B$ is complete, and the all paths from A and B are through C only then (A, B, C) form a decomposition and C is said to be separator. A sequence of subgraphs that cannot be further decomposed are the *prime components* of a graph and a graph is decomposible only if all its prime components are complete.

Carvalho and Scott (2009b) developed Wishart g-prior, a default version of the hyper-inverse Wishart prior for restricted covariance matrices and showed how it corresponds to the implied fractional prior for selecting a graph using fractional Bayes factors. Then they applied a class of priors that automatically handles the problem of multiple hypothesis testing. They demonstrated that the combined use of a multiplicity-correction prior on graphs and fractional Bayes factors for computing marginal likelihoods yields better performance.

How well these graphical methods do when there is no prior knowledge of the underlying graph structure is not studied yet. Furthermore, these methods don't work for any type of graph. Existing non-graph theory Bayesian methods 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 those methods are based on sampling the elements of the matrix one at a time which is not efficient and not attractive for high-dimensional data, especially when d is large. 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. Graphs theory methods, however, does not work for all graphical structures limiting their use. We will focus on non-graph theory approach.

1.3 Multivariate Response Regression Model

The model for multivariate regression is

$$Y = XB + \epsilon, \tag{1.6}$$

where Y is the $n \times d$ matrix of responses, X is the $n \times p$ matrix of predictors, B is the $p \times d$ matrix of regression coefficients, and ϵ is the $n \times d$ matrix of random errors. Alternatively, we can write

$$y_{ik} = \sum_{j=1}^{p} x_{ij}\beta_{jk} + \epsilon_{ik},$$

where i is the subject index ($i = 1, \ldots n$), j is the predictor index ($j = 1, \ldots p$), and k is the response index ($k = 1, \ldots d$). Error terms ϵ_{ik} and $\epsilon_{ik'}, (k \neq k')$ represent the different responses within a subject (e.g., fMRI signals from regions k and k' of subject i) and are likely to be correlated, while e_{ik} and $e_{i'k}, (i \neq i')$ represent the same response from different subjects (e.g., fMRI signals from regions k of subjects i and i') and are assumed to be independent.

The emergence of high-dimensional data in genomics, imaging, econometrics, chemometrics and other quantitative area has presented us with a large number of predictors along with a large number of response variables that calls for simultaneous variable selection and estimation of both the mean and covariance parameters. Traditionally, subset selection was used for variable selection and least squares was used for estimation when presented with large number of predictors. However, the subset selection approach is unstable due to discontinuity of the process. When either the dimension d of the covariance matrix or the number of predictors p is larger than the sample size the model is not identifiable, leading to failure of the traditional methods like least squares or maximum likelihood. For $p > n$ the subset selection method can be unstable because the procedure is not continuous (Breiman, 1996). Even when the sample size is larger than both the dimension of the covariance matrix and the number of predictors traditional methods are stable only when both $\frac{d}{n}$ and $\frac{p}{n}$ are reasonably small. When $\frac{d}{n} < 1$, but not small enough, the condition numbers of the maximum likelihood estimator S of the covariance matrix can be unusually large leading to unstable estimators for the precision matrix (Khondker et al., 2011). When $\frac{p}{n} < 1$, but not small enough, the condition numbers of the matrix $X^{\mathrm{T}}X$, where X is the covariate matrix, can be unusually large leading to unstable least squares estimators for the mean parameters.

Best single-model variable selection is inherently unstable and Bayesian model averaging provides a robust prediction remedy. Under squared error loss optimal prediction takes the form of Bayes model averaging (see Brown et al. (2002) and references therein). The shrinkage approaches for estimation of B can be divided into two major groups - (1) without decomposition and (2) via decomposiotn.

1.3.1 Traditional Estimation of Regression Coefficients

The curse of dimensionality boils down to dealing with too many parameters than the sample size reasonably permits. When dimension is larger than the sample size the model is unidentifiable and all the parameters are not estimable. Even when the dimension is smaller than the sample size but dimension to sample size ratio is not small enough or there is colinearity among the predictors the estimators are unstable. The early approaches involved separation approach− variable selection to reduce dimension and then parameter estimation. While subset selection techniques were used in variable selection, estimation was typically done by least squares regression. However, the subset selection approach is unstable due to discontinuity of the process; so is the best single-model variable selection (Breiman, 1996). Bayesian model averaging provides a robust prediction remedy regarding stability; under squared error loss optimal prediction takes the form of Bayes model averaging. Brown et al. (2002) introduced Bayes model averaging incorporating variable selection allowing for fast computation for dimensions up to several hundreds.

Another approach is shrinkage− shrinking the parameters to achieve stability and improve performance. The most popular among them are the L_1 and L_2 priors and their variants. The L_2 priors (penalties) tend to shrink the regression coefficients to achieve stability; it forces the coefficients of highly correlated covariates towards each other by inflating the diagonals of the $X^{\mathrm{T}}X$, where X is the matrix of covariates. The most common example of L_2 priors are normal (Ridge regression) and Cauchy priors. Recently there has been a surge in *black hole* priors− priors that create a singularity at the origin with a black hole around. The prior forces the maximum aposteriori (MAP) estimates of the smaller coefficients to singularity without creating discontinuity to perform simultaneous variable selection and estimation. The most common of them

are lasso (Tibshirani, 1996), adaptive lasso (Zou, 2006), smoothly clipped absolute deviation (Fan and Li, 2001), and double Pareto (Armagan et al., 2011). Some hybrids and other variants of these priors include grouped lasso, fused lasso, elastic net, etc. Rothman et al. (2010) proposed simultaneous estimation of sparse coefficient matrix and sparse covariance matrix to improve on estimation error under L_1 penalty. Their approach does not take advantage of the potential correlation among the coefficients.

Breiman and Friedman (1997) considered the problem of predicting several response variables from the same set of explanatory variables and showed that even when the random error terms e_{ik} and $e_{ik'}$ are independent for different responses the responses y_{ik} and $y_{ik'}$ of a sample i can be correlated due to their dependence on the same predictor set X_i. They introduced shrinkage estimation called "Cards and Whey" that predicts the multivariate response with an optimal linear combination of the ordinary least squares predictors method to take advantage of the correlation in the responses arising from shared random predictors as well as correlated errors. This is a multivariate generalization of proportional shrinkage based on cross-validation and derives its power by shrinking in the right co-ordinate system (canonical co-ordinates).

Rothman et al. (2010) proposed simultaneous estimation of sparse mean parameters and the covariance matrix called multivariate regression with covariance estimation (MRCE). They improve prediction in the multivariate regression problem while allowing for interpretable models in terms of the predictors. They reduced the number of parameters using the $L-1$ penalties on both the mean parameter B and covariance parameter Ω in optimizing the likelihood. MRCE assumes the predictors are not random and focused on the conditional distribution of Y given X, althout, the formulas would be the same with random predictors. Unlike in the Curds and Whey framework, the MRCE assumes that correlation of the response variables arises only from the correlation in the

errors.

1.3.2 Low Rank Estimation

Principal component analysis (PCA) is arguably the most widely used statistical tool for data analysis and dimensionality reduction for multivariate response. A number of natural approaches to robustifying PCA have been explored and proposed in the literature over several decades including influence function techniques, multivariate trimming, alternating minimization, random sampling techniques, etc. (Candés et al., 2009; Jolliffe, 2002). A convenient approach is via decompose the data matrix into a diagonal matrix of singular values Δ and two unitary matrices U and V, that is

$$Y = U\Delta V = \sum_{l=1}^{r} \delta_l u_l v_l^{\mathrm{T}}. \tag{1.7}$$

A common convention is to list the singular values in descending order. The rank is reduced by minimizing the dimension r of Δ.

Candés et al. (2009) applied robust principal component analysis when response is a superposition of a low-rank component and a sparse component ($Y = Y_0 + S_0$) to recover the two components individually. Their method Principal Component Pursuit is used to recover the low rank component $Y_0 = U\Delta V^{\mathrm{T}} = \sum_{l=1}^{r} \delta_l u_l v_l^{\mathrm{T}}$.

There is another less explored approach that can exploit commonality of the coefficients to achieve shrinkage in the presence of colinearity among regressors as well as among responses. The emergence of high-dimensional data in genomics, imaging, econometrics, chemometrics and other quantitative area has presented us with a large number of predictors along with a large number of response variables, often with strong

correlations. In quantitative trait studies problems arise when high-dimensional response sets such as fMRI signals or volumes of each voxel in the brain are predicted by high-dimensional covariate sets such as gene expression or SNPs. Genes or SNPs may co-conspire, working in unison, to produce similar patterns of fMRI signals in the brain. In such situations regression coefficients are both vertically and horizontally correlated with rank smaller than dimension. For example, Alzheimer's Disease Neuroimaging Initiative (ADNI) collects clinical, imaging, genetic, and biospecimen data on elderly controls, mildly cognitive impaired, and Alzheimer's patients. To study the impact of SNPs on the volumes of certain regions of interest one has to expect some common pattern of correlation among the regression coefficients. The responses and predictors may be associated through fewer channels than the dimensions of the coefficient matrix leading to a reduced rank of the mean parameter B.

The relationship among correlated responses and predictors may be exploited by dimension reduction via reduced rank decomposition of the regression parameters to greatly reduce the number of parameters and facilitate efficient estimation of the coefficient matrix. This factorization has started to receive more attention in recent years. Several authors have explored the decomposition of the response matrix Y (see Ding et al. (2011) and the references therein). Others took the latent model approach to restrict the rank of the coefficient matrix (Izenman (1975), Reinsel and Velu (1998)) and sparsity-inducing regularization techniques to reduce the number of parameters (Tibshirani (1996), Turlach et al. (2005), Peng et al. (2010)). Chen et al. (2012) has used singular value decomposition of the coefficient matrix B with L_1 penalty on the singular vectors U and V and computed the posterior modes for orthogonal design matrix. There method, however, is limited to orthogonal design matrix where columns of X must be independent. We relax the assumption to remove the orthonormality of U and V and allow correlated covariates as the regression coefficients are likely to be correlated both

ways. The coefficients in each row can be correlated as they are the effect of the same covariate on different responses for a particular subject. Moreover, the coefficients in each column can be correlated as they are the effect of different covariates on the same response for a particular subject. Our approach exploits this two-way correlation structure in cases where regressors can be correlated and random.

Breiman and Friedman (1997) introduced shrinkage estimation called Cards and Whey (C&W) method to improve on prediction error when the same set of predictors is used for multivariate response. They showed that even when the random error terms e_{ik} and $e_{ik'}$ are independent for different responses the responses y_{ik} and $y_{ik'}$ of a sample i can be correlated due to their dependence on the same set of predictors X_i. The C&W linear predictor has the form $\tilde{Y} = \hat{Y}^{OLS}M$, where M is a $d \times d$ shrinkage matrix estimated from the data to exploit correlation in the responses arising from shared random predictors.

Ding et al. (2011) extended the idea in Bayesian framework and introduced a rank recovery mechanism; their low rank component is modeled as $Y_0 = U(Z\Delta)V^{\mathrm{T}} = \sum_{l=1}^{r} \delta_l z_l u_l v_l^{\mathrm{T}}$, where Z is diagonals matrix with $z_l \in \{0, 1\}$. They claimed that restrictions on U, V, and Δ comes at a greater computational cost without any remarkable benefit. Relaxing the orthonormality assumptions of U and V and non-negativity assumption on Δ, that allows for a more flexible prior specification, they used normal priors for u_l, v_l and δ_l to achieve shrinkage. A binomial prior is used for z_l, which is introduced for rank learning.

Chen et al. (2012) used singular value decomposition of the mean parameter B for multivariate response model (1.7) under $L - 1$ (adaptive lasso) penalty. The SVD representation shows that B is composed of r orthogonal unit-rank layers of decreasing

importance, and each layer provides a distinct channel relating the responses to the predictors, which parsimoniously reveals the structure imposed by (1.7). They sought to seek a $\hat{B}$ with sparse SVD structure in the vicinity of some initial consistent estimator $\tilde{B}$ by decomposing the rank-r problem into r parallel sparse unit-rank regression problems, by forming r "exclusive layers".

None of the existing approches address the simultaneous estimation of high- dimensional mean parmeter matrix and high- dimensional covariance matrix.

1.3.3 Low Rank Estimation Under Longitudinal Setting

Many longitudinal biomedical studies, such as genomics and neuroimaging, repeatedly collect a large number of responses and covariates from a small set of subjects and focus on establishing associations among them. For instance, in imaging genetics, various imaging measures, such as volumes of regions of interest (ROIs), are repeatedly measured and may be predicted by high-dimensional covariate vectors, such as single nucleotide polymorphisms (SNPs) or gene expressions. These imaging measures can serve as important endotraits that may ultimately lead to discoveries of genes for some complex mental and neurological disorders, such as schizophrenia, since imaging data provides the most effective measures of brain structure and function (Scharinger et al., 2010; Paus, 2010; Peper et al., 2007; Chiang et al., 2011b,a). This motivates us to develop a longitudinal low rank regression model for the analysis of longitudinal high-dimensional responses and covariates.

Modeling longitudinal high-dimensional covariates and responses involve four challenges (i) a large number of regression coefficients, (ii) spatial correlation, (iii) temporal correlation, and (iv) multicollinearity among predictors. When the dimension of responses and the number of covariates, which are denoted by d and p, respectively, are

even moderately high, fitting a multivariate linear model usually requires estimating a $d \times p$ matrix of regression coefficients, whose number pd can be much larger than the sample size. At each given time, accounting for complicated spatial correlation among multiple responses is important for improving prediction accuracy of multivariate analysis (Breiman and Friedman, 1997). Accounting for temporal correlation is important for both prediction and estimation accuracy. Moreover, the collinearity among genetic predictors can cause issues of over-fitting and model misidentification (Fan and Lv, 2010).

Under the cross-sectional settings, several approaches explored new methods for high-dimensional responses and covariates. Breiman and Friedman (1997) introduced a Cards and Whey (C&W) to improve prediction error by accounting for correlations among the response variables when both p and d are moderate compared to the sample size. Peng et al. (2010) proposed a variant of the elastic net to enforce sparsity in the high-dimensional regression coefficient matrix, but they did not account for correlations among responses. Rothman et al. (2010) proposed a simultaneous estimation of a sparse coefficient matrix and a sparse covariance matrix to improve on estimation error under the L_1 penalty. Vounou et al. (2010) considered the singular value decomposition of the coefficient matrix and used the LASSO-type penalty on both the left and right singular vectors to ensure its sparse structure. They, however, do not model longitudinal data and do not provide a standard inference tool (e.g., standard error) on the nonzero components of the left and right singular vectors or the coefficient matrix.

Several attempts have been made to investigate the effect of genotypes on longitudinal phenotypes. Chen and Wang (2011) proposed penalized spline based methods for functional mixed effects models with varying coefficients, but they focus on small p and d under a low-dimensional setting. Wang et al. (2012) used sparse multitask

regression to examine the association between genetic markers and longitudinal neuroimaging phenotypes. However, their multi-task regression model considered subjects with the same number of repeated measures and ignore spatial-temporal correlations of imaging phenotypes, and thus it leads to loss of statistical power in detecting gene-imaging associations. Vounou et al. (2011) and Silver et al. (2012) proposed various sparse reduced-rank regression models by using penalized regression methods for the detection of genetic associations with longitudinal phenotypes. They, however, ignore the spatio-temporal correlations of longitudinal phenotypes, which are important for both estimation and prediction accuracy. Moreover, none of them explore the gene and time interaction, which can reveal important genetic traits altering time affects on longitudinal phenotypes.

1.4 Motivating Examples

Consider the challenges in the analysis of genetic and imaging data collected by the NIH ADNI. The NIH ADNI is an ongoing public-private initiative to test whether genetic, clinical, functional and structural neuroimaging data can be combined to measure the progression of mild cognitive impairment (MCI) and early Alzheimer's disease (AD) . ADNI initiative is recruiting study subjects over 50 sites across the United States and Canada. The genetic and clinical data along with corresponding structural brain MRI data from baseline and follow-up were obtained from the ADNI publicly available database (http://adni.loni.ucla.edu/). Our interest is to perform genome-wide searches for establishing the association between the SNPs collected on top genes reported by AlzGen (http://www.alzgene.org/) and the brain volumes of 93 regions of interest (ROIs), while accounting for other time-varying covariates, such as age, and baseline covariates, such as gender, as well as spatiotemporal correlation among responses. By using the Bayesian GLRR for repeated measures data, we can easily carry out formal statistical inferences,such as the identification of significant SNPs or SNPs that interact with aging

on the differences among all 93 ROI volumes between AD and normal controls.

The MRI data was collected across a variety of 1.5 Tesla MRI scanners with individualized protocols for each scanner. To obtain standard T1-weighted images volumetric 3-dimensional sagittal MPRAGE or equivalent protocols with varying resolutions were used. The typical protocol included: inversion time (TI) = 1000 ms, repetition time (TR) = 2400 ms, flip angle = 8^o, and field of view (FOV) = 24 cm with a $256 \times 256 \times 170$ acquisition matrix in the $x-$, $y-$, and $z-$dimensions yielding a voxel size of $1.25 \times 1.26 \times 1.2$ mm^3. Standard steps including anterior commissure and posterior commissure correction, skull-stripping, cerebellum removing, intensity inhomogeneity correction, segmentation and registration (Shen and Davatzikos, 2004) were used to preprocessed the MRI data. We then carried out automatic regional labeling by labeling the template and by transferring the labels following the deformable registration of subject images. After labeling 93 ROIs, we were able to compute volumes for each of these ROIs for each subject.

To genotype subjects in the ADNI database, the Human 610-Quad BeadChip (Illumina, Inc., San Diego, CA) was used, which resulted in a set of 620,901 SNP and copy number variation (CNV) markers. Since the Apolipoprotein E (APOE) SNPs, rs429358 and rs7412, are not on the Human 610-Quad Bead-Chip, they were genotyped separately. These two SNPs together define a 3 allele haplotype, namely the $\epsilon 2$, $\epsilon 3$, and $\epsilon 4$ variants and the presence of each of these variants was available in the ADNI database for all the individuals. The software EIGENSRAIT in the package of EIGENSOFT 3.0 was used to calculate the population stratification coefficients of all subjects. To reduce population stratification effects, we only used 749 Caucasians from all 818 subjects who had at least one imaging sample available.

We also performed quality control on this initial set of genotypes. In order to impute the missing genotypes in our sample, we used MACH4 version 1.0.16 with default parameters to infer the haplotype phase. We also included the APOE-ϵ_4 variant, coded as the number of observed ϵ_4 variants. We dropped SNPs with more than 5% missing values and imputed the mode for the missing SNP for the remaining. In the final quality controlled genotype data, we dropped the SNPs with minor allele frequency smaller than 0.1 and Hardy-Weinberg p-value $< 10^{-6}$.

The data is multivariate whose covariance needs to be estimated in order to obtain a more precise estimate for the regression coefficients and build nework among the regions of interest. The covariates are hig-dimensional deserving special techniques for fitting feasible regression models. The responses are measureed repeatedly calling for accomodating speciotemporal correlation as well as age effect on response as well as genotype-phenotype relationship. We developed a series of three papaers to address these issues.

1.5 Methods Background

Our first paper introduces a genaralized double-gamma prior that can be reduced to commonly used frequentist methods. Then we develop a Bayesian lasso estimator for the covariance matrix and propose a metropolis-based sampling scheme. A major hurdle in covariance estimation is the positive-definiteness constrain. Our columnwise sampling scheme allowes sampling positive-definite matrices while opening the floodgate for for many differrent priors. This development is motivated by functional network exploration for the entire brain from magnetic resonance imaging (MRI) data.

Next we propose a Bayesian generalized low rank regression model (GLRR) for the mean parameter estimation where the regression coefficient matrix is separated into

single-rank laers. Then we combine this with factor loading method of covariance estimation to capture the spatial correlation among the responses and jointly estimate the mean and covariance parameters. We explore model evaluation and optimal rank selection that allowes for inference on each layer of the coefficient matrix. This development is motivated by performing genome-wide searches for associations between genetic variants and brain imaging phenotypes from data collected by Alzheimer's Disease Neuroimaging Initiative (ADNI).

Finally, we extend GLRR to longitudinal setting and propose a Bayesian longitudinal generalized low rank regression (LGLRR) to account for spatiotemporal correlation among the responses as well as estimation of full-rank coefficient matrix for standard prognostic factors. This development is motivated by genome-wide searches for associations between genetic variants and brain imaging phenotypes observed over time. Our primary focus is to fit nonparametric curves for age effect and model the age-genotype interaction to explore their effect on brain volume over time.

CHAPTER 2

THE BAYESIAN COVARIANCE LASSO

2.1 Introduction

In our firts paper we propose generalized priors which include common frequentist penalties like the adaptive lasso penalty of Fan et al. (2009), the lasso (L_1) penalty of Friedman et al. (2008a), and the SPICE penalty of Rothman et al. (2010) 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 Imaging (MRI) 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 among ROIs when there is no prior information about the underlying structure of the network or graph.

2.2 The General Method

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

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

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

2.2.1 Proposed Priors

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

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

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

$$\begin{aligned}\log p(\Theta|Y) =& \frac{n}{2}\log\det\Theta - \frac{n}{2}tr(S\Theta) \\ & - \sum_{k=1}^{d}\beta_k\theta_{kk} - \sum_{k=2}^{d}\sum_{k'=1}^{k-1} b_{kk'}|\theta_{kk'}| + C,\end{aligned} \tag{2.1}$$

where C is a constant independent of Θ. The popular frequentist penalized likelihoods including ACLASSO, CLASSO and SPICE can be derived from (2.1) as special cases as follows. If we choose $\beta_k = nd\lambda_{kk}/2$ and $b_{kk'} = nd\lambda_{kk'}$ (for $k > k'$), then (2.1) reduces to

$$\frac{n}{2}\{\log\det\Theta - tr(S\Theta) - \sum_{k=1}^{d}\sum_{k'=1}^{d} d\lambda_{kk'}|\theta_{kk'}|\} + C. \tag{2.2}$$

Fan et al. (2009) optimized equation (2.2) as the objective function in the ACLASSO method, which can be interpreted as the posterior mode under $\text{Exp}(nd\lambda_{kk}/2)$ priors for the diagonal elements and $\text{Laplace}(nd\lambda_{kk'})$ priors for the off-diagonal elements of the precision matrix Θ.

If we set $b_{kk'} = 2\beta_k = n\rho$, the priors for θ_{kk} are $i.i.d$ $\text{Exp}(n\rho/2)$ and the $\theta_{kk'}$ are $i.i.d$ $\text{Laplace}(n\rho)$ for $k > k'$. Then (2.1) reduces to

$$\log p(\Theta|Y) = \frac{n}{2}\left\{\log\det\Theta - tr(S\Theta) - \rho||\Theta||_{l_1}\right\} + C, \tag{2.3}$$

where $||\Theta||_{l_1} = \sum_{k=1}^{d}\sum_{k'=1}^{p}|\theta_{kk'}|$ is the l_1 norm of Θ. Banerjee et al. (2007) optimized equation (2.3) in their covariance selection method (ignoring $n/2$), while Friedman et al. (2008a) also optimized equation (2.3) in their CLASSO method, which is essentially the posterior mode under $\text{Exp}(n\rho/2)$ priors for the diagonal elements and $\text{Laplace}(n\rho)$ priors for the off-diagonal elements of Θ. Banerjee et al. (2007) has shown that (2.3) is concave in Θ, which yields that the posterior distribution of Θ is unimodal. Hence, we will use $\text{Exp}(n\rho/2)$ priors for the diagonal elements and $\text{Laplace}(n\rho)$ priors for the off-diagonal

elements of Θ so that our log-posterior is the same as the objective function of CLASSO in (2.3).

If we choose not to penalize the diagonal elements of Θ, then we can let the hyper-parameter β_k approach 0 ($\beta_k \to 0$) or equivalently choose improper uniform priors on $(0, \infty)$ for the diagonal elements of Θ. In that case, (2.3) further reduces to

$$\log p(\Theta|Y) = \frac{n}{2}\left\{\log\det\Theta - tr(S\Theta) - \rho||\Theta^{-}||_{l_1}\right\} + C, \tag{2.4}$$

where Θ^{-} has the same off-diagonal elements as Θ but all the diagonal elements are zero. Yuan and Lin (2007) and Rothman et al. (2010) used equation (2.4) as their objective function (ignoring $n/2$ and C) and calculated the posterior mode in their SPICE method.

2.2.2 Full Conditionals

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

$$\Theta = \begin{pmatrix} \Theta_{-kk} & \boldsymbol{\theta}_k \\ \boldsymbol{\theta}_k^{\mathrm{T}} & \theta_{kk} \end{pmatrix} \text{ and } S = \begin{pmatrix} S_{-kk} & \boldsymbol{s}_k \\ \boldsymbol{s}_k^{\mathrm{T}} & s_{kk} \end{pmatrix}, \tag{2.5}$$

where θ_{kk} is the kth diagonal element of Θ, $\boldsymbol{\theta}_k = (\theta_{k1}, \ldots, \theta_{k,k-1}, \theta_{k,k+1}, \ldots, \theta_{kd})^{\mathrm{T}}$ is the vector of all off-diagonal elements of the kth column, and Θ_{-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 Θ. By using the Schur decomposition (Schur, 1909), we have $\det(\Theta) = \det(\Theta_{-kk})D_k$, where $D_k = (\theta_{kk} - C_k)$ and $C_k = \boldsymbol{\theta}_k^{\mathrm{T}}\Theta_{-kk}^{-1}\boldsymbol{\theta}_k$ are scalar quantities. Similarly, s_{kk} is the kth diagonal element of S, $\boldsymbol{s}_k$ 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

Θ for $k = 1, \ldots, d$. It follows from (2.3) that the conditional densities for θ_{kk} and $\boldsymbol{\theta}_k$ can be written as follows:

$$\begin{aligned} p(\theta_{kk}|Y, \boldsymbol{\theta}_k, \Theta_{-kk}, \rho) \propto & D_k^{\frac{n}{2}} \exp\{-\frac{n}{2}(s_{kk} + \rho)\theta_{kk}\}, \\ p(\boldsymbol{\theta}_k|Y, \theta_{kk}, \Theta_{-kk}, \rho) \propto & D_k^{\frac{n}{2}} \exp\{-\frac{n}{2}(s_k^{\mathrm{T}}\boldsymbol{\theta}_k + \rho||\boldsymbol{\theta}_k||_{l_1})\} \\ & \times I(D_k > 0), \end{aligned} \tag{2.6}$$

where $I(A)$ is the indicator function of the event A. 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

$$p(\theta_{kk}|Y, \boldsymbol{\theta}_k, \Theta_{-kk}, \rho) \propto D_k^{\frac{n}{2}} \exp\{-\frac{n}{2}s_{kk}\theta_{kk}\}.$$

Note that in (2.6), we could replace D_k by $\det(\Theta)$ which is computed faster than D_k since $\boldsymbol{\theta}_k^{\mathrm{T}}\Theta_{-kk}^{-1}\boldsymbol{\theta}_k$ 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 $\boldsymbol{\theta}_k^{\mathrm{T}}\Theta_{-kk}^{-1}\boldsymbol{\theta}_k$ 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.

Theorem 1: Suppose we start with a positive definite current value of Θ and sample from

$$\begin{aligned} p(\theta_{kk}|Y, \boldsymbol{\theta}_k, \Theta_{-kk}) &\propto D_k^{\frac{n}{2}} \exp\{-\frac{n}{2}(s_{kk} + \rho)\theta_{kk}\}, \\ p(\boldsymbol{\theta}_k|Y, \theta_{kk}, \Theta_{-kk}) &\propto D_k^{\frac{n}{2}} \exp\{-\frac{n}{2}(\boldsymbol{s}_k + \rho\boldsymbol{\gamma}_k)^{\mathrm{T}}\boldsymbol{\theta}_k\} I(D_k > 0), \end{aligned}$$

where $\boldsymbol{\gamma}_k = (\gamma_{k1}, \ldots, \gamma_{kd})^{\mathrm{T}}$ and $\gamma_{kk'} = \mathrm{sign}(\theta_{kk'})$ for $k' = 1, \ldots, d$. This sampling process guarantees that we sample positive definite values of Θ at all subsequent steps.

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

2.2.3 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 Θ and the choice of the variance of the proposal density is determined from the Hessian matrix. We can write

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

The first-order derivative of the logarithm of full conditional distribution with respect to $\boldsymbol{\theta}_k$ is $0.5n\left\{D_k^{-1}D_k^{(1)} - (\boldsymbol{s}_k + \rho\boldsymbol{\gamma}_k)\right\}$, where $D_k^{(1)} = -2\Theta_{-kk}^{-1}\boldsymbol{\theta}_k$ 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

$$-0.5n\{D_k^{-1}(D_k^{-1}D_k^{(1)}D_k^{(1)^{\mathrm{T}}} + D_k^{(2)})\},$$

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 $V_k = cD_k(D_k^{-1}D_k^{(1)}D_k^{(1)^{\mathrm{T}}} - D_k^{(2)})^{-1}|_{\Theta=Q}$, where Q is a suitable estimate of Θ (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 $q(\boldsymbol{\theta}_k) \equiv N_{d-1}(\boldsymbol{\theta}_k^t, V_k)$, where $\boldsymbol{\theta}_k^t$ is the current value of the k-th off-diagonal column at iteration t. If x is the proposed value for $\boldsymbol{\theta}_k^{t+1}$, then the Metropolis-Hastings acceptance probability is $\alpha = \min\left\{1, p(\boldsymbol{x}|Y, \theta_{kk}, \Theta_{-kk})/p(\boldsymbol{\theta}_k^t|Y, \theta_{kk}, \Theta_{-kk})\right\}$. Therefore, we set

$\boldsymbol{\theta}_k^{t+1} = \boldsymbol{x}$ with probability α and $\boldsymbol{\theta}_k^{t+1} = \boldsymbol{\theta}_k^t$ with probability $1 - \alpha$.

There are several possible sampling strategies. We could sample Θ 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 Θ 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 ($\theta_{kk'}$'s) are higher than the acceptance rates of the columns $\boldsymbol{\theta}_k$. Our computations show that the replacement rate is roughly $(1-\text{acceptance rate})^2$, 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 Θ 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 $\boldsymbol{\theta}_k$ 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 Liu et al. (2000) 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 Θ. Figure 2.1 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, $\boldsymbol{w}_1, \ldots, \boldsymbol{w}_m$ from the symmetric proposal density $N_{d-1}(\boldsymbol{\theta}_k^t, V_k)$, where m is the number of tries; in our simulation we choose $m = 5$.

2. If $I(\theta_{kk} - \boldsymbol{w}_j^T \Theta_{-kk}^{-1} \boldsymbol{w}_j > 0) = 0$ for all $j = 1, \ldots, m$ then do not replace $\boldsymbol{\theta}_k$ and stop; otherwise select $\boldsymbol{w}_j$ from $\boldsymbol{w}_1, \ldots, \boldsymbol{w}_m$ with probability proportional to $p(\boldsymbol{w}_j | \theta_{kk}, \Theta_{-kk})$. Denote the selected vector as $\boldsymbol{w}$.

3. Draw $\boldsymbol{x}_1^*, \ldots, \boldsymbol{x}_{m-1}^*$ from $N_{d-1}(\boldsymbol{w}, V_k)$, and denote $\boldsymbol{x}_m^* = \boldsymbol{\theta}_k^t$.

4. Replace $\boldsymbol{\theta}_k^t$ by $\boldsymbol{w}$ with probability

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

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

Note that, in the above scheme V_k remains constant for all MCMC samples; $p(\boldsymbol{w}_j|\theta_{kk}, \Theta_{-kk})$ and $p(\boldsymbol{x}_j^*|\theta_{kk}, \Theta_{-kk})$ are in the same form as (2.6) where $\boldsymbol{\theta}_k$ is replaced by $\boldsymbol{w}_j$ 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 $\rho \sim Gamma(\alpha_0, \beta_0)$, then it could be sampled using the Gibbs sampler. The full conditional of ρ is $\rho|\alpha_0, \beta_0, \Theta, Y \sim Gamma(\alpha_0, \beta_0 + ||\Theta||_{l_1})$. This choice requires choosing appropriate values of the hyperparameters α_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 (Yang and Berger, 2007). Since BCLASSOm estimates all of the elements of Θ 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.

2.2.4 Credible Regions

Suppose we have E MCMC samples $\Theta_1, \ldots, \Theta_E$ from the posterior distribution of the d dimensional precision matrix Θ and let $\Psi_e = \log(\Theta_e)$ be the matrix logarithm of the $e-$th sample and $\Theta_e = \exp(\Psi_e)$ be the matrix exponential of Ψ_e. Note that, if $\lambda_1, \ldots, \lambda_d$ are the eigenvalues of Θ and $\gamma_1, \ldots, \gamma_d$ are the eigenvalues of Ψ, then $\boldsymbol{\gamma}_k = \log(\lambda_k)$ for $k = 1, \ldots, d$. Now, let $\bar{\Psi}$ is the posterior arithmetic mean of $\Psi_1, \ldots, \Psi_E$ then $\bar{\Theta}_G = \exp(\overline{\Psi})$ is the posterior geometric mean of Θ_e. We define the Euclidean distance between $\Psi_e = (\psi_{e,kk'})$ and the posterior mean $\bar{\Psi} = (\bar{\psi}_{kk'})$ given by

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

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 Θ.

2.3 Simulation Study

We used simulations to compare the performance of our BCLASSOm and BCLASSOs estimators with the three frequentist penalized likelihood methods namely, CLASSO (Friedman et al., 2008a), ACLASSO (Fan et al., 2009), and CSCAD (Fan et al., 2009). Among the Bayesian methods, the Yang and Berger (2007) 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 Θ is larger than the sample size (hence the matrix is singular). In Smith and Kohn (2002) and Wong et al. (2003), 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 $A = (a_{kk'})$

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 $\Theta = AA^{\mathrm{T}}$ and $\Sigma = \Theta^{-1}$. 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 $ks = 2, \ldots, d$. 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.

2.3.1 Criteria for comparison

There are several loss measures proposed for evaluating the performance in estimation of the precision and covariance matrices as discussed in Yang and Berger (2007). 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 Θ are defined as

$$\begin{aligned} L_1(\Theta, \hat{\Theta}) &= tr(\Theta^{-1}\hat{\Theta}) - \log\det(\Theta^{-1}\hat{\Theta}) - d, \\ L_2(\Theta, \hat{\Theta}) &= tr(\Theta^{-1}\hat{\Theta} - I)^2. \end{aligned} \tag{2.7}$$

where $\mathrm{vec}(A) = (a_{11}, \cdots, a_{1d}, \cdots, a_{d1}, \cdots, a_{dd})^T$ for any $d \times d$ matrix $A = (a_{kk'})$. Similar loss functions for Σ will result in the Bayes estimators $\hat{\Sigma}_{L_1} = \{E(\Theta|Y)\}^{-1}$ and

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

respectively. We use $\hat{\Theta}_{L_1} = \{E(\Sigma|Y)\}^{-1}$ and $\hat{\Sigma}_{L_1} = \{E(\Theta|Y)\}^{-1}$ in our simulation studies as the BCLASSO estimators for Θ and Σ, respectively. Since $\hat{\Theta}_{L_2}$ and $\hat{\Sigma}_{L_2}$ 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 $\hat{\Theta}_{L_1}$ and $\hat{\Sigma}_{L_1}$ 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 $\hat{\Theta}_{L_1}$ and $\hat{\Sigma}_{L_1}$ 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 Escoufer (1973) as $R(\Theta, \hat{\Theta}) = \text{tr}(\Theta\hat{\Theta})/\{\text{tr}(\Theta\Theta)\text{tr}(\hat{\Theta}\hat{\Theta})\}^{1/2}$. In this measure, the closer the estimator $\hat{\Theta}$ is to Θ, the higher the value of $R(\Theta, \hat{\Theta})$. We compared our estimates $\hat{\Theta}_{L_1}$ and $\hat{\Sigma}_{L_1}$ 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 Σ in a similar fashion. The optimal estimators minimize these loss functions. Yang and Berger (2007) showed that the Bayes (hence minimax) estimators of Θ under L_1 and L_2 are, respectively, given by

$$\hat{\Theta}_{L_1} = \hat{\Theta}_{L_1} = \{E(\Sigma|Y)\}^{-1},$$
$$\text{vec}(\hat{\Theta}_{L_2}) = \{E(\Sigma \otimes \Sigma|Y)\}^{-1}\text{vec}\{E(\Sigma|Y)\}^{-1}.$$

2.3.2 Results

Table 2.1 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 2.2 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 2.3 summarizes the mean matrix correlations and their standard deviations. In terms of the matrix correlation measure $R(\Theta, \hat{\Theta})$, 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.

2.4 Application to Real Data

Example 1: The first dataset is flow cytometry data on $d = 11$ proteins on $n = 7466$ cells from Sachs et al. (2003). In Sachs et al. (2003), 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 Friedman et al. (2008a) for comparison of the agreements

Table 2.1: 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)

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 L_1 minimax estimator; BCLASSOs = Bayesian covariance lasso with sparsity.

Table 2.2: 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 Sparse	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)
	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 2.3: 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)

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 Θ). For the maximum likelihood network in Figure 2.3(b), we used a hard-threshold that gives the same number of connections as those of Sachs et al. (2003). The penalties for the CLASSO in Figure 2.3(d), ACLASSO in Figure 2.3(e) and CSCAD in Figure 2.3(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 Sachs03, we decided to fix the penalty manually to get the same number of connections. The results are shown in Figures 2.3 (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 2.4; the Bayesian network that has the closest number of connections to that of Sachs is shown on Figure 2.3. The level of agreement of each of these 4 methods to those of Sachs' results were computed and reported in Table 2.4. The results indicate similar agreement between the networks of BCLASSO, ACLASSO and CLASSO and Sachs et al. (2003)'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

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

The posterior distribution of Θ under the lasso penalty can be written as in (2.3) and the full conditionals are given in (2.6). For the penalty parameter ρ, we take $\rho \sim \text{Gamma}(1, 1)$. 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 2.4 and the networks in Figure 2.5. 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 2.5.

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 (Gao et al., 2009). 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 2.5. 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.

2.5 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 Θ, 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 $p(\boldsymbol{\theta}_k|y, \theta_{kk}, \rho)$. 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; $\theta_{kk} \sim \text{gamma}(\alpha_k, \beta_k)$ and independent double gamma priors for the off-diagonal elements $\theta_{kk'} \sim \text{double gamma}(0, a_{kk'}, b_{kk'})$ 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 Θ is given by

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

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.

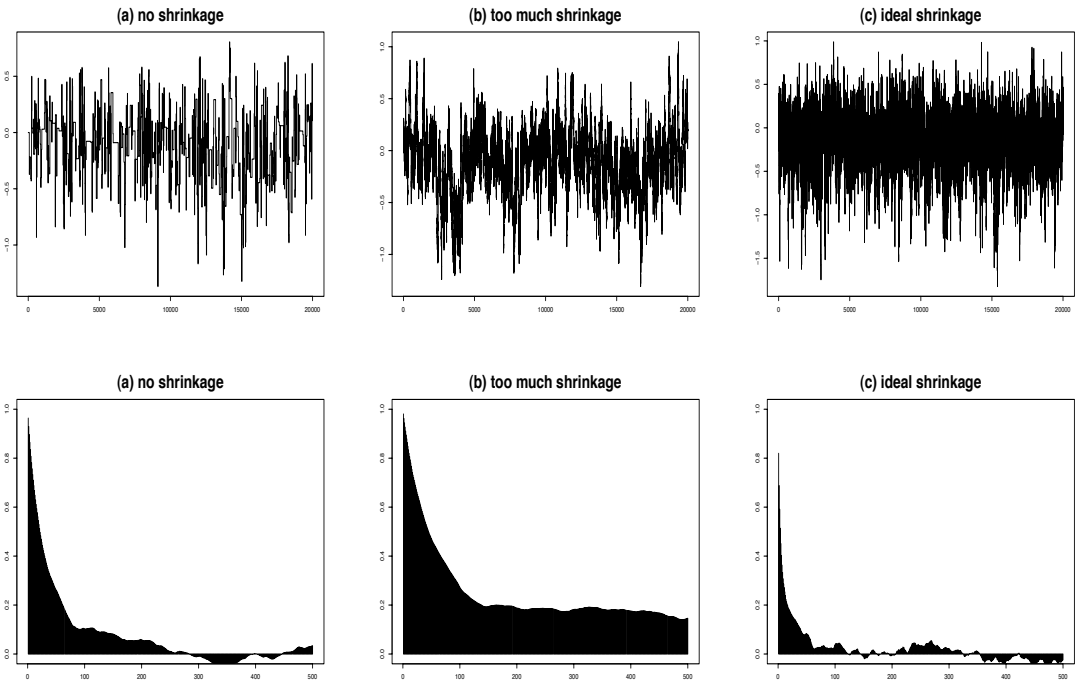

Figure 2.1: Trace plots (top row) and autocorrelation plots (bottom row) of θ_{12} for $d = 5$ and $n = 10$ showing the impact of variance tuning of the proposal density.

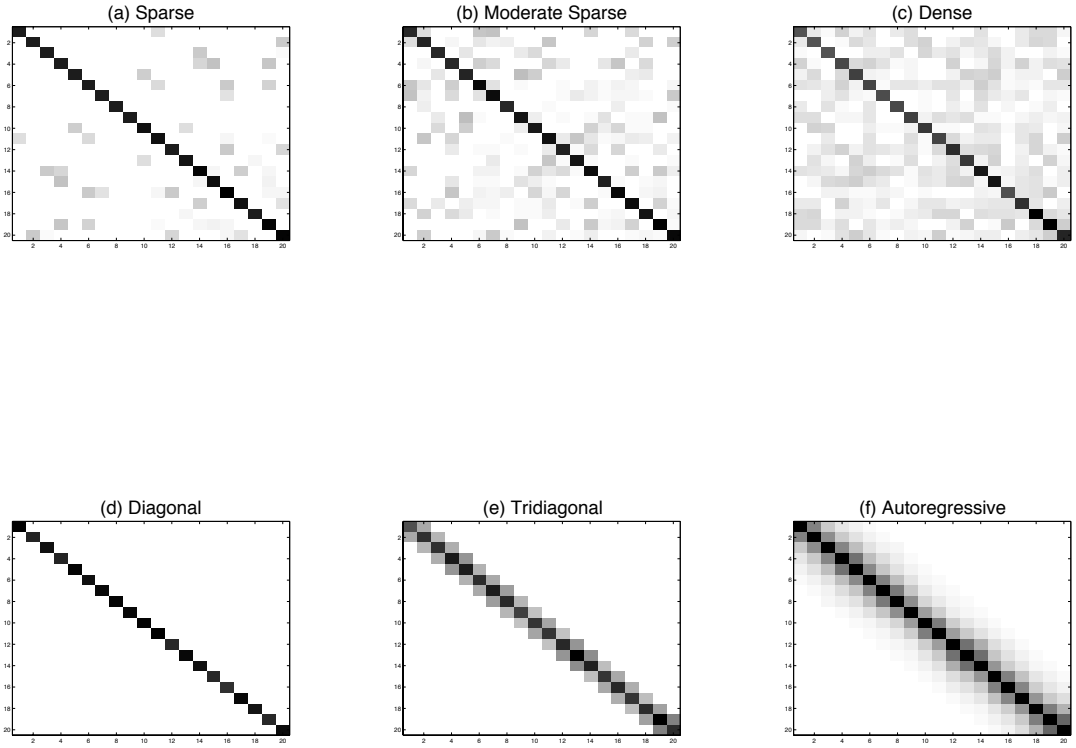

Figure 2.2: Image plots of the six types of precision matrices (Θ) considered in the simulation study. The top 3 are unstructured and the bottom 3 are structured.

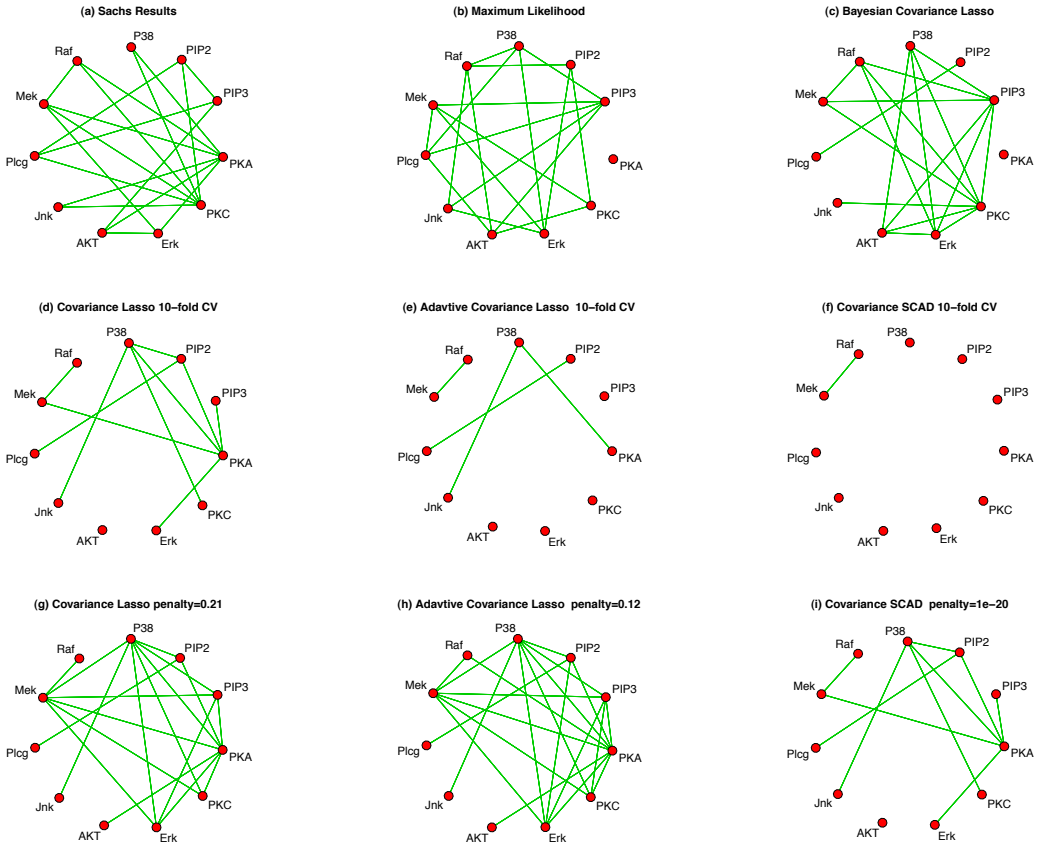

Figure 2.3: Networks for 11 proteins from Sachs et al. (2003)

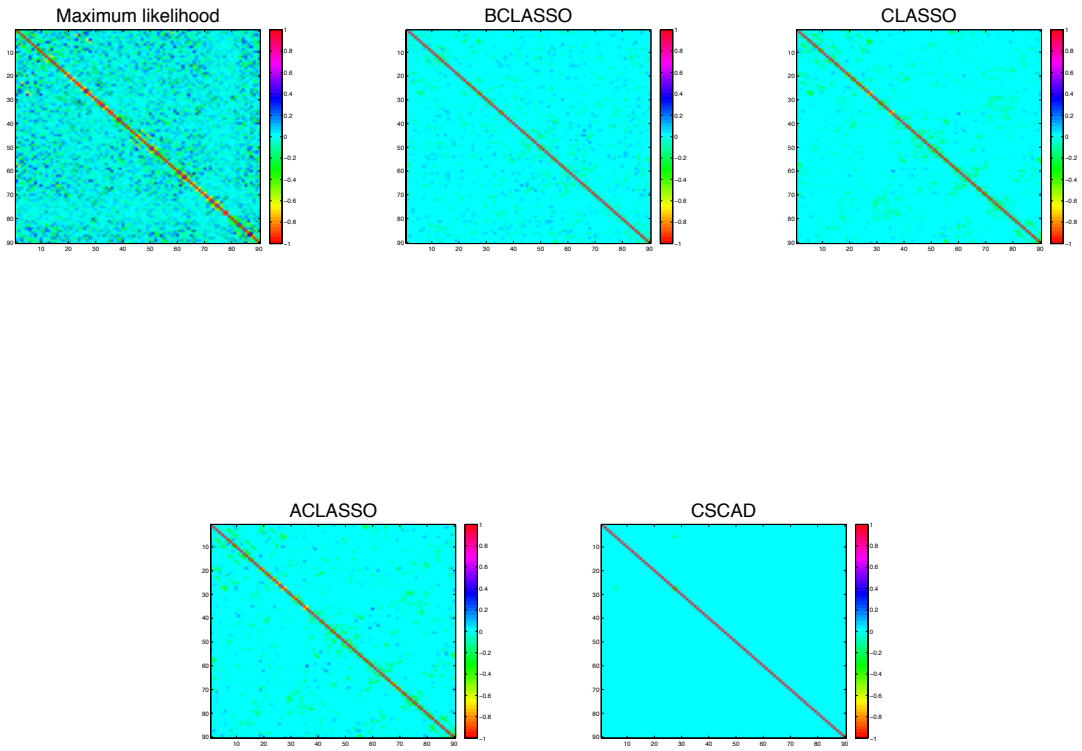

Figure 2.4: Image plots of the partial correlation matrices for 90 regions of 2-year old children' brains using the five different methods

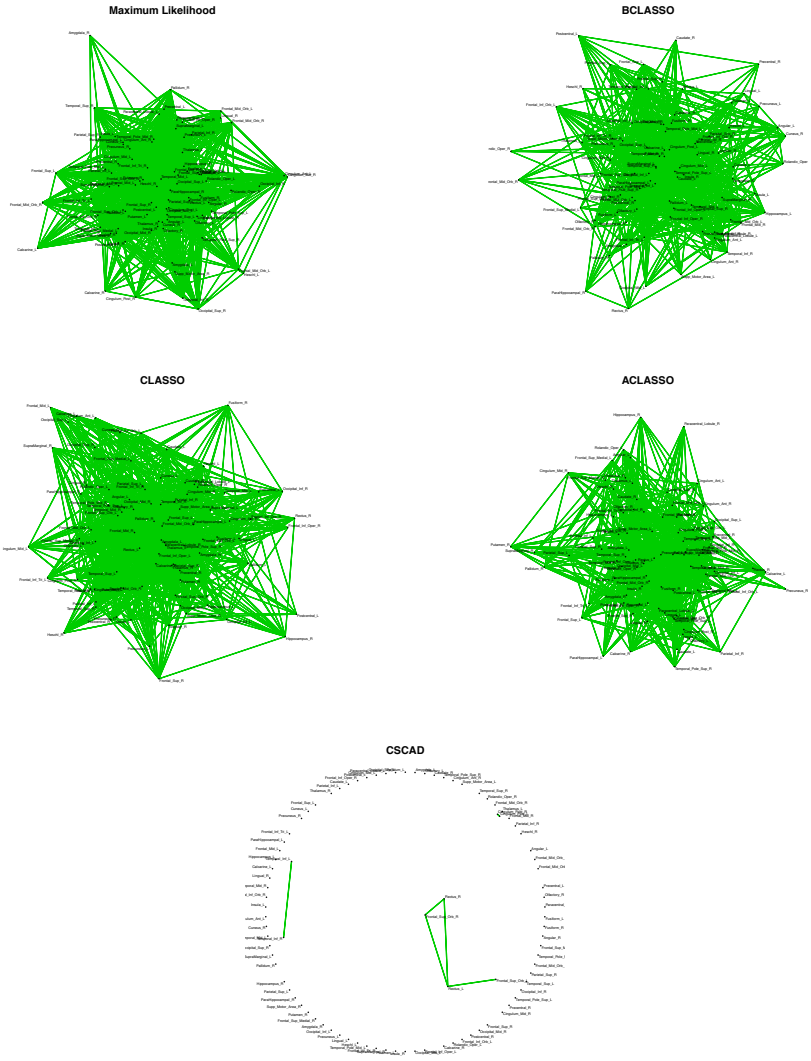

Figure 2.5: Networks for 90 regions of 2-year old children' brains using the different methods

Table 2.4: Agreement of Methods with the Results from Sachs et al. (2003)

Method	Connections	Sensitivity	Specificity	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

CI = credible interval; PPV = Positive predictive value;
NPV = Negative predictive value.

Table 2.5: ROIs 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
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 Oper R	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

CHAPTER 3

BAYESIAN GENERALIZED LOW RANK REGRESSION

3.1 Introduction

The emergence of high-dimensional data in genomics and neuroimaging, among other areas, has presented us with a large number of predictors as well as many response variables, which may have strong correlations. For instance, in imaging genetics as an emerging field, such problems frequently arise when multivariate imaging measures, such as volumes of cortical and subcortical regions of interest (ROIs), are predicted by high-dimensional covariate vectors, such as gene expressions or single nucleotide polymorphisms (SNPs). The joint analysis of imaging and genetic data may ultimately lead to discoveries of genes for some complex mental and neurological disorders, such as autism and schizophrenia (Cannon and Keller, 2006; Turner et al., 2006; Scharinger et al., 2010; Paus, 2010; Peper et al., 2007; Chiang et al., 2011a,b). This motivates us to develop low rank regression models (GLRR) for the analysis of high-dimensional responses and covariates under the high-dimension-low-sample-size setting.

Developing models for high-dimensional responses and covariates poses at least four major challenges including (i) a large number of regression parameters, (ii) a large covariance matrix, (iii) correlations among responses, and (iv) multicollinearity among predictors. When the number of responses and the number of covariates, which are

denoted by d and p, respectively, are even moderately high, fitting conventional multivariate response regression models (MRRM) usually requires estimating a $d \times p$ matrix of regression coefficients, whose number pd can be much larger than the sample size. Although accounting for complicated correlation among multiple responses is important for improving the overall prediction accuracy of multivariate analysis (Breiman and Friedman, 1997), it requires estimating $d(d+1)/2$ unknown parameters in a $d \times d$ unstructured covariance matrix. Another notorious difficulty is that the collinearity among a large number of predictors can cause issues of over-fitting and model misidentification (Fan and Lv, 2010).

There is a great interest in developing new statistical methods to handle these challenges for MRRMs. The early developments involve a separation approach- variable selection to reduce dimension and then parameter estimation, when both p and d are moderate compared to the sample size (Breiman and Friedman, 1997). For instance, Brown et al. (2002) introduced Bayesian model averaging incorporating variable selection for prediction, which allows for fast computation for dimensions up to several hundred. Recently, much attention has been given to shrinkage methods for achieving better stability and improving performance (Tibshirani, 1996). Notably, the most popular ones are the L_1 and L_2 penalties. The L_2 penalty forces the coefficients of highly correlated covariates towards each other, whereas the L_1 penalty usually selects only one predictor from a highly correlated group while ignoring the others. L_1 priors can be seen as sparse priors since they create a singularity at the origin whose gravity pulls the smaller coefficients to zero under maximum a posteriori (MAP) estimation. There are fully Bayesian approaches with sparse priors for univariate responses like the Bayesian LASSO (Park and Casella, 2008), a generalization of the LASSO (Kyung et al., 2010), and the double Pareto (Armagan et al., 2011), among many others. These methods, however, are primarily developed under the univariate-response-high-dimensional-covariate setting.

There have been several attempts in developing new methods under the high-dimensional-response-and-covariate setting. When both p and d are moderate compared to the sample size, Breiman and Friedman (1997) introduced a Curds and Whey (C&W) method to improve prediction error by accounting for correlations among the response variables. Peng et al. (2010) proposed a variant of the elastic net to enforce sparsity in the high-dimensional regression coefficient matrix, but they did not account for correlations among responses. Rothman et al. (2010) proposed a simultaneous estimation of a sparse coefficient matrix and sparse covariance matrix to improve on estimation error under the L_1 penalty. Similarly, Yin and Li (2011) presented a sparse conditional Gaussian graphical model in order to study the conditional independent relationships among a set of gene expressions adjusting for possible genetic effects. Furthermore, several authors have explored the low rank decomposition of the regression coefficient matrix and then use sparsity-inducing regularization techniques to reduce the number of parameters (Izenman, 1975; Reinsel and Velu, 1998; Tibshirani, 1996; Turlach et al., 2005; Chen et al., 2012; Vounou et al., 2010). For instance, Chen et al. (2012) and Vounou et al. (2010) considered the singular value decomposition of the coefficient matrix and used the LASSO-type penalty on both the left and right singular vectors to ensure its sparse structure. Since all variable selection methods require a selection of a proper amount of regularization for consistent variable selection, some methods, such as stability selection and cross validation, are needed for such selection (Meinshausen and Buhlmann, 2010). They, however, do not provide a standard inference tool (e.g., standard deviation) on the nonzero components of the left and right singular vectors or the coefficient matrix. Moreover, frequentist inference is the primary approach for making statistical inferences in the high-dimensional-response-and-covariate setting.

In this paper, we propose a new Bayesian GLRR to model the association between

genetic variants and brain imaging phenotypes. A low rank regression model is introduced to characterize associations between genetic variants and brain imaging phenotypes, while accounting for the impact of other covariates. We assume shrinkage priors on the singular values of the regression coefficient matrix, while not explicitly requiring orthonormality of left and right singular vectors. This facilitates fast computation of the regression coefficient matrix. We consider a sparse latent factor model to more flexibly capture the within-subject correlation structure and assume a multiplicative gamma process shrinkage priors on the factor loadings, which allow for the introduction of infinitely many factors (Bhattacharya and Dunson, 2011). We propose Bayesian local hypothesis testing to identify significant effects of genetic markers on imaging phenotypes, while controlling for multiple comparisons. Posterior computation proceeds via an efficient Markov chain Monte Carlo (MCMC) algorithm.

In Section 2, we introduce the NIH Alzheimer's Disease Neuroimaging Initiative (ADNI) dataset. In Section 3, we introduce GLRR and its associated Bayesian estimation procedure. In Section 4, we conduct simulation studies with a known ground truth to examine the finite sample performance of GLRR and compare it with the conventional LASSO method. Section 5 illustrates an application of GLRR in the joint analysis of imaging, genetic, and clinical data from ADNI. Section 6 presents concluding remarks.

3.2 Generalized Low Rank Regression Models

3.2.1 Model Setup

Consider imaging genetic data from n independent subjects in ADNI. For each subject, we observe a $d \times 1$ vector of imaging measures, denoted by $Y_i = (y_{i1}, \ldots, y_{id})^{\mathrm{T}}$, and a $p \times 1$ vector of clinical and genetic predictors, denoted by $X_i = (x_{i1}, \ldots, x_{ip})^{\mathrm{T}}$, for $i = 1, \ldots, n$. Let $\boldsymbol{Y} = (y_{ik})$ be an $n \times d$ matrix of mean centered responses, $\boldsymbol{X} = (x_{ij})$ be an $n \times p$ matrix of standardized predictors, $B = (\beta_{jk})$ be a $p \times d$ matrix of regression

coefficients, and $E = (\epsilon_{ik})$ be an $n \times d$ matrix of residuals. We consider a multivariate response regression model given by

$$Y_i = B^{\mathrm{T}} X_i + \boldsymbol{\epsilon}_i, \text{ or } \boldsymbol{Y} = \boldsymbol{X} B + E, \tag{3.1}$$

where $\boldsymbol{\epsilon}_i \sim N_d(\mathbf{0}, \Sigma = \Theta^{-1})$, in which $\Theta = \Sigma^{-1}$ is the $d \times d$ precision matrix. There are several statistical challenges in fitting model (3.1) to real data. When both p and d are relatively large compared to n, the number of parameters in B equals $p \times d$ and can be much larger than n. Furthermore, the number of unknown parameters in Σ equals $d(d+1)/2$. In addition to the number of unknown parameters, there are some additional complexities arising from practical applications, including different scales for different response variables and collinearity among the predictors.

In this model multiple responses are measured from the same subject and share a set of common predictors. Therefore, the regression coefficient matrix B can have two-way linear dependence coming from both the correlated responses and covariates. This shared mean structure can lead to a low rank mean parameter matrix B. We exploit this shared structure of B by decomposing it as

$$B = U \Delta V^{\mathrm{T}} = \sum_{l=1}^{r} B_l = \sum_{l=1}^{r} \delta_l \boldsymbol{u}_l {\varepsilon_l}^{\mathrm{T}}, \tag{3.2}$$

where r is the rank of B, $B_l = \delta_l \boldsymbol{u}_l {\varepsilon_l}^{\mathrm{T}}$ is the $l-$th layer for $l = 1, \ldots, r$, $\Delta = \mathrm{diag}(\delta_1, \ldots, \delta_r)$, $U = [\boldsymbol{u}_1, \cdots, \boldsymbol{u}_r]$ is a $p \times r$ matrix, and $V = [\varepsilon_1, \cdots, \varepsilon_r]$ is a $d \times r$ matrix. Since it is expected that only a small set of genetic variates are associated with phenotypes, a small rank of B may be able to capture the major dependence structure.

Given the large number of parameters in Σ, we consider a Bayesian factor model to

relate the random effects $\boldsymbol{\epsilon}_i$ to the latent factors $\boldsymbol{\eta}_i$ as

$$\boldsymbol{\epsilon}_i = \Lambda \boldsymbol{\eta}_i + \boldsymbol{\xi}_i, \tag{3.3}$$

where Λ is a $d \times \infty$ factor loading matrix, $\boldsymbol{\eta}_i \sim N_\infty(\mathbf{0}, \boldsymbol{I}_\infty)$, and $\boldsymbol{\xi}_i \sim N(\mathbf{0}, \Sigma_\xi)$ with $\Sigma_\xi = \text{diag}(\sigma_1^2, \ldots, \sigma_d^2)$. To achieve dimensionality reduction, one would typically restrict the dimension of the latent factor vector $\boldsymbol{\eta}_i$ to be orders of magnitude less than that of $\boldsymbol{\epsilon}_i$. By following Bhattacharya and Dunson (2011), we choose a prior that shrinks the elements of Λ to zero as the column index increases. Thus, it bypasses the challenging issue of selecting the number of factors. Finally, our GLRR integrates the low rank model (3.2) and the Bayesian factor model (3.3). Specifically, our GLRR can be written as

$$Y_i = \sum_{l=1}^{r} X_i^{\mathrm{T}} \delta_l \varepsilon_l \boldsymbol{u}_l{}^{\mathrm{T}} + \Lambda \boldsymbol{\eta}_i + \boldsymbol{\xi}_i. \tag{3.4}$$

Other than genetic markers, such as SNP's, it is common that X_i has a subvector, denoted by X_{Pi}, consisting of several prognostic variables, such as age, gender, and disease status in real applications. There are two different methods to deal with prognostic factors in the presence of genetic markers. The first method is a two-step approach. The first step is to fit the MRRM solely with these prognostics factors as covariates and then calculate the fitted residuals as adjusted responses. The final step is to fit model (3.4) to the adjusted responses with genetic markers as X. The second method is to fit model (3.4) with both prognostic factors and genetic markers as covariates. Let B_P be the $p_P \times d$ matrix of coefficients associated with the prognostic factors and X_{Si} and B_S be, respectively, the subvector of X_i and the submatrix of B associated with genetic markers. It may be reasonable to assume that B_P may be unstructured and B_S admits the decomposition given by $B_S = U_S \Delta_S V_S^{\mathrm{T}} = \sum_{l=1}^{r} B_{S,l}$. In this case, the model can be

written as

$$Y_i = B_P^{\mathrm{T}} X_{Pi} + \sum_{l=1}^{r} B_{S,l}^{\mathrm{T}} X_{Si} + \boldsymbol{\epsilon}_i. \tag{3.5}$$

We take the second approach and fit model (3.5) in real data analysis.

3.2.2 Low Rank Approximation

The decomposition (3.2) is similar to the standard singular value decomposition (SVD), but it differs from SVD. Specifically, it is unnecessary that the columns of U and V in (3.2) are orthonormal and this allows that u_{jl} and v_{jl} can take any value in $(-\infty, \infty)$, since identifiability is not critical for making inference on B. Thus, the decomposition (3.2) can be regarded as a generalization of SVD in Chen et al. (2012). Moreover, compared to SVD, this decomposition leads to better computational efficiency, since sampling a unit vector in a high-dimensional sphere is computationally difficult. Nevertheless, each layer B_l is a factorization with unit rank, which amounts to estimating a common $p \times 1$ vector of distinct regression coefficients and making the rest of the coefficients some linear combinations of this vector with d additional parameters. Within the l-th layer, each column of B_l shares the same $\boldsymbol{u}_l$ and δ_l, which facilitates the exploitation of a common dependence structure among the covariates collected from the same set of subjects. Similarly, each row of B_l shares the same ε_l and δ_l facilitating the exploitation of a common dependence structure among the responses from the same set of subjects. The number of parameters at each layer is $p + d$ and the total number of parameters equals $r \times (p + d)$. Since $r << \min(p, d)$, the use of the decomposition (3.2) leads to a huge dimension reduction.

The decomposition (3.2) differs from two other popular methods including multivariate response models and stepwise unit rank regression models. Multivariate response

models estimate a separate $p \times 1$ vector of coefficients for each response totaling $p \times d$ parameters. In frequentist analysis (Chen et al., 2012), it is common to sequentially explore each layer of B based on the ordering of Δ, which leads to stepwise unit rank regressions (SURR). Specifically, one first fits the unit rank ($r = 1$) regression with the observed $\boldsymbol{Y}$ as the response to estimate the first layer $\hat{B}_1$ and $\hat{\boldsymbol{Y}} = \boldsymbol{X}\hat{B}_1$. Subsequently, one fits another unit rank regression with $\boldsymbol{Y} - \hat{\boldsymbol{Y}}$ as the response to estimate the second layer $\hat{B}_2$. One can continue this process until the r-th rank. Thus, SURR can be viewed as a special case of GLRR.

3.2.3 Covariance Structure

The covariance structure for $\boldsymbol{Y}_i$ is given by

$$\Sigma = \Theta^{-1} = \Lambda\Lambda^{\mathrm{T}} + \Sigma_{\xi}. \tag{3.6}$$

It is common to impose a constraint on Λ to define a unique model free from identification problems, since Σ is invariant under the transformation $\Lambda^* = \Lambda P$ for any semi-orthogonal matrix P with $PP^{\mathrm{T}} = I$. For instance, for identifiability purposes, one may impose a full rank lower triangular constraint, which implicitly specifies an order dependence among the responses (Geweke and Zhou, 1996). However, it is unnecessary to impose such a constraint on Λ if our primary interest is on covariance matrix estimation. Specifically, we will specify a multiplicative gamma process shrinkage prior in (4.6) on a parameter expanded loading matrix with redundant parameters. The induced prior on Σ is invariant to the ordering of the responses. This shrinkage prior adaptively selects a truncation of the infinite loadings to one having finite columns. Thus, it facilitates the posterior computation and provides an accurate approximation to the infinite factor model.

3.2.4 Priors

We first consider the priors on the elements of all layers B_l. When dealing with two highly correlated covariates, the L_1 prior tends to pick one and drop the other since it is typically a least angle selection approach to force some coefficients to zero, whereas the L_2 prior tends to force the coefficients towards each other to produce two highly correlated coefficients. In GLRR, since our primary interest is to exploit the potential two-way correlations among the estimated coefficients, we choose the L_2 prior. Let $\mathrm{Ga}(a, b)$ be a gamma distribution with scale a and shape b. Specifically, we choose

$$
\begin{aligned}
&\delta_l \sim N(0, \tau_\delta^{-1}) \text{ with } \tau_\delta \sim \mathrm{Ga}(a_0, b_0),\\
&\boldsymbol{u}_l \sim N_p(0, \tau_u^{-1} I_p) \text{ with } \tau_u \sim p + \mathrm{Ga}(c_0, d_0),\\
&\varepsilon_l \sim N_d\left(0, \mathrm{diag}(\tau_{v,1}^{-1}, \ldots, \tau_{v,d}^{-1})\right) \text{ with } \tau_{v,1}, \ldots, \tau_{v,d} \sim d + \mathrm{Ga}(e_0, f_0),
\end{aligned}
$$

where a_0, b_0, c_0, d_0, e_0, and f_0 are prefixed hyper-parameters. The number of predictors p is included in the hyperprior of τ_u to have a positive-definite covariance matrix of high dimensional $\boldsymbol{u}_l$ and fix the scale of $\boldsymbol{u}_l$. Similarly, we add the dimension d to all hyper-priors for $\tau_{v,l}$. Moreover, we standardize all predictors to have zero mean and unit variance, and thus a single prior is sensible for all elements of $\boldsymbol{u}_l$. The varying dispersions $\tau_{v,1}, \ldots, \tau_{v,d}$ are chosen to account for different scales of different responses. For example, the volumes of different ROIs vary dramatically across ROIs, so it is more sensible to use separate dispersions for different ROIs.

We place the multiplicative gamma process shrinkage prior (Bhattacharya and Dunson, 2011) on Λ in order to increasingly shrink the factor loadings towards zero with the column index. Such shrinkage priors avoid the drawback of order dependence from the lower triangular constraint on Λ for identifiability. We use inverse gamma priors on the

diagonal elements of Σ_ξ. Specifically, these priors are given as follows:

$$\begin{aligned} \Lambda &= \{\lambda_{kh}\}, \;\; k = 1, \ldots, d \,; h = 1, \ldots, \infty, \\ \lambda_{kh}|\phi_{kh}, \tau_{\lambda h} &\sim N(0, \phi_{kh}^{-1}\tau_{\lambda h}^{-1}), \;\; \phi_{kh} \sim \mathrm{Ga}(v/2, v/2), \sigma_k^{-2} \sim \mathrm{Ga}(a_{\sigma k}, b_{\sigma k}), \\ \psi_1 &\sim \mathrm{Ga}(a_1, 1), \;\; \psi_g \sim \mathrm{Ga}(a_2, 1), \;\; g \geq 2, \; \tau_{\lambda h} = \prod_{l=1}^{h} \psi_l, \end{aligned} \tag{3.7}$$

where ψ_g for $g = 1, \ldots, \infty$ are independent random variables, $\tau_{\lambda h}$ is a global shrinkage parameter for the h-th column, and the ϕ_{kh}s are local shrinkage parameters for the elements in the h-th column. Moreover, v, a_1,, a_2, $a_{\sigma k}$ and $b_{\sigma k}$ are prefixed hyper-parameters. When $a_2 > 1$, the $\tau_{\lambda h}$'s increase stochastically with the column index h, which indicates more shrinkage favored over the columns of higher indices. The loading component specific prior precision $\phi_{kh}^{-1}\tau_{\lambda h}^{-1}$ allows shrinking the components of Λ. Straightforward Gibbs sampler can be applied for posterior computation.

3.2.5 Posterior Computation

We propose a straightforward Gibbs sampler for posterior computation after truncating the loadings matrix to have $k_* << d$ columns. An adaptive strategy for inference on the truncation level k_* has been described in (Bhattacharya and Dunson, 2011). The Gibbs sampler is computationally efficient and mixes rapidly. Starting from the initiation step, the Gibbs sampler at the truncated level k_* proceeds as follows:

1. Update $(\boldsymbol{u}_l, \tau_u)$ according to their conditional distributions

$$\begin{aligned} p(\boldsymbol{u}_l|-) \sim & \boldsymbol{N}_p\left(\delta_l \Sigma_{\boldsymbol{u}_l} \boldsymbol{X}^{\mathrm{T}} Y_l \Theta \varepsilon_l, \Sigma_{\boldsymbol{u}_l}\right), \\ & p(\tau_u|-) \sim p + \mathrm{Ga}\left(c_0, d_0 + 0.5 \sum_{l=1}^{r} \boldsymbol{u}_l^{\mathrm{T}} \boldsymbol{u}_l\right), \end{aligned}$$

where $\Sigma_{\boldsymbol{u}_l} = \left\{\tau_u I_p + \delta_l^2 (\varepsilon_l^{\mathrm{T}} \Theta \varepsilon_l) \boldsymbol{X}^{\mathrm{T}} \boldsymbol{X}\right\}^{-1}$.

2. Update $(\varepsilon_l, \tau_{v,k})$ according to their conditional distributions

$$p(\varepsilon_l|-) \sim \boldsymbol{N}_d(\delta_l \Sigma_{\varepsilon_l} \Theta Y_l^{\mathrm{T}} \boldsymbol{X} \boldsymbol{u}_l, \Sigma_{\varepsilon_l}),$$
$$p(\tau_{v,k}|-) \sim d + \mathrm{Ga}\left(e_0, f_0 + 0.5 \sum_{l=1}^{r} \varepsilon_l^2\right)$$

for $k = 1, \ldots, d$, where $\Sigma_{\varepsilon_l} = \left\{\mathrm{diag}(\tau_{v,1}, \ldots, \tau_{v,d}) + \delta_l^2 (\boldsymbol{u}_l^{\mathrm{T}} \boldsymbol{X}^{\mathrm{T}} \boldsymbol{X} \boldsymbol{u}_l)\Theta\right\}^{-1}$.

3. Update (δ_l, τ_δ) according to their conditional distributions

$$p(\delta_l|-) \sim N(\sigma_{\delta_l}^2 \boldsymbol{u}_l^{\mathrm{T}} \boldsymbol{X}^{\mathrm{T}} E_l \Theta \varepsilon_l, \sigma_{\delta_l}^2),$$
$$p(\tau_\delta|-) \sim \mathrm{Ga}\left(a_0, b_0 + 0.5 \sum_{l=1}^{r} \delta_l^2\right),$$

where $\sigma_{\delta_l}^2 = \left\{\tau_\delta + (\varepsilon_l \Theta^{\mathrm{T}} \varepsilon_l)(\boldsymbol{u}_l^{\mathrm{T}} \boldsymbol{X}^{\mathrm{T}} \boldsymbol{X} \boldsymbol{u}_l)\right\}^{-1}$.

4. Update the kth row of Λ_{k_*}, denoted by λ_k, from its conditional distribution

$$p(\lambda_k|-) \sim \boldsymbol{N}((\sigma_k^{-2} \boldsymbol{\eta}^{\mathrm{T}} \boldsymbol{\eta} + D_k^{-1})^{-1} \boldsymbol{\eta}^{\mathrm{T}} \sigma_k^{-2} E_k, (\sigma_k^{-2} \eta^{\mathrm{T}} \eta + D_k^{-1})^{-1}),$$

where $\boldsymbol{\eta} = (\eta_1, \ldots, \eta_n)^{\mathrm{T}}$, $E_k = (\epsilon_{1k}, \ldots, \epsilon_{nk})^{\mathrm{T}}$ is the kth column of $E = \boldsymbol{Y} - \boldsymbol{X} B$, and $D_k = \mathrm{diag}(\phi_{k1}^{-1} \tau_{\lambda 1}^{-1}, \ldots, \phi_{kk_*}^{-1} \tau_{\lambda k_*}^{-1})$ for $k = 1, \ldots, d$.

5. Update ϕ_{kh} from its conditional distribution

$$p(\phi_{kh}|-) \sim \mathrm{Ga}\left(\frac{v+1}{2}, \frac{v + \lambda_{kh}^2 \tau_{\lambda h}}{2}\right).$$

6. Update ψ_1 from its conditional distribution

$$p(\psi_1|-) \sim \mathrm{Ga}\left(a_1 + \frac{1}{2} d k_*, 1 + \frac{1}{2} \sum_{g=1}^{k_*} \tau_{\lambda g}^{(h)} \sum_{k=1}^{d} \phi_{kg} \lambda_{kg}^2\right),$$

and update ψ_h, $h \geq 2$ from its conditional distribution

$$p(\psi_h|-) \sim \text{Ga}\left(a_2 + \frac{1}{2}d(k_* - h + 1), 1 + \frac{1}{2}\sum_{g=h}^{k^*}\tau_{\lambda g}^{(h)}\sum_{k=1}^{d}\phi_{kg}\lambda_{kg}^2\right),$$

where $\tau_{\lambda g}^{(h)} = \prod_{t=1,t\neq h}^{g}\psi_t$ for $h = 1, \ldots, k_*$.

7. Update σ_k^{-2}, $k = 1, \ldots, d$, from its conditional distribution

$$p(\sigma_k^{-2}|-) \sim \text{Ga}(a_{\sigma k} + \frac{n}{2}, b_{\sigma k} + \frac{1}{2}\sum_{i=1}^{n}(y_{ik} - \lambda_k^{\text{T}}\eta_i)^2).$$

8. Update η_i, $i = 1, \ldots, n$, from conditionally independent posteriors

$$p(\eta_i|-) \sim N((I_{k_*} + \Lambda^{\text{T}}\Theta\Lambda)^{-1}\Lambda_{k_*}^{\text{T}}\Theta\boldsymbol{\epsilon}_i, (I_{k_*} + \Lambda^{\text{T}}\Theta\Lambda)^{-1}),$$

where $\boldsymbol{\epsilon}_i$ is the ith row of E.

3.2.6 Determining the Rank of B

We consider different methods for determining the rank of B. For frequentist inference, many regularization methods have been developed to recover the low rank structure of a matrix, such as B, by shrinking δ_ℓ's to zero in (3.2) (Chen et al., 2012). For Bayesian inference, it may be tempting to use Bayesian model averaging and allow varying number of layers in order to improve prediction performance, but it limits us on making statistical inference on each layer of B, U, and V. We take a fixed-rank approach and use some selection criteria to choose an optimal value of r. Specifically, we consider five different selection criteria including the Akaike information criterion (AIC), the Bayesian information criterion (BIC), the normalized prediction error (PEN), the multivariate R^2, and the normalized model error (MEN) for GLRR.

Let $\hat{\boldsymbol{Y}} = \boldsymbol{X}\hat{B}$, where $\hat{B}$ is the posterior estimate of B based on the MCMC samples. Let $\text{SSE} = \text{tr}((\widehat{\boldsymbol{Y}} - \boldsymbol{Y})^{\text{T}}(\widehat{\boldsymbol{Y}} - \boldsymbol{Y}))$ be the error sum of squares and $p_* = r(p+d)$ be the number of parameters in B. The five evaluation criteria are, respectively, given by

$$\begin{aligned}
&\text{AIC} = \log(\text{SSE}) + 2\frac{p_*}{nd},\ \text{BIC} = \log(\text{SSE}) + \frac{\log(nd)}{nd}p_*, \\
&\text{PEN}(\widehat{\boldsymbol{Y}}, \boldsymbol{Y}) = \frac{\text{SSE}}{\text{tr}(\boldsymbol{Y}^{\text{T}}\boldsymbol{Y})} \times 100, R^2(\widehat{\boldsymbol{Y}}, \boldsymbol{Y}) = \frac{\text{tr}\left(\widehat{\boldsymbol{Y}}^{\text{T}}\widehat{\boldsymbol{Y}}\right)}{\text{tr}(\boldsymbol{Y}^{\text{T}}\boldsymbol{Y})} \times 100, \qquad (3.8) \\
&\text{MEN}(\hat{B}, B) = \frac{\text{tr}((\hat{B} - B)^{\text{T}}\Sigma_X(\hat{B} - B))}{\text{tr}(B^{\text{T}}\Sigma_X B)} \times 100.
\end{aligned}$$

The numerator and denominator of the MEN are, respectively, the model error and measurement error of model (3.4) (Yuan et al., 2007). Thus, the MEN is the ratio of the model error over the measurement error as a percentage of the total magnitude of all parameters. Similarly, the PEN and R^2 are defined as percentages, which makes comparisons more meaningful and readily comparable across studies.

To illustrate the effectiveness of all five criteria, we independently simulated 100 data sets from model (3.4) with $(n, p, d) = (100, 200, 100)$ and a rank 5 matrix B. For each simulated data set, we used the Gibbs sampler to draw posterior samples to estimate B and then calculated the five selection criteria in (3.8) as the rank varied from 1 to 10. Finally, based on all 100 simulated data sets, we calculated the mean and standard deviation of each selection criterion as the rank varied from 1 to 10. As shown in Figure 3.1, PEN, MEN, R^2, and AIC stabilize around the true rank, whereas BIC reaches the minimum at the true rank. This may indicate that BIC outperforms other selection criteria for determining the true rank of B.

3.2.7 Thresholding

Based on the MCMC samples obtained from the Gibbs sampler, we are able to identify three different sets of information including (i) SNPs that significantly contribute to a large portion of imaging phenotypes, (ii) imaging phenotypes that are associated with those SNPs in (i), and (iii) important individual SNP effects on individual imaging phenotypes. Statistically, (i), (ii), and (iii) can be formulated as testing significant elements in U, V, and B, respectively. For the sake of space, we focus on (i). Suppose that we draw a set of MCMC samples $U^{(m)} = (u_{jl}^{(m)})$ for $m = 1, \ldots, M$. Due to the magnitude ambiguity of U, we normalize each column of $U = (u_{jl})$ to calculate $U^* = (u_{jl}^*)$. Moreover, we develop a specific strategy to deal with the sign ambiguity of U^*. For the $l-$th column of U^*, we use the normalized MCMC samples $U^{(m)*} = (u_{jl}^{(m)*})$ to empirically determine the j_0-th row such that $P(|u_{j_0l}^*| = \max_j |\tilde{u}_{jl}^*|) \geq P(|\tilde{u}_{j'l}^*| = \max_j |\tilde{u}_{jl}^*|)$ for all $j' \neq j_0$. Then, we fix $u_{j_0l}^{(m)*}$ to be positive for $l = 1, \ldots, r$ and $m = 1, \ldots, M$.

To detect SNPs in (i), we suggest to calculate the median and median absolute deviation (MAD) of $u_{jl}^{(m)*}$, denoted by $\hat{u}_{jl}^*$ and $s_{u,jl}$, respectively, since the MCMC samples $\{u_{jl}^{(m)}\}$ may oscillate dramatically between the positive solution and the negative solution due to the sign ambiguity for all j, l. Then, one may formulate it as testing the local null and alternative hypotheses for $|u_{jl}^*|$ relative to $s_{u,jl}$ given by

$$H_{0,jl} : \left|u_{jl}^*\right| \leq T* \text{ versus } H_{1,jl} : \left|u_{jl}^*\right| > T*,$$

where $T*$ is a specific threshold for each u_{jl}^*. One may calculate the probability of $\left|u_{jl}^*\right| T* = |\hat{u}_{jl}^*|/(1.4826 s_{u,jl})$ given the observed data and then adjust for multiple comparisons (Müller et al., 2004; Wang and Dunson, 2010). Another approach is to directly calculate $t_{u,jl}$ and apply standard multiple comparison methods, such as the false discovery rate, to determine $T*$ (Benjamini and Hochberg, 1995). We have found that

these two methods lead to similar results, and thus we take the second approach. Moreover, this Bayesian thresholding method works well even when different responses are not on the same scale. Compared to the 'hard' thresholding methods used in shrinkage methods (Chen et al., 2012; Peng et al., 2010; Rothman et al., 2010; Yin and Li, 2011), this Bayesian thresholding method accounts for the variation of each u_{jl}^* and has a probabilistic interpretation.

3.3 Simulation Study

3.3.1 Simulation Setup

We carried out some simulation studies to examine the finite-sample performance of the GLRR and its posterior computation. We generated all simulated data according to model (3.4). The simulation studies were designed to establish the association between a relatively high-dimensional phenotype vector with a set of continuous covariates or a set of commonly used genetic markers (e.g., SNP). For each case, 100 simulated data sets were generated.

We simulated $\boldsymbol{\epsilon}_i \sim N_d(0, \Sigma)$ and used two types of covariates including (i) continuous covariates generated from $X_i \sim N_p(0, \Sigma_X)$ and (ii) actual SNPs from ADNI data set. We determined Σ and Σ_X as follows. Let p_0 be the binomial probability, which controls the sparsity of the precision matrix. We first generated a $p \times p$ matrix $A = (a_{jj'})$ with $a_{jj} = 1$ and $a_{jj'} = \text{uniform}(0,1) \times \text{binomial}(1, p_0)$ for $j \neq j'$, set $\Sigma_X = AA^{\mathrm{T}}$, and standardized Σ_X into a correlation matrix such that $\Sigma_{X,jj} = 1$ for $j = 1, \ldots, d$. Similarly, we used the same method to generate Σ, the covariance matrix of $\boldsymbol{\epsilon}_i$. For both Σ and Σ_X, we set about 20% of the elements of Σ^{-1} and Σ_X^{-1} to be zero, yielding that the means of the absolute correlations of Σ and Σ_X are close to 0.40, respectively. We chose actual SNPs from the ADNI data set. Specifically, we only considered the 10,479 SNPs collected on chromosome 19, screened out all SNPs with more than 5% missing data

and minor allele frequency (MAF)$<$ 0.05, and randomly selected 400 SNPs from the remaining SNPs. For $n = 1,000$ case, 500 subjects were randomly chosen and then replicated twice, whereas for the n=100 case, 100 subjects were randomly chosen from ADNI data set.

We considered five structures of B in order to examine the finite-sample performance of GLRR under different scenarios.

- Case 1: $X_i \sim N_p(0, \Sigma_X)$ and a "+" structure was preset for B with $(p, d) = (100, 100)$ with the elements of B being set as either 0 or 1.

- Case 2: $X_i \sim N_p(0, \Sigma_X)$ and B was set as a 200×100 matrix with the true rank $r_0 = 5$. Specifically, we set $B = U\Delta V$ with $U = (u_{jl})$, $\Delta = \text{diag}(\delta_{ll}) = \text{diag}(100, 80, 60, 40, 20)$, and $V = (v_{lk})$ being 200×5, 5×5, and 5×100 matrices, respectively. Moreover, we generated all elements u_{jl} and v_{lk} independently from a $N(0, 1)$ generator and then orthonormalized U and V.

- Case 3: Covariates are actual SNPs and B has the same structure as that in Case 2 but with $(p, d) = (400, 100)$.

- Case 4: $X_i \sim N_p(0, \Sigma_X)$ and B was set as a 200×100 matrix with high degrees of correlation among elements with an average absolute correlation of 0.8, and then 20% of the elements of B were randomly forced to 0. After enforcing zeros, the true rank is 100 and the average absolute correlation is close to 0.7.

- Case 5: Covariates are actual SNPs and B is the same as that in Case 4 with $(p, d) = (400, 100)$.

We chose noninformative priors for the hyperparameters of B and set $\alpha_0 = \beta_0 = a_0 = b_0 = c_0 = d_0 = e_0 = f_0 = 10^{-6}$. Since shrinkage is achieved through dimension reduction by choosing $r << \min(d, p)$, these noninformative choices of the hyperparameters suit

well. For the hyperparameters of Σ, we chose somewhat informative priors in order to impose the positive-definiteness constraint and set $\nu = a_1 = a_2 = a_{\sigma k} = a_{\sigma k} = 1$ for $k = 1, \ldots, d$. For each simulated dataset, we ran the Gibbs sampler for 10,000 iterations with $5,000$ burn-in iterations.

As a comparison, we considered a multivariate version of LASSO (Peng et al., 2010), Bayesian LASSO (BLASSO) (Park and Casella, 2008), and group-sparse multitask regression and feature selection (G-SMuRFS) (Wang et al., 2012) for all simulated data. For LASSO, we fitted d separate LASSO regressions to each response with a single tuning parameter across all responses by using a 5-fold cross validation. Since variances of all columns X and E are relatively equal, the variances of all columns of Y should be close to each other. In this case, a single tuning parameter is sensible. For BLASSO, we chose single priors for each column of the response matrix by setting all hyperparameters to unity. For G-SMuRFS, we used single group and selected the optimal values of the penalty parameters by using a 5-fold cross validation.

To compare different methods, we calculated their sensitivity and specificity scores under each scenario. For all regularization methods, since we choose all possible values of the tuning parameters for calculating their sensitivity and specificity scores, it is unnecessary to use the cross validation method to select the tuning parameters. Let $I(\cdot)$ be an indicator function of an event and $t_{jk} = \hat{\beta}_{jk}/s_{\beta,jk}$, where $\hat{\beta}_{jk}$ and $s_{\beta,jk}$ denote the posterior mean and standard deviation of β_{jk}, respectively. Specifically, for a given threshold T_0, sensitivity and specificity scores are, respectively, given by

$$\mathrm{Se}(T_0) = \frac{\mathrm{TP}(T_0)}{\mathrm{TP}(T_0) + \mathrm{FN}(T_0)}, \text{ and } \mathrm{Sp}(T_0) = \frac{\mathrm{TN}(T_0)}{\mathrm{TN}(T_0) + \mathrm{FP}(T_0)},$$

where $\mathrm{TP}(T_0)$, $\mathrm{TP}(T_0)$, $\mathrm{TP}(T_0)$, and $\mathrm{TP}(T_0)$ are, respectively, the numbers of true positives, false positives, true negatives, and false negatives, given by

$$\begin{aligned}
\mathrm{TP}(T_0) &= \sum_{j,k} I(|t_{jk}| > T_0) I(\beta_{jk} \neq 0),\ \mathrm{FP}(T_0) = \sum_{j,k} I(|t_{jk}| > T_0) I(\beta_{jk} = 0), \\
\mathrm{TN}(T_0) &= \sum_{j,k} I(|t_{jk}| \leq T_0) I(\beta_{jk} = 0),\ \mathrm{FN}(T_0) = \sum_{j,k} I(|t_{jk}| \leq T_0) I(\beta_{jk} \neq 0).
\end{aligned}$$

Varying T_0 gives different sensitivity and specificity scores, which allow us to create receiver operating characteristic (ROC) curves. In each ROC curve, sensitivity is plotted against 1-specificity. The larger the area under the ROC curve, the better a method in identifying the true positives while controlling for the false positives.

3.3.2 Results

We first performed a preliminary analysis by using five data sets simulated according to the five structures of B and $n = 1,000$. See Figure 3.2 for the true B and estimated $\hat{B}$ by using GLRR3 (GLRR with $r = 3$), GLRR5 (GLRR with $r = 5$), BLASSO, G-SMuRFS, and LASSO under Case 1-Case 5. Inspecting Figure 3.2 reveals that for relatively large sample sizes, the fitted GLRR with r close to the true rank does a better job in recovering the underlying structure of B, while BLASSO and G-SMuRFS perform reasonably well for all cases. For the "+" structure of B with the true rank $r_0 = 2$ in Case 1, GLRR3 performs the best, whereas LASSO does a poor job. For B with the true rank $r_0 = 5$ in Cases 2 and 3, GLRR5 performs the best. The LASSO method performs reasonably well in recovering B for continuous X, when B is a 200×100 matrix, whereas it performs poorly when X is the SNP matrix. For the high-rank B in Cases 4 and 5, LASSO performs the best in recovering B, while GLRR3 and GLRR5 perform reasonably well.

Secondly, we examined the finite sample performance of LASSO, BLASSO, G-SMuRFS,

GLRR3, and GLRR5 under Cases 1-5 for $n = 100$. In each case, 100 simulated data sets were used and the mean and standard deviation of each of the five selection criteria were calculated. The results are presented in Table 3.1. Inspecting Table 3.1 reveals that GLRRs outperform LASSO in most cases. As p increases, GLRRs outperform LASSO in terms of MEN, PEN, and R^2. Under Cases 3 and 5, GLRRs outperform LASSO with much smaller errors as well as lower standard deviations for MEN and PEN. LASSO performs much better for continuous covariates than for discrete SNPs, but such patterns do not appear for GLRRs. The results of GLRRs and BLASSO are comparable in terms of both AIC and BIC, but the number of parameters under GLRRs is much smaller than that under BLASSO. BLASSO and G-SMuRFS perform well in terms of both model error and prediction, but that comes at a higher cost since these methods have all non-zero estimates to push BIC very high. The high R^2 and low prediction error of BLASSO and G-SMuRFS in the high dimension cases may be caused by over-fitting and model misidentification (Fan and Lv, 2010).

Thirdly, we used the ROC curve to compare LASSO, BLASSO, G-SMuRFS, GLRR3, and GLRR5 under Cases 1-5. See Figure 3.3 for details. For Case 1, BLASSO demonstrates consistently the best power for almost every level of specificity, while G-SMuRFS is the second best. GLRR3 and GLRR5 fall in the middle. For Case 4, all the methods appear to be comparable with GLRR3 and GLRR5. For Cases 2, 3, and 5, GLRRs consistently outperform all other methods.

We also compared the timing of each method in a personal laptop with Intel Core i5 1.7 GHz processor and 4 GB memory. It takes LASSO and G-SMuRFS roughly 5 minutes to choose the optimal penalty and calculate estimates for a single sample of Case 5. All Bayesian methods take much longer since one has to sample many MCMC samples. Specifically, BLASSO takes about 2.75 hours to generate 10,000 samples plus

Table 3.1: Empirical comparison of GLRR3, GLRR5, LASSO, BLASSO and G-SMuRFS under Cases 1-5 based on the five selection criteria. The means and standard deviations of these criteria are also calculated and their standard deviations are presented in parentheses. Moreover, UN denotes the unstructured B.

Case/(p,d,n)/X	B/r_0	Method	MEN	PEN	R^2	AIC	BIC
1	"+"	LASSO	6.21 (0.83)	3.98 (0.73)	89.50 (2.03)	9.54 (0.08)	12.45 (0.24)
(100, 100, 100)	2	BLASSO	4.63 (0.48)	4.19 (0.57)	92.09 (1.13)	10.82 (0.08)	18.02 (0.07)
Continuous		G-SMuRFS	5.74 (0.65)	1.88 (0.30)	94.62 (0.71)	10.01 (0.10)	17.22 (0.10)
		GLRR3	11.81 (10.96)	3.17 (7.35)	94.71 (7.34)	8.27 (0.47)	8.70 (0.47)
		GLRR5	6.81 (6.87)	2.14 (3.69)	94.73 (3.72)	8.16 (0.31)	8.88 (0.31)
2	$U\Delta V$	LASSO	26.64 (2.15)	1.26 (1.49)	97.37 (6.01)	11.94 (0.77)	18.25 (0.93)
(200, 100, 100)	5	BLASSO	22.38 (1.63)	1.24 (0.21)	98.55 (0.33)	13.39 (0.41)	27.78 (0.41)
Continuous		G-SMuRFS	21.87 (1.69)	1.11 (0.11)	98.20 (0.07)	9.95 (0.10)	24.38 (0.24)
		GLRR3	31.56 (6.56)	14.60 (9.24)	83.77 (9.75)	12.88 (0.61)	13.53 (0.61)
		GLRR5	21.69 (1.87)	0.36 (1.93)	98.54 (2.22)	8.90 (0.47)	8.88 (0.47)
3	$U\Delta V$	LASSO	50.41 (12.00)	50.87 (11.94)	19.59 (7.78)	12.81 (0.15)	13.66 (0.23)
(400, 100, 100)	5	BLASSO	25.57 (0.02)	10.08 (2.04)	93.24 (3.29)	15.69 (0.39)	43.53 (0.39)
SNPs		G-SMuRFS	24.28 (0.02)	10.27 (2.01)	91.69 (4.01)	16.96 (0.01)	42.80 (0.01)
		GLRR3	20.23 (4.25)	21.39 (13.96)	76.74 (14.20)	11.86 (0.60)	12.93 (0.60)
		GLRR5	13.64 (7.88)	4.07 (7.77)	93.60 (7.97)	10.19 (0.60)	11.99 (0.60)
4	UN	LASSO	22.16 (1.93)	1.27 (1.49)	94.40 (6.16)	12.45 (0.88)	19.11 (1.12)
(200, 100, 100)	100	BLASSO	19.44 (1.16)	1.22 (0.21)	98.32 (0.74)	13.21 (0.79)	27.63 (0.79)
Continuous		G-SMuRFS	15.00 (1.44)	1.10 (0.01)	98.40 (0.08)	9.74 (0.11)	24.16 (0.11)
		GLRR3	18.32 (1.53)	5.16 (0.64)	93.93 (0.73)	12.20 (0.04)	12.85 (0.04)
		GLRR5	16.02 (1.59)	4.30 (0.56)	94.26 (0.71)	12.14 (0.03)	13.22 (0.03)
5	UN	LASSO	39.14 (12.93)	39.11 (12.95)	24.53 (9.53)	14.60 (0.12)	17.22 (0.62)
(400, 100, 100)	100	BLASSO	22.43 (1.13)	12.07 (0.03)	89.31 (0.24)	15.60 (0.35)	41.44 (0.35)
SNPs		G-SMuRFS	18.58 (0.02)	12.27 (0.01)	88.69 (0.01)	16.96 (0.01)	42.21 (0.01)
		GLRR3	19.88 (0.01)	20.01 (0.03)	77.15 (0.04)	13.56 (0.00)	14.64 (0.00)
		GLRR5	17.87 (0.01)	17.98 (0.03)	77.81 (0.04)	13.65 (0.00)	14.45 (0.00)

10,000 thousand burn-ins. For the same number of samples, GLRR3 takes about 30 minutes and GLRR5 takes about 40 minutes.

3.4 The Alzheimer's Disease Neuroimaging Initiative

3.4.1 Imaging Genetic Data

Imaging genetics is an emergent trans-disciplinary research field to primarily evaluate the association between genetic variation and imaging measures as continuous phenotypes. Compared to traditional case control status, since imaging phenotypes may be closer to the underlying biological etiology of many neurodegenerative and neuropsychiatric diseases (e.g., Alzheimer), it may be easier to identify underlying genes of those diseases (Cannon and Keller, 2006; Turner et al., 2006; Scharinger et al., 2010; Paus, 2010; Peper et al., 2007; Chiang et al., 2011b,a). A challenging analytical issue of imaging genetics is that the numbers of imaging phenotypes and genetic markers can be relatively high. The aim of this data analysis is to use GLRR to specifically identify strong associations between imaging phenotypes and SNP genotypes in imaging genetic studies.

The development of GLRR is motivated by the analysis of imaging, genetic, and clinical data collected by ADNI. The ADNI is an ongoing public-private partnership to test whether genetic, structural and functional neuroimaging, and clinical data can be combined to measure the progression of mild cognitive impairment (MCI) and early Alzheimer's disease (AD). Subjects in the ADNI data have been recruited from over 50 sites across the United States and Canada. The structural brain MRI data and corresponding clinical and genetic data from baseline and follow-up were downloaded from the ADNI publicly available database (http://adni.loni.ucla.edu/). Our problem of interest is to perform genome-wide searches for establishing the association between SNPs on the top 40 AD candidate genes as listed on the AlzGene database (www.alzgene.org) as of June 10, 2010 and the brain volumes of 93 regions of interest, whose names and

abbreviation are given in the supplementary document, while accounting for other covariates, such as age and gender. By using the Bayesian GLRR, we can easily carry out formal statistical inferences, such as the identification of significant SNPs on the differences among all 93 ROI volumes.

The MRI data, collected across a variety of 1.5 Tesla MRI scanners with protocols individualized for each scanner, included standard T1-weighted images obtained using volumetric 3-dimensional sagittal MPRAGE or equivalent protocols with varying resolutions. The typical protocol included: repetition time (TR) = 2400 ms, inversion time (TI) = 1000 ms, flip angle = 8^o, and field of view (FOV) = 24 cm with a $256\times256\times170$ acquisition matrix in the $x-$, $y-$, and $z-$dimensions yielding a voxel size of $1.25\times1.26\times1.2$ mm^3. The MRI data were preprocessed by standard steps including anterior commissure and posterior commissure correction, skull-stripping, cerebellum removing, intensity inhomogeneity correction, segmentation, and registration (Shen and Davatzikos, 2004). Subsequently, we carried out automatic regional labeling by labeling the template and by transferring the labels following the deformable registration of subject images. After labeling 93 ROIs, we were able to compute volumes for each of these ROIs for each subject.

The Human 610-Quad BeadChip (Illumina, Inc., San Diego, CA) was used to genotype 818 subjects in the ADNI database, which resulted in a set of 620,901 SNP and copy number variation (CNV) markers. Since the Apolipoprotein E (APOE) SNPs, rs429358 and rs7412, are not on the Human 610-Quad Bead-Chip, they were genotyped separately. These two SNPs together define a 3 allele haplotype, namely the $\epsilon2$, $\epsilon3$, and $\epsilon4$ variants and the presence of each of these variants was available in the ADNI database for all the individuals. The software EIGENSRAIT in the package of EIGENSOFT 3.0 was used to calculate the population stratification coefficients of all subjects. To reduce

population stratification effects, we only used 761 Caucasians from all 818 subjects. We used the baseline T1 MRI scans and genetic data from all 742 Caucasians.

By following Wang et al. (2012), we selected SNPs belonging to the top 40 AD candidate genes by using quality control methods. The first line quality control steps include (i) call rate check per subject and per SNP marker, (ii) gender check, (iii) sibling pair identification, (iv) the Hardy-Weinberg equilibrium test, (v) marker removal by the minor allele frequency, and (vi) population stratification. The second line preprocessing steps include removal of SNPs with (i) more than 5% missing values, (ii) minor allele frequency smaller than 10%, and (iii) Hardy-Weinberg equilibrium $p-$value $< 10^{-6}$. This left us with 1,071 SNPs on 37 genes. We used the 1071 SNP and APOE-$\epsilon 4$ to form X, that gives $p = 1,072$.

3.4.2 Results

We fitted GLRR (3.5) with all the baseline volumes of 93 ROIs in 749 subjects as a multivariate response vector, the 1,072 selected SNPs as X matrix, and age, intracerebroventricular volume (ICV), gender, education and handedness as prognostic related covariates. To determine the rank of B, GLRR was fitted for up to $r = 10$ layers. By comparing the five different selection criteria, we chose $r = 3$ layers for the final data analysis. We ran the Gibbs sampler for $20,000$ iterations after $20,000$ burn-in iterations. Based on the MCMC samples, we calculated the posterior median and maximum absolute deviation (MAD) of the normalized U and V, and B, and then we used the standard normal approximation to calculate the $p-$values of each component of U, V, and B. The upper left panel of Figure 3.4 presents the estimated posterior median map of B, in which the elements with their $p-$values greater than 0.01 were set to zero, which reveals sparsely distributed points along the horizontal and vertical directions in the estimated B, indicating that the low-rank model would fit the ADNI data reasonably well.

We used $1.426 \times$ MAD to compute robust standard errors from the posterior median based MAD for each element of B and used a normal approximation to compute its $-\log_{10}(p)$. Specifically, we created two new matrices based on the estimated B in order to detect important ROIs and SNPs. We first applied this thresholding method to B in order to compute a new matrix B_{bin}, in which β_{jk} was set at zero if its $-\log_{10}(p)$ is less than 10, and set to 1 otherwise. Then, we calculated a 93×93 matrix $B_{bin}^T B_{bin}$ and a 1072×1072 matrix $B_{bin} B_{bin}^T$. See the upper middle and right panels of Figure 3.4. The second row of Figure 3.4 presents the $-\log_{10}(p)$ maps of B, U, and V, respectively.

We selected the top ROIs corresponding to the largest diagonal elementsof $B_{bin}^T B_{bin}$, which are listed in the first column of Table 3.3. We also picked the top ROIs based on the $-\log_{10}(p)-$values in each column of V, which are shown in the second, third, and fourth columns in Table 3.3. The locations of these ROIs are shown in Figure 3.5. Among these ROIs, the left and right sides rank close to each other, which may indicate structural brain symmetry.

We ranked the SNPs in the $B_{bin} B_{bin}^{\mathrm{T}}$ according to the sum of the columns, and in the first three columns of the U matrix by their $-\log_{10}(p)-$values. The top 20 most significant SNPs and their corresponding genes are listed in Table 3.2 under columns $B_{bin} B_{bin}^{\mathrm{T}}$, U_1, U_2, and U_3, respectively. To investigate the top SNPs and their relationship with ROI volumes in the coefficient matrix. we retained SNPs, which are correlated with at least one ROI at a significant level smaller than $10^{-6.3}$. For each SNP, we highlighted the locations of ROIs with correlation at a significant level smaller than $10^{-6.3}$, which are shown in Figure 3.6. There are different patterns of SNPs' effects on ROIs: i) rs10792821 (PICALM), rs9791189 (NEDD9), rs9376660 (LOC651924), and rs17310467 (PRNP) are significantly correlated with a small number of ROIs with relative large

coefficients; ii) rs4933497 (CH25H) and rs1927976 (DAPK1) are significantly correlated with a small number of ROIs with relative small coefficients; iii) rs1411290 (SORCS1), rs406322 (IL33), and rs1018374 (NEDD9) are significantly correlated with a large number of ROIs with medium coefficients; iv) rs1411290 (SORCS1), rs406322 (IL33) is significantly correlated with a large number of ROIs with small coefficients. Figure 3.7 shows the heatmap of coefficients among these 10 SNPs and the ROIs on the left and right hemispheres, respectively. The ROIs are chosen such that each ROI is significantly correlated to at least one of the 10 SNPs at a significance level small than $10^{-6.3}$. Most of these SNPs were not revealed in the literature of genome-wide association studies, which did not take into account the imaging phenotypes.

We were able to detect some additional SNPs, such as rs439401 (gene APOE), among others, which are not identified in existing genome-wide association studies. However, most GWA studies mainly used case-control status as the response and fitted a simple model, such as a logistic regression model. In contrast, the use of imaging measures as endophenotype may dramatically increase statistical power in detecting much more informative SNPs and genes, which deserve further investigation in Alzheimer's research.

3.5 Discussion

We have developed a Bayesian analysis GLRR to model the association between high-dimensional responses and high-dimensional covariates with an novel application in imaging genetic data. We have introduced a low rank regression model to approximate the large association matrix through the standard SVD. We have used a sparse latent factor model to more flexibly capture the complex spatial correlation structure among high-dimensional responses. We have proposed Bayesian local hypothesis testing to identify significant effects of genetic markers on imaging phenotypes, while controlling for multiple comparisons. GLRR dramatically reduces the number of parameters

Table 3.2: Ranked top SNPs based on the diagonal of $B_{bin}B_{bin}^T$ and columns of U.

$B_{bin}B_{bin}^T$		U_1		U_2		U_3	
SNP	gene	SNP	gene	SNP	gene	SNP	gene
rs9376660	LOC651924	rs9389952	LOC651924	rs439401	APOE	rs1057490	ENTPD7
rs878183	SORCS1	rs1927976	DAPK1	rs1018374	NEDD9	rs406322	IL33
rs717751	SORCS1	rs659023	PICALM	rs4713379	NEDD9	rs6441961	CCR2
rs4713379	NEDD9	rs729211	CALHM1	rs9376660	LOC651924	rs913778	DAPK1
rs1411290	SORCS1	rs6037908	PRNP	rs878183	SORCS1	rs6457200	NEDD9
rs1930057	DAPK1	rs3014554	LDLR	rs1411290	SORCS1	rs1018374	NEDD9
rs10792821	PICALM	rs11757904	NEDD9	rs717751	ORCS1	rs10422797	EXOC3L2
rs406322	IL33	rs1336269	LOC651924	rs2327389	NEDD9	rs17310467	PRNP
rs9791189	NEDD9	rs1316801	CLU	rs10884402	SORCS1	rs11193593	SORCS1
rs1018374	NEDD9	rs744970	NEDD9	rs1251753	SORCS1	rs10792821	PICALM
rs386880	IL33	rs1799898	LDLR	rs4796412	TNK1	rs9395285	CD2AP
rs17310467	PRNP	rs6133145	PRNP	rs2125071	SORCS1	rs2418960	SORCS1
rs4846048	MTHFR	rs7910584	SORCS1	rs6609709	OTC	rs10787011	SORCS1
rs10884402	SORCS1	rs1441279	SORCS1	rs4846048	MTHFR	rs7749883	LOC651924
rs3014554	LDLR	rs7067538	SORCS1	rs1360246	SORCS1	rs7025417	IL33
rs439401	APOE	rs10512188	DAPK1	rs9791189	NEDD9	rs10429166	TFAM
rs4878112	DAPK1	rs10491052	SORCS1	rs12625444	PRNP	rs1958938	DAPK1
rs10429166	TFAM	rs7929057	MS4A4E	rs10792821	PICALM	rs16871166	NEDD9
rs10787011	SORCS1	rs17475756	DAPK1	rs7918637	SORCS1	rs10948367	CD2AP
rs7095427	SORCS1	rs9496146	LOC651924	rs6088662	PRNP	rs13031703	BIN1

to be sampled and tested leading to a remarkably faster sampling scheme and efficient inference. We have shown good finite-sample performance of GLRR in both the simulation studies and ADNI data analysis. Our data analysis results have confirmed the important role of well-known genes such as APOE-$\epsilon 4$ in the pathology of ADNI, while highlighting other potential candidates that warrant further investigation.

Many issues still merit further research. First, it is interesting to incorporate common variant and rare variant genetic markers in GLRR (Bansal et al., 2010). Second, it is important to consider the joint of genetic markers and environmental factors on high-dimensional imaging phenotypes (Thomas, 2010). Third, the key features of GLRR can be adapted to more complex data structures (e.g., longitudinal, twin and family) and other parametric and semiparametric models. For instance, for longitudinal neuroimaging data, we may develop a GLRR to explicitly model the temporal association between high-dimensional responses and high-dimensional covariates, while accounting for complex temporal and spatial correction structures. Fourthly, it is important to combine different imaging phenotypes calculated from other imaging modalities, such as diffusion tensor imaging, functional magnetic resonance imaging (fMRI), and electroencephalography (EEG), in imaging genetic studies.

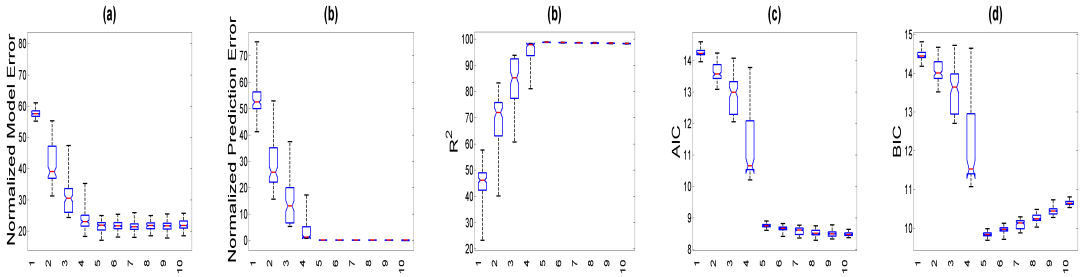

Figure 3.1: Simulation results: the box plots of five selection criteria including MEN($\hat{B}, B$), PEN($\widehat{\boldsymbol{Y}}, \boldsymbol{Y}$), $R^2(\widehat{\boldsymbol{Y}}, \boldsymbol{Y})$, AIC, and BIC against rank r from the left to the right based on 100 simulated data sets simulated from model (3.4) with $(n, p, d) = (100, 200, 100)$ and the true rank $r_0 = 5$.

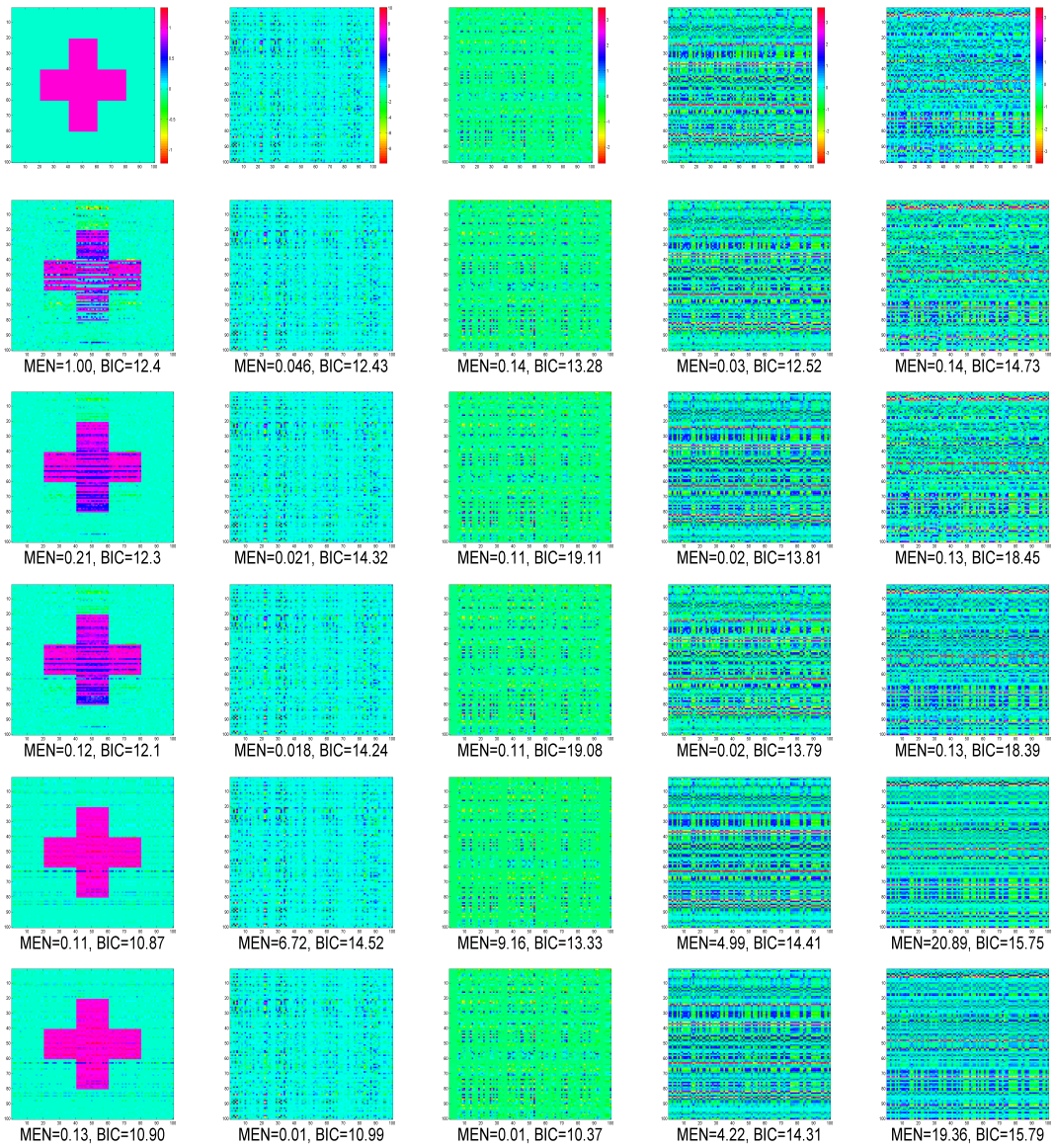

Figure 3.2: Simulation results: comparisons of true B image and estimated true B images by using LASSO, BLASSO, G-SMuRFS, GLRR3, and GLRR5 under five different scenarios. MEN$(B, \hat{B})$ and BIC were calculated for each estimated $\hat{B}$. The sample size is $n = 1000$. Columns 1-5 correspond to Cases 1-5, respectively. The true ranks of B under Cases 1-5 are, respectively, $2, 5, 5, 100$ and 100. The top row contains true B maps under Cases 1-5 and rows 2-6 correspond to the estimated $\hat{B}$ under LASSO, Bayesian LASSO, G-SMuRFS, GLRR3, and GLRR5, respectively. For simplicity, only the first 100 rows and 100 columns of B were presented. Moreover, all plots in the same column are on the same scale.

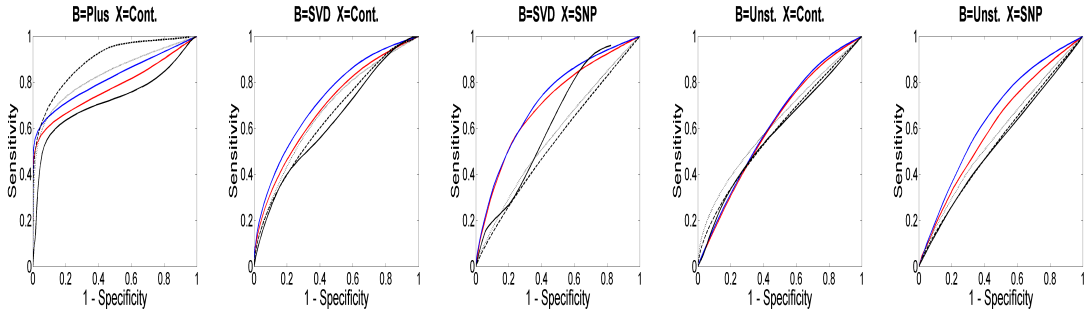

Figure 3.3: Comparisons of GLRR3, GLRR5, and LASSO under Cases 1-5: mean ROC curves based on GLRR3 (red line), GLRR5 (blue line), LASSO (black line), G-SMuRFS (dottedd line) and BLASSO (dashed line). For each case, 100 simulated data sets of size $n = 100$ each were used.

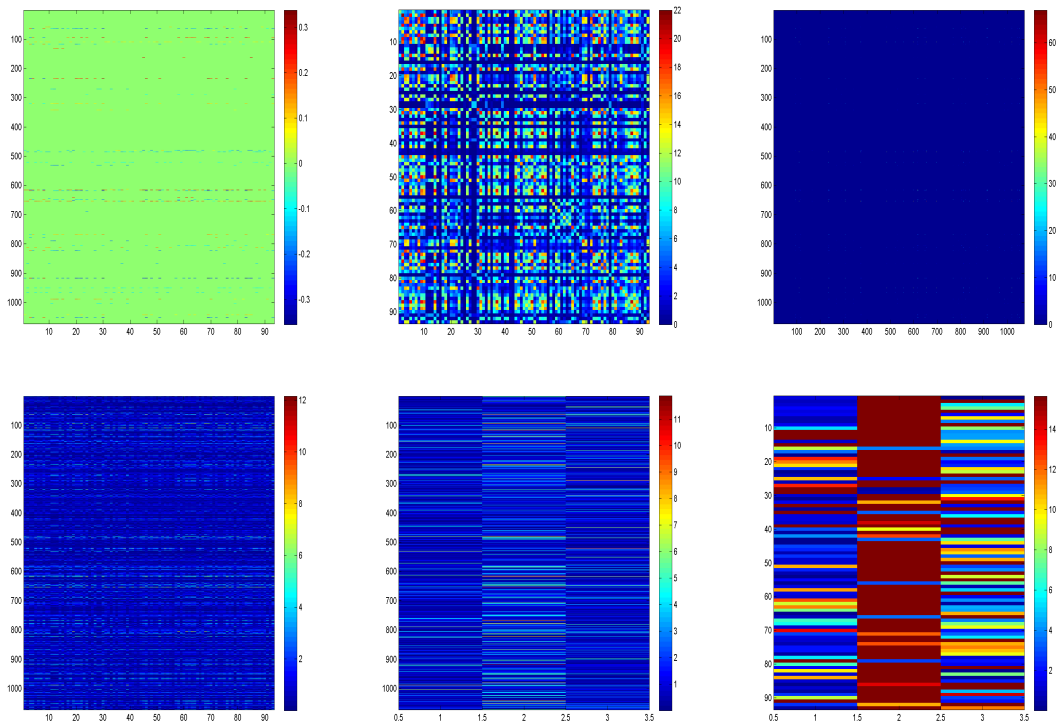

Figure 3.4: Results of ADNI data: the posterior estimate of $\hat{B}$ matrix after thresholding out elements whose $p-$ values are greater than 0.001 (left panel), $B_{bin}^T B_{bin}$ (middle panel) and $B_{bin} B_{bin}^T$ (right panel) in the first row; and the $-\log_{10} p-$ value matrices corresponding to B (left panel), U (middle panel), and V (right panel) in the second row.

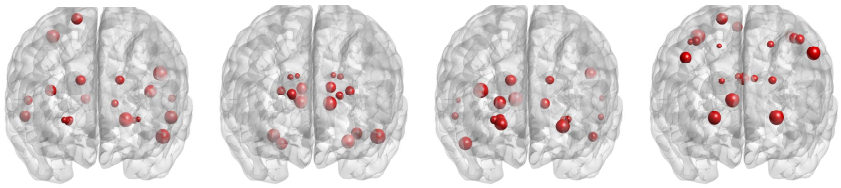

Figure 3.5: Results of ADNI data: the top 20 ROIs based on $B_{bin}^T B_{bin}$ and the first 3 columns of V. The sizes of the dots represent the rank of the ROIs.

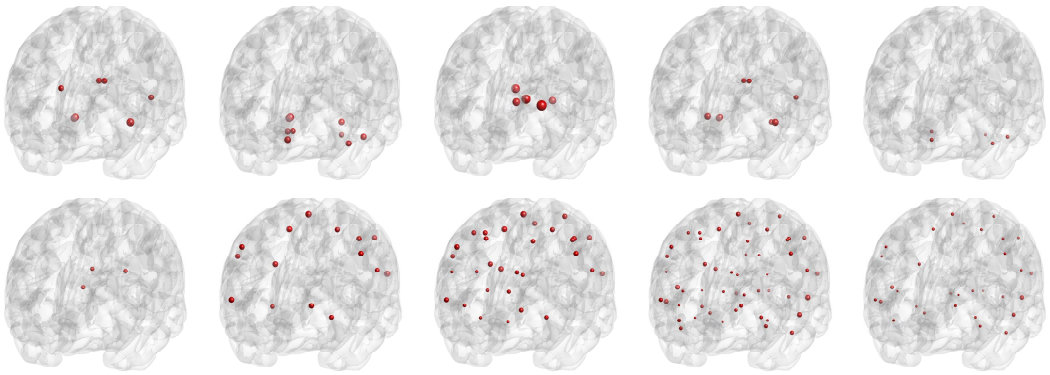

Figure 3.6: Results of ADNI data: at a -$\log_{10}(p)$ significance level greater than 6.3, the top row depicts the locations of ROIs that are correlated with SNPs rs10792821 (PICALM), rs9791189 (NEDD9), rs9376660 (LOC651924), rs17310467 (PRNP), rs4933497 (CH25H), respectively; the bottom row shows the ROIs correlated with SNPs rs1927976 (DAPK1), rs1411290 (SORCS1), rs406322 (IL33), rs1018374 (NEDD9), and rs439401 (APOE). The sizes of the dots represent the absolute magnitudes of the regression coefficients.

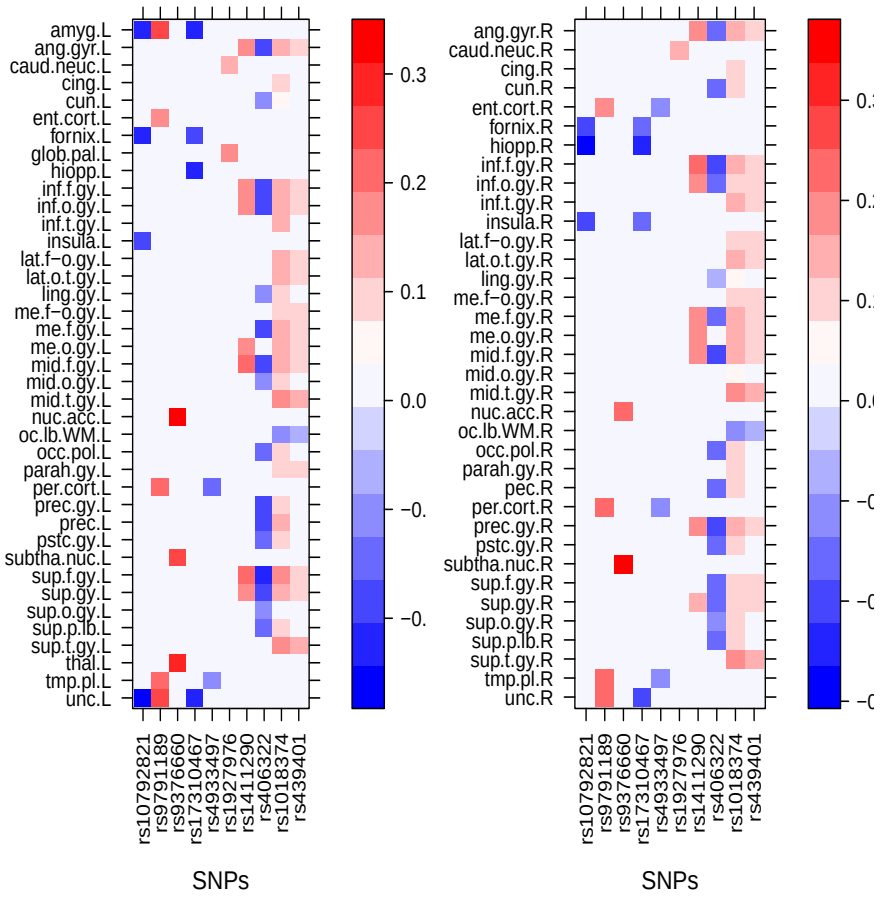

Figure 3.7: Heatmaps of coefficients between SNPs and ROIs on the left (left panel) and right (right panel) hemispheres. Coefficients with $-\log_{10}(p)$
-value smaller than 6.3 are set to 0.

Table 3.3: Ranked top ROIs based on the diagonal of $B_{bin}^T B_{bin}$ and columns of V.

$B_{bin}^T B_{bin}$	V_1	V_2	V_3
hiopp.R	caud.neuc.L	sup.t.gy.L	sup.p.lb.L
hiopp.L	caud.neuc.R	sup.t.gy.R	pstc.gy.L
amyg.R	post.limb.L	mid.t.gy.R	sup.o.gy.L
unc.L	post.limb.R	hiopp.R	prec.L
subtha.nuc.R	glob.pal.R	mid.t.gy.L	sup.p.lb.R
sup.t.gy.R	ant.caps.R	hiopp.L	pec.R
amyg.L	glob.pal.L	amyg.R	sup.o.gy.R
sup.t.gy.L	putamen.L	lat.ve.R	prec.gy.L
lat.ve.R	putamen.R	inf.t.gy.R	pstc.gy.R
nuc.acc.L	ant.caps.L	subtha.nuc.R	prec.gy.R
lat.ve.L	thal.R	amyg.L	me.f.gy.L
mid.t.gy.L	thal.L	unc.L	mid.f.gy.R
insula.L	tmp.pl.R	lat.ve.L	ang.gyr.L
sup.f.gy.L	subtha.nuc.L	inf.f.gy.R	sup.f.gy.L
insula.R	per.cort.L	lat.f-o.gy.L	fornix.L
mid.t.gy.R	tmp.pl.L	parah.gy.L	occ.pol.L
mid.f.gy.L	subtha.nuc.R	inf.t.gy.L	ang.gyr.R
unc.R	per.cort.R	parah.gy.R	cun.L
inf.t.gy.R	nuc.acc.L	nuc.acc.L	occ.pol.R
inf.f.gy.R	inf.t.gy.R	insula.L	mid.f.gy.L

CHAPTER 4

BAYESIAN LONGITUDINAL LOW RANK REGRESSION

4.1 Introduction

Many longitudinal biomedical studies, such as genomics and neuroimaging, repeatedly collect a large number of responses and covariates from a small set of subjects and focus on establishing associations among them. For instance, in imaging genetics, various imaging measures, such as volumes of regions of interest (ROIs), are repeatedly measured and may be predicted by high-dimensional covariate vectors, such as single nucleotide polymorphisms (SNPs) or gene expressions. These imaging measures can serve as important endotraits that may ultimately lead to discoveries of genes for some complex mental and neurological disorders, such as schizophrenia, since imaging data provides the most effective measures of brain structure and function (Scharinger et al., 2010; Paus, 2010; Peper et al., 2007; Chiang et al., 2011b,a). This motivates us to develop a longitudinal low rank regression model for the analysis of longitudinal high-dimensional responses and covariates.

Modeling longitudinal high-dimensional covariates and responses involve four challenges (i) a large number of regression coefficients, (ii) spatial correlation, (iii) temporal correlation, and (iv) multicollinearity among predictors. When the dimension of responses and the number of covariates, which are denoted by d and p, respectively, are even moderately high, fitting a multivariate linear model usually requires estimating a

$d \times p$ matrix of regression coefficients, whose number pd can be much larger than the sample size. At each given time, accounting for complicated spatial correlation among multiple responses is important for improving prediction accuracy of multivariate analysis (Breiman and Friedman, 1997). Accounting for temporal correlation is important for both prediction and estimation accuracy. Moreover, the collinearity among genetic predictors can cause issues of over-fitting and model misidentification (Fan and Lv, 2010).

Under the cross-sectional settings, several approaches explored new methods for high-dimensional responses and covariates. Breiman and Friedman (1997) introduced a Cards and Whey (C&W) to improve prediction error by accounting for correlations among the response variables when both p and d are moderate compared to the sample size. Peng et al. (2010) proposed a variant of the elastic net to enforce sparsity in the high-dimensional regression coefficient matrix, but they did not account for correlations among responses. Rothman et al. (2010) proposed a simultaneous estimation of a sparse coefficient matrix and a sparse covariance matrix to improve on estimation error under the L_1 penalty. Vounou et al. (2010) considered the singular value decomposition of the coefficient matrix and used the LASSO-type penalty on both the left and right singular vectors to ensure its sparse structure. They, however, do not model longitudinal data and do not provide a standard inference tool (e.g., standard error) on the nonzero components of the left and right singular vectors or the coefficient matrix.

Several attempts have been made to investigate the effect of genotypes on longitudinal phenotypes. Chen and Wang (2011) proposed penalized spline based methods for functional mixed effects models with varying coefficients, but they focus on small p and d under a low-dimensional setting. Wang et al. (2012) used sparse multitask regression to examine the association between genetic markers and longitudinal neuroimaging phenotypes. However, their multi-task regression model considered subjects

with the same number of repeated measures and ignore spatial-temporal correlations of imaging phenotypes, and thus it leads to loss of statistical power in detecting gene-imaging associations. Vounou et al. (2011) and Silver et al. (2012) proposed various sparse reduced-rank regression models by using penalized regression methods for the detection of genetic associations with longitudinal phenotypes. They, however, ignore the spatio-temporal correlations of longitudinal phenotypes, which are important for both estimation and prediction accuracy. Moreover, none of them explore the gene and time interaction, which can reveal important genetic traits altering time affects on longitudinal phenotypes.

In this paper, we propose a new Bayesian L2R2 to model the associations between a large number of predictors and multivariate longitudinal responses. A low rank regression model is introduced to characterize the low rank structure of a regression coefficient matrix between genetic variants and longitudinal imaging phenotypes, while accounting for the effects of other covariates. For the low-rank structure, we assume shrinkage priors on the singular values of the regression coefficient matrix, while not explicitly requiring orthonormality of left and right singular vectors. This facilitates fast computation of the regression coefficient matrix via the exact conditionals. We consider a penalized spline based method for delineating time-varying covariates such as age, a random effects model for capturing the within-subject temporal correlations of longitudinal responses, and a sparse factor model (Bhattacharya and Dunson, 2011) for capturing the unstructured within-subject spatial correlations of multivariate responses. We propose Bayesian local hypothesis testing to identify significant predictor effects on longitudinal responses, while controlling for multiple comparisons. Posterior computation proceeds via an efficient Markov chain Monte Carlo (MCMC) algorithm.

4.2 The Alzheimer's Disease Neuroimaging Initiative

As an emerging interdisciplinary research field, imaging genetics primarily aims to evaluate the association between genetic variates and imaging phenotypes. Since imaging phenotypes may be closer to the underlying etiology of many neuropsychiatric and neurodegenerative diseases (e.g., Alzheimer's), it may be more powerful to use imaging phenotypes for the detection of underlying genes of those diseases compared with traditional case control status (Peper et al., 2007). Due to the high dimension of imaging phenotypes and genotypes, it is analytically and computationally challenging for most statistical methods. The aim of this data analysis is to develop L2R2 in identifying associations between longitudinal phenotypes and SNP genotypes collected by the NIH ADNI, while capturing varying coefficients of time effect on longitudinal phenotypes and spatio-temporal correlations among longitudinal phenotypes.

The NIH ADNI is an ongoing public-private initiative to test whether genetic, clinical, and functional and structural neuroimaging data can be integrated to measure the progression of mild cognitive impairment (MCI) and early Alzheimer's disease (AD). ADNI initiative is recruiting study subjects over 50 sites across the United States and Canada. The genetic and clinical data along with corresponding structural brain MRI data from baseline and follow-up were obtained from the ADNI publicly available database (http://adni.loni.ucla.edu/). Our interest is to perform genome-wide searches for establishing the association between the SNPs collected on top genes reported by AlzGene (http://www.alzgene.org/) and the brain volumes of 93 regions of interest (ROIs), while accounting for other time-varying covariates, such as age, and baseline covariates, such as gender. By using L2R2 we can easily carry out formal statistical inferences, such as the identification of significant SNPs or SNPs that interact with aging on the differences among all 93 ROI volumes between AD and normal controls.

The MRI data was collected across a variety of MRI scanners with individualized protocols for each scanner to obtain standard T1-weighted images volumetric 3-dimensional sagittal MPRAGE or equivalent protocols with varying resolutions. The typical protocol included: inversion time (TI) = 1,000 ms, repetition time (TR) = 2,400 ms, flip angle = 8^o, and field of view (FOV) = 24 cm with a $256 \times 256 \times 170$ acquisition matrix in the $x-$, $y-$, and $z-$dimensions yielding a voxel size of $1.25 \times 1.26 \times 1.2$ mm^3. Standard steps including anterior commissure and posterior commissure correction, skull-stripping, cerebellum removing, intensity inhomogeneity correction, segmentation and registration (Shen and Davatzikos, 2004) were used to preprocess the MRI data. We then carried out automatic regional labeling by labeling the template and by transferring the labels following the deformable registration of subject images. After labeling 93 ROIs, we were able to compute their volumes for each subject.

To genotype subjects in the ADNI database, the Human 610-Quad BeadChip (Illumina, Inc., San Diego, CA) was used, which resulted in a set of 620,901 SNP and copy number variation (CNV) markers. Since the Apolipoprotein E (APOE) SNPs, rs429358 and rs7412, are not on the Human 610-Quad Bead-Chip, they were genotyped separately. These two SNPs together define a 3 allele haplotype, namely the $\epsilon 2$, $\epsilon 3$, and $\epsilon 4$ variants and the presence of each of these variants was available in the ADNI database for all the individuals. The software EIGENSRAIT in the package of EIGENSOFT 3.0 was used to calculate the population stratification coefficients of all subjects. To reduce population stratification effects, we only used 749 Caucasians from all 818 subjects who had at least one imaging sample available.

We also performed quality control on this initial set of genotypes (Wang et al., 2012). In order to impute the missing genotypes in our sample, we used MACH4 version 1.0.16 with default parameters to infer the haplotype phase. We also included the APOE-ϵ_4

variant, coded as the number of observed ϵ_4 variants. We dropped SNPs with more than 5% missing values and imputed the mode for the missing SNP for the remaining. In the final quality controlled genotype data, we dropped the SNPs with minor allele frequency smaller than 0.1 and Hardy-Weinberg p-value $< 10^{-6}$.

4.3 Methods

4.3.1 Model Setup

Consider n independent subjects and imaging genetic data collected at m_i time points t_{ij} for $j = 1, \ldots, m_i$ from the i-th subject for $i = 1, \ldots, n$. Let $N = \sum_{i=1}^{n} m_i$ be the total number of observations for each response. For the $i-$th subject, we observe an $m_i \times d$ matrix of imaging measures $Y_i = (y_{ik}(t_{ij}))$ for $j = 1, \ldots, m_i$ and $k = 1, \ldots, d$, a $p \times 1$ vector of time-invariant genetic predictors $\boldsymbol{x}_i = (x_{i1}, \cdots, x_{ip})^T$, and a $q_2 \times 1$ vector of time-variant prognostic factors $\boldsymbol{w}_{2,i}(t) = (w_{2,i1}(t), \cdots, w_{2,iq_2}(t))^T$ for $j_2 = 1, \ldots, q_2$. The original model of L2R2 can be written as

$$y_{ik}(t) = \boldsymbol{x}_i^{\mathrm{T}} \boldsymbol{\beta}_k + \mu_k(t) + \boldsymbol{w}_{2,i}(t)^{\mathrm{T}} \boldsymbol{\gamma}_{2,k} + \boldsymbol{z}_i(t)^{\mathrm{T}} \mathbf{b}_{ik} + \epsilon_{ik}(t) \text{ for } k = 1, \ldots, d, \tag{4.1}$$

where $\boldsymbol{\beta}_k$ is a $p \times 1$ vector of coefficients, $\mu_k(\cdot)$ is an unknown function of t, $\boldsymbol{\gamma}_{2,k}$ is a $q_2 \times 1$ vector of coefficients, $\boldsymbol{z}_i(t)$ is a $p_b \times 1$ vector of time-varying covariates measured at time point t, $\mathbf{b}_{ik}$ is a $p_b \times 1$ vector of random effects, and $\epsilon_{ik}(t)$ is an error term at time point t. It is assumed that $\mathbf{b}_i = (\mathbf{b}_{ik}) \sim N(\mathbf{0}, \Sigma_b)$ and $\boldsymbol{\epsilon}_i(t) = (\epsilon_{ik}(t)) \sim N_d(\mathbf{0}, \Sigma_e = \Theta^{-1})$, where Σ_b and Σ_e are $p_b d \times p_b d$ and $d \times d$ covariance matrices. For simplicity, it is assumed that $\Sigma_b = \mathrm{diag}(\Sigma_{b,1}, \cdots, \Sigma_{b,d})$ is a block diagonal matrix with each imaging phenotype forming a block, which captures the temporal correlations of each longitudinal phenotype, and $\Sigma_e = \Theta^{-1}$ is an unstructured precision matrix, which captures the spatial correlations among responses at each time point. Since $\boldsymbol{w}_{2,i}(t)$ may vary across t, it allows us to delineate the joint gene$\times$time effects on longitudinal phenotypes.

We first consider a low rank model for the $p \times d$ coefficient matrix $B = [\boldsymbol{\beta}_1 \cdots \boldsymbol{\beta}_d]$. Since both responses and predictors are measured from each subject, they are likely correlated with each other, and thus B may have a two-way linear association structure. This shared structure of B can be exploited to approximate a low rank model B as follows:

$$B = U\Delta V^{\mathrm{T}} = \sum_{l=1}^{r} B_l = \sum_{l=1}^{r} \delta_l \boldsymbol{u}_l {\varepsilon_l}^{\mathrm{T}}, \tag{4.2}$$

where r is the rank or number of layers of B, $B_l = \delta_l \boldsymbol{u}_l {\varepsilon_l}^{\mathrm{T}}$ is the lth layer for $l = 1, \ldots, r$, $\Delta = \mathrm{diag}(\delta_1, \ldots, \delta_r)$, $U = [\boldsymbol{u}_1, \cdots, \boldsymbol{u}_r]$ is a $p \times r$ matrix, and $V = [\varepsilon_1, \cdots, \varepsilon_r]$ is a $d \times r$ matrix. Since it is expected that only a small set of genetic variates are associated with longitudinal phenotypes, we expect that r is relatively small and V has a sparse structure. By using the low rank model, we are able to reduce the number of unknown parameters of B from pd to $(d + p)r$, which leads to a huge dimension reduction when $r << \min(p, d)$.

We consider different models for $\mu(\cdot)$. If $\mu(t)$ has a parametric form (e.g., linear or exponential), then model (4.1) reduces to a parametric random effects model. From now on, we consider a penalized spline model for $\mu(t)$ with a polynomial of degree s given by

$$\mu_k(t) = \gamma_{k,0} + \gamma_{k,1} t + \cdots + \gamma_{k,s} t^s + \sum_{m=1}^{q_1 - s - 1} \gamma_{k,s+m} (t - T_{0,m})_+^s = \boldsymbol{w}_1(t)^{\mathrm{T}} \boldsymbol{\gamma}_{1,k}, \tag{4.3}$$

where $\boldsymbol{\gamma}_{1,k} = (\gamma_{k,0}, \cdots, \gamma_{k,q_1-1})^T$, $\boldsymbol{w}_1(t) = (1, t, \cdots, t^s, (t - T_{0,1})_+^s, \cdots, (t - T_{0,q_1-1})_+^s)^T$, and $T_{0,1} \leq \cdots \leq T_{0,q_1-s-1}$ are a dense set of pre-determined knots over the range of the t_{ij}'s and are typically chosen to mimic the distribution of the t_{ij}'s, such as their $q_1 - s - 1$ tiles. Moreover, we may choose other basis functions, such as wavelet basis, to represent $\mu_k(t)$, but most methods presented below are directly applicable.

We consider a sparse factor model for $\boldsymbol{\epsilon}_i(t)$ as follows:

$$\boldsymbol{\epsilon}_i(t) = \Lambda \boldsymbol{\eta}_i(t) + \boldsymbol{\xi}_i(t), \tag{4.4}$$

where Λ is a $d \times \infty$ factor loading matrix, $\boldsymbol{\eta}_i(t) \sim N_\infty(\mathbf{0}, I_\infty)$, and $\boldsymbol{\xi}_i(t) \sim N(\mathbf{0}, \Sigma_\xi)$ with $\Sigma_\xi = \text{diag}(\sigma^2_{\xi,1}, \ldots, \sigma^2_{\xi,d})$. To achieve dimension reduction, one would typically restrict the dimension of the latent factor vector $\boldsymbol{\eta}_i(t)$ to be orders of magnitude less than that of $\boldsymbol{\epsilon}_{it}$. By following Bhattacharya and Dunson (2011), we choose a prior that shrinks the elements of Λ to zero as the column index increases. Thus, it bypasses the challenging issue of selecting the number of factors.

Our L2R2 integrates assumptions (4.1)-(4.4) and can be written in a matrix form as follows:

$$Y = XU\Delta V^{\mathrm{T}} + W\Gamma + Z\boldsymbol{b} + E, \tag{4.5}$$

where $Y = (y_{ik}(t_{ij}))$ is an $N \times d$ matrix of responses, $X = (\boldsymbol{x}_i)$ is an $N \times p$ matrix of genetic predictors with $\boldsymbol{x}_i$ repeated m_i times, $Z = (\boldsymbol{z}_i(t_{ij}))$ is an $N \times p_b$ matrix of covariates, and $E = (\boldsymbol{\epsilon}_i(t_{ij}))$ is an $N \times d$ matrix of error terms. Moreover, let $q = q_1 + q_2$, W is an $N \times q$ matrix, whose first q_1 columns consist of $\boldsymbol{w}_1(t_{ij})$ and second q_2 columns consist of $\boldsymbol{w}_{2,i}(t_{ij})$ for all i, j. Similarly, Γ is a $q \times d$ matrix and its first q_1 rows consist of $\boldsymbol{\gamma}_{1,k}$ and its second q_2 columns consist of $\boldsymbol{\gamma}_{2,k}$. In model (4.5), our primary interest is to make statistical inference on both B (or (U, Δ, V)) and Γ.

4.3.2 Priors

We consider the priors on the elements of B. Let $\text{Ga}(a, b)$ be a gamma distribution with scale a and shape b. We choose the L_2 priors on the parameters at each layer B_l

as follows:

$$\delta_l \sim N(0, \tau_\delta^{-1}),\ \tau_\delta \sim \text{Ga}(a_0, b_0),\ \boldsymbol{u}_l \sim N_p(0, p^{-1} I_p), \text{and } \varepsilon_l \sim N_d(0, d^{-1} I_d),$$

where a_0 and b_0 are pre-specified hyper-parameters. Although, we have used the same precision parameter for each element of $\boldsymbol{u}_l$, one could easily incorporate group information by choosing separate precision for each group. The number of predictors p (or d) is included in the hyperprior of $\boldsymbol{u}_l$ (or ε_l) to have a positive-definite covariance matrix of high dimensional $\boldsymbol{u}_l$ (or ε_l). Moreover, this data driven approach on priors for $\boldsymbol{u}_l$ and ε_l requires no additional hyper-parameters to choose. Since we standardize all predictors to have zero mean and unit variance, a single prior is sensible for all elements of $\boldsymbol{u}_l$. Moreover, since we rescale all responses, we use the same dispersion for all components of ε_l. Since we focus on exploiting the potential two-way correlations among the estimated coefficients, we choose the L_2 priors, which tend to borrow strength from correlated neighbors and force the coefficients towards each other to produce two highly correlated coefficients. Moreover, posterior computations are simpler and faster under the L_2 priors.

We consider the priors on the elements of $\Gamma = (\gamma_{jk})$. We also choose the L_2 prior on γ_{jk}'s as follows:

$$\gamma_{jk} \sim N(0, \tau_\gamma^{-1}) \text{ and } \tau_\gamma \sim \text{Ga}(c_0, d_0)$$

for $j = 1, \ldots, q$ and $k = 1, \ldots, d$, where c_0 and d_0 are hyperparameters. For the subject specific random coefficients we also choose independent and identically distributed normal priors as

$$b_{ik} \sim N(0, \tau_b^{-1} I_{q^*}) \text{ and } \tau_b \sim \text{Ga}(e_0, f_0)$$

for $i = 1, \ldots, n$ and $k = 1, \ldots, d$, where e_0 and f_0 are hyperparameters and q^* is the number of random effects in b_{ik}.

We consider the priors on the elements of Λ and Σ_ξ. We place the multiplicative gamma process shrinkage prior (Bhattacharya and Dunson, 2011) on Λ in order to increasingly shrink the factor loadings towards zero with the column index. Such shrinkage priors avoid the drawback of order dependence from the lower triangular constraint on Λ for identifiability. We use inverse gamma priors on the diagonal elements of Σ_ξ. Specifically, these priors are given as follows:

$$\begin{aligned}
\Lambda &= \{\lambda_{kh}\}, \;\; k = 1, \ldots, d \,; h = 1, \ldots, \infty, \\
\lambda_{kh}|\phi_{kh}, \tau_{\lambda h} &\sim N(0, \phi_{kh}^{-1}\tau_{\lambda h}^{-1}), \;\; \phi_{kh} \sim \mathrm{Ga}(v/2, v/2), \sigma_k^{-2} \sim \mathrm{Ga}(a_{\sigma k}, b_{\sigma k}), \\
\psi_1 &\sim \mathrm{Ga}(a_1, 1), \;\; \psi_g \sim \mathrm{Ga}(a_2, 1), \;\; g \geq 2, \; \tau_{\lambda h} = \prod_{l=1}^{h} \psi_l,
\end{aligned}$$

where ψ_g for $g = 1, \ldots, \infty$ are independent random variables, $\tau_{\lambda h}$ is a global shrinkage parameter for the h-th column, and the ϕ_{kh}s are local shrinkage parameters for the elements in the h-th column. Moreover, v, a_1,, a_2, $a_{\sigma k}$ and $b_{\sigma k}$ are prefixed hyperparameters. When $a_2 > 1$, the $\tau_{\lambda h}$'s increase stochastically with the column index h, which indicates more shrinkage favored over the columns of higher indices. The loading component specific prior precision $\phi_{kh}^{-1}\tau_{\lambda h}^{-1}$ allows shrinking the components of Λ.

We consider the priors on the elements of $\Sigma_{k,b}$'s. We place the independent Wishart prior on $\Sigma_{k,b}$ with $p(\Sigma_{k,b}^{-1}) \sim W[S_{b,0}, \rho_0, p_b]$ for $k = 1, \cdots, d$, where ρ_0 and the positive definite matrix $S_{b,0}$ are the given hyperparameters. Under this formulation all conditionals have closed forms and the posterior computation can be implemented via efficient Gibbs sampler.

4.3.3 Posterior Computation

The joint posterior for L2R2 with the above priors can be written as

$$
\begin{aligned}
&p(U,\Delta,V,\Gamma,\boldsymbol{b},\Theta|\text{data}) \qquad (4.6)\\
&\propto \tau_\delta^{a_o+\frac{1}{2}r-1}\tau_\gamma^{c_o+\frac{1}{2}q-1}\tau_b^{c_o+\frac{1}{2}nq^*-1}e^{-b_0\tau_\delta-d_0\tau_\gamma-f_0\tau_b}\\
&\quad \text{etr}\left(-\frac{1}{2}\left[\sum_{l=1}^{r}\left\{\tau_\delta\delta_l^2+\varepsilon_l^{\mathrm{T}}(dI_d)\varepsilon_l+\boldsymbol{u}_l^{\mathrm{T}}(pI_p)\boldsymbol{u}_l\right\}+\Gamma(\tau_\gamma I_q)\Gamma^{\mathrm{T}}+b(\tau_b I_{nq^*})b^{\mathrm{T}}\right]\right)\\
&\quad \text{etr}\left(-\frac{1}{2}(Y-XU\Delta V-W\Gamma-Z\boldsymbol{b})\Theta(Y-XU\Delta V-W\Gamma-Z\boldsymbol{b})^{\mathrm{T}}\right)
\end{aligned}
$$

We propose a straightforward Gibbs sampler for posterior computation, which converges rapidly. Starting from the initiation step, the Gibbs sampler at each iteration proceeds as follows:

1. For $l=1,\ldots,r$ update $\boldsymbol{u}_l$ from the conditional distributions

   $p(\boldsymbol{u}_l|-)\sim N_p\left(\delta_l\Sigma_{\boldsymbol{u}_l}X^{\mathrm{T}}Y_{B,l}\Theta\varepsilon_l,\Sigma_{\boldsymbol{u}_l}\right)$,

   where $\Sigma_{\boldsymbol{u}_l}=\left\{pI_p+\delta_l^2(\varepsilon_l^{\mathrm{T}}\Theta\varepsilon_l)X^{\mathrm{T}}X\right\}^{-1}$ and $Y_{B,l}=Y-X\sum_{l'\neq l}^{r}B_l-W\Gamma-Zb$.

   Update ε_l from $p(\varepsilon_l|-)\sim N_d(\delta_l\Sigma_{\varepsilon_l}\Theta Y_{B,l}^{\mathrm{T}}X\boldsymbol{u}_l,\Sigma_{\varepsilon_l})$,

   where $\Sigma_{\varepsilon_l}=\left\{dI_d+\delta_l^2(\boldsymbol{u}_l^{\mathrm{T}}X^{\mathrm{T}}X\boldsymbol{u}_l)\Theta\right\}^{-1}$.

   Update δ_l from $p(\delta_l|-)\sim N\left(\sigma_{\delta_l}^2\boldsymbol{u}_l^{\mathrm{T}}X^{\mathrm{T}}Y_{B,l}\Theta\varepsilon_l,\sigma_{\delta_l}^2\right)$,

   where $\sigma_{\delta_l}^2=\left\{\tau_\delta+(\varepsilon_l\Theta^{\mathrm{T}}\varepsilon_l)(\boldsymbol{u}_l^{\mathrm{T}}X^{\mathrm{T}}X\boldsymbol{u}_l)\right\}^{-1}$

2. Update τ_δ from $p(\tau_\delta|-)\sim \text{Ga}(a_0+0.5r,b_0+0.5\sum_{l=1}^{r}\delta_l^2)$

3. Update Γ_k from $p(\Gamma_k|-)\sim N(\Sigma_{\Gamma_k}W^{\mathrm{T}}\{y_{\Gamma,k}-(Y_{\Gamma,-k}-W\Gamma_{-k})\theta_k\},\Sigma_{\Gamma_k}^2)$, where $\Sigma_{\Gamma_k}=\{\theta_{kk}W^{\mathrm{T}}W+\gamma I_q\}^{-1}$, $y_{\Gamma,k}$ is the kth column of $Y_\Gamma=Y-XB-Zb$, $Y_{\Gamma,-k}$ is the matrix after dropping the kth column of Y_Γ, θ_{kk} is the element at kth row and kth column of Θ, θ_k is the kth column of Θ after dropping θ_{kk}, and Θ_{-kk} is the matrix after dropping kth row and kth column of Θ. This partioning was motivated

by (Khondker et al., 2013); this columnwise sampling scheme allows computionally efficient sampling with a feasible dimension of the conditinal covariance matrix.

Update τ_γ from $p(\tau_\gamma|-) \sim \text{Ga}(c_o + 0.5qd, d_0 + 0.5\sum_{j=1}^{q}\sum_{k=1}^{d}\gamma_{j'k}^2)$

4. Update b_k from $p(b_k|-) \sim N(\Sigma_{b_k}W^{\mathrm{T}}\{y_{b,k} - (Y_{b,-k} - Zb_{-k})\theta_k\}, \Sigma_{b_k}^2)$, where $\Sigma_{b_k} = (\theta_{kk}Z^{\mathrm{T}}Z + \tau_\gamma I_{nq^*})^{-1}$, $y_{b,k}$ is the kth column of $Y_b = Y - XB - W\Gamma$ and $Y_{b,-k}$ is the matrix after dropping the kth column of Y_b.

   Update τ_b from $p(\tau_b|-) \sim \text{Ga}(e_o + 0.5nq^*d, d_0 + 0.5\sum_{i=1}^{n}\sum_{j^*=1}^{q^*}\sum_{k=1}^{d} b_{ij^*k}^2)$

5. Update the kth row of Λ_{k_*}, denoted by λ_k, from its conditional distribution $p(\lambda_k|-) \sim \boldsymbol{N}((\sigma_k^{-2}\boldsymbol{\eta}^{\mathrm{T}}\boldsymbol{\eta} + D_k^{-1})^{-1}\boldsymbol{\eta}^{\mathrm{T}}\sigma_k^{-2}E_k, (\sigma_k^{-2}\eta^{\mathrm{T}}\eta + D_k^{-1})^{-1})$, where $\boldsymbol{\eta} = (\eta_1, \ldots, \eta_n)^{\mathrm{T}}$, $E_k = (\epsilon_{1k}, \ldots, \epsilon_{nk})^{\mathrm{T}}$ is the kth column of $E = Y - XB - W\Gamma - Zb$, and $D_k = \text{diag}(\phi_{k1}^{-1}\tau_{\lambda 1}^{-1}, \ldots, \phi_{kk_*}^{-1}\tau_{\lambda k_*}^{-1})$ for $k = 1, \ldots, d$. Update ϕ_{kh} from its conditional distribution $p(\phi_{kh}|-) \sim \text{Ga}(\frac{v+1}{2}, \frac{v+\lambda_{kh}^2\tau_{\lambda h}}{2})$ and σ_k^{-2}, $k = 1, \ldots, d$, from its conditional distribution $p(\sigma_k^{-2}|-) \sim \text{Ga}(a_{\sigma k} + \frac{n}{2}, b_{\sigma k} + \frac{1}{2}\sum_{i=1}^{n}(y_{ik} - \lambda_k^{\mathrm{T}}\eta_i)^2)$. Update ψ_1 from its conditional distribution $p(\psi_1|-) \sim \text{Ga}(a_1 + \frac{1}{2}dk_*, 1 + \frac{1}{2}\sum_{g=1}^{k_*}\tau_{\lambda g}^{(h)}\sum_{k=1}^{d}\phi_{kg}\lambda_{kg}^2)$,

   Update ψ_h, $h \geq 2$ from its conditional distribution $p(\psi_h|-) \sim \text{Ga}(a_2 + \frac{1}{2}d(k_* - h + 1), 1 + \frac{1}{2}\sum_{g=h}^{k^*}\tau_{\lambda g}^{(h)}\sum_{k=1}^{d}\phi_{kg}\lambda_{kg}^2)$, where $\tau_{\lambda g}^{(h)} = \prod_{t=1, t\neq h}^{g}\psi_t$ for $h = 1, \ldots, k_*$.

   Update η_i, $i = 1, \ldots, n$, from conditionally independent posteriors $p(\eta_i|-) \sim N((I_{k_*} + \Lambda^{\mathrm{T}}\Theta\Lambda)^{-1}\Lambda_{k_*}^{\mathrm{T}}\Theta\boldsymbol{\epsilon}_i, (I_{k_*} + \Lambda^{\mathrm{T}}\Theta\Lambda)^{-1})$, where $\boldsymbol{\epsilon}_i$ is the ith row of E.

4.4 Simulation Study

4.4.1 Simulation Setup

We carried out simulation studies for model (4.5) to examine the finite-sample performance of the L2R2 and its posterior computation. The simulation studies were designed

to establish the association between relatively high-dimensional longitudinal phenotypes with a set of commonly used genetic markers (e.g., SNP). Specifically, we selected all age records of the first $n = 100$ subjects ($N = 422$ records) of the 749 subjects from the ADNI imaging data. Then we standardized age, formed cubic polynomial ($s = 3$) and splines with eleven knots ($q_1 - 1 = 11$) and added standardized intracranial volume (ICV), gender, and education to form W so that $q = 15$. The first column of W is a column of 1s, the second column is age, the third column is age^2, the fourth column is age^3, columns fifth through twelve form the B-spline basis of age where the knots are based on every 10th percentile. Then we formed Z matrix with random effect z_{ih} as the standardized time in years from the baseline to visit h for the ith subject. Although, the elements of $\boldsymbol{b}$ were independently generated from $N(0, 1)$. We formed X matrix with actual SNPs in the ADNI data from the corresponding 100 subjects each repeating m_i times.

We simulated $\boldsymbol{\epsilon}_i(t) \sim N_d(0, \Sigma_e)$, where Σ_e was determined as follows. Let p_0 be a binomial probability, which controls the sparsity of the precision matrix. We first generated a $d \times d$ matrix $A = (a_{jj'})$ with $a_{jj} = 1$ and $a_{jj'} = \text{uniform}(0, 1) \times \text{binomial}(1, p_0)$ for $j \neq j'$, set $\Sigma_e = AA^{\mathrm{T}}$, and standardized Σ into a correlation matrix. The value of p_0 was tuned so that about 20% of the elements in Σ_e were zeros, yielding that the mean of the absolute correlations of Σ is about 0.40.

The low-rank coefficient matrix for SNPs B was generated with the true rank $r_0 = 5$ for two cases (i) moderately sparse case with 25% zero elements and (ii) extremely sparse case with 95% zero elements. Specifically, we set $B = U\Delta V$ with $U = (u_{jl})$, $\Delta = \text{diag}(\delta_{ll}) = \text{diag}(100, 80, 60, 40, 20)$, and $V = (v_{lk})$ being $p \times 5$, 5×5, and $5 \times d$ matrices, respectively. Moreover, we generated all elements u_{jl} and v_{lk} independently from a $N(0, 1) \times \text{binomial}(1, p_0)$ generator and then normalized the columns of U and

V to have zero mean and unit variance. The value of p_0 was tuned so that about 25% of the elements of B were zeros for the moderately sparse case and about 95% of the elements were zeros for the extremely sparse case. Each element of Γ was independently generated as $\gamma_{jk} \sim N(0,1) \times \text{binomial}(1, 0.8)$; some zero elements are required to compute specificity discussed in the next section.

For posterior computation, we chose non-informative priors for the hyperparameters and set $a_0 = b_0 = 10^{-6}$. Since shrinkage for B is achieved through dimension reduction by choosing $r << \min(d, p)$, these noninformative choices of the hyperparameters suit well. For covariance parameter Σ_e, we chose somewhat informative priors in order to impose the positive-definiteness constraint, we do not repeat the details in this paper. Similarly, for the hyperparameters of Γ and b, we chose informative priors and allowed larger shrinkage for larger dimension-to-sample size. For each simulated data set, we ran the Gibbs sampler collected $10,000$ iterations after $5,000$ burn-in iterations. We considered four cases with varying dimensions and priors below. For all cases, the true rank of B was set to $r = 5$. For each case, 100 simulated data sets were generated. The following cases were considered for simulation:

- Case 1: $p = 50, d = 50, n = 100, c_0 = d_0 = 0.5, e_0 = f_0 = 1$.
- Case 2: $p = 100, d = 100, n = 100, c_0 = d_0 = 1, e_0 = f_0 = 2$.
- Case 3: $p = 200, d = 100, n = 100, c_0 = d_0 = 2, e_0 = f_0 = 4$.
- Case 4: $p = 400, d = 100, n = 100, c_0 = d_0 = 4, e_0 = f_0 = 8$.

The results do not vary considerably for more or less informative priors. Generally, for higher dimension-to-sample size, more informative priors should produce better results. We compared our results with group-sparse multitask regression and feature selection (G-SMuRFS) (Wang et al., 2012) using a single group. Since the existing results

in (Wang et al., 2012) suggest that G-SMuRFS does comparably or outperform lasso and Bayesian lasso, we restrict our comparison to G-SMuRFS only. For G-SMuRFS, to avoid grid search, we first set both penalty parameters equal. This reduces the search for optimal parameter to one-way search, which was performed via a 5-fold cross validation from a series of values. The G-SMuRFS method does not allow for separation of the coefficient matrices like L2R2. We separated out estimated Γ and B in order to compute model performance.

4.4.2 Comparison of Results

For performance evaluation we used six different selection criteria including the Akaike information criterion (AIC), the Bayesian information criterion (BIC), the normalized prediction error (PEN), the joint multivariate R^2, the average of R^2 measures for individual responses, and the normalized model error (MEN) for B and Γ. Let $\hat{Y} = X\hat{B} + W\hat{\Gamma}$, where $\hat{B}$ is the posterior estimate of B based on the MCMC samples. Let $\text{SSE} = \text{tr}((\widehat{Y} - Y)^{\text{T}}(\widehat{Y} - Y))$ be the error sum of squares and $p_* = r(p + d) + qd$ be the combined number non-zero of parameters in B and Γ. For traditional approaches like G-SMuRFS the total number of non-zero parameters is $p_* = (p + q)d$. The six evaluation criteria are, respectively, given by

$$\begin{aligned}
&\text{AIC} = \log(\text{SSE}) + 2\tfrac{p_*}{nd}, && \text{BIC} = \log(\text{SSE}) + \tfrac{\log(nd)}{nd}p_*, \\
&\text{PEN}(\widehat{Y}, Y) = \frac{\text{SSE}}{\text{tr}(Y^{\text{T}}Y)} \times 100, && \text{Joint}R^2(\widehat{Y}, Y) = \frac{\text{tr}\big(\widehat{Y}^{\text{T}}\widehat{Y}\big)}{\text{tr}(Y^{\text{T}}Y)} \times 100, \\
&\text{MEN}(\hat{B}, B) = \frac{\text{tr}((\hat{B}-B)^{\text{T}}\Sigma_X(\hat{B}-B))}{\text{tr}(B^{\text{T}}\Sigma_X B)} \times 100, && \text{MEN}(\hat{\Gamma}, \Gamma) = \frac{\text{tr}((\hat{\Gamma}-\Gamma)^{\text{T}}\Sigma_X(\hat{\Gamma}-\Gamma))}{\text{tr}(\Gamma^{\text{T}}\Sigma_X\Gamma)} \times 100.
\end{aligned} \tag{4.7}$$

The numerator and denominator of the MEN are, respectively, the model error and measurement error of model (Yuan et al., 2007). Thus, the MEN is the ratio of the model error over the measurement error as a percentage of the total magnitude of all parameters. Similarly, the PEN and joint R^2 are defined as percentages. Normalization gives us unit-free measures, which makes comparisons more meaningful and readily

comparable across studies. We also used average R^2, which the mean of k individual coefficients of determination corresponding to each response.

In addition, we calculated the sensitivity and specificity scores for each method. Let $I(\cdot)$ be an indicator function of an event and $t_{jk} = \hat{\beta}_{jk}/s_{\beta,jk}$, where $\hat{\beta}_{jk}$ and $s_{\beta,jk}$ denote the posterior mean and standard deviation of β_{jk}, respectively. Specifically, for a given threshold T_0, sensitivity and specificity scores are, respectively, given by

$$\text{Se}(T_0) = \frac{\text{TP}(T_0)}{\text{TP}(T_0) + \text{FN}(T_0)} \text{ and } \text{Sp}(T_0) = \frac{\text{TN}(T_0)}{\text{TN}(T_0) + \text{FP}(T_0)},$$

where $\text{TP}(T_0)$, $\text{TP}(T_0)$, $\text{TP}(T_0)$, and $\text{TP}(T_0)$ are, respectively, the numbers of true positives, false positives, true negatives, and false negatives. T_0 gives different sensitivity and specificity scores, which allow us to create receiver operating characteristic (ROC) curves. In each ROC curve, sensitivity is plotted against 1-specificity. Varying sensitivity and specificity values were obtained after using a series of thresholding values from zero to the maximum value for G-SMuRFS. For L2R2, we used a varying standard deviation multiplier for thresholding to obtain varying sensitivity and specificity values. A larger area under the ROC curve indicates a better method in identifying the true positives, while controlling for the false positives.

4.4.3 Results

The simulation results in Table 4.1 show that the L2R2 does better in controlling the model error. In terms of prediction error and explanatory power (joint and average R^2) the results are similar. However, in terms of AIC L2R2 performs slightly better and much better in terms BIC as it requires fewer parameters in the model. These results indicate the advantage of borrowing strength from correlated phonetypes as well as accounting for spatiotemporal correlations among longitudinal phenotypes. We also

compared these methods in terms of ROC curves (Figure 4.1). The ROC curves reveal that L2R2 substantially outperforms in controlling false positives and false negatives as evident from larger area covered under the ROC curves. The results are comparatively better for the sparser coefficient matrix B. This is also supported by better performance of L2R2 in terms of model error. The results of L2R2 are comparatively better in terms of ROC, suggesting that a probabilistic thresholding may outperform a constant-based one. For the extremely sparse coefficient matrix, which may be of much lower rank than the dimension, the use of L2R2 may be a better choice. Figure 4.2 plotted selected spline functions from true coefficients (Γ) and estimated coefficients from both models. L2R2 more closely estimates the underlying spline functions in both settings of B matrix. Figure 4.3 plotted true coefficients matrix (B) and estimated coefficients from both models. L2R2 more closely estimates the true coefficient matrix and picks up less noise under both settings of B matrix.

4.5 Application to ADNI Data

In ADNI database, we included all 749 Caucasian subjects subjects with at least one non-missing structural MRI measures giving an unbalanced data set with $n = 749$ subjects and $N = 2817$ MRI measures. Among them, 41 subjects have only one observation and another 67 subjects have only two observations. It leads us to consider a single random effect, since adding more than one random effect will entail heavy penalty, especially for those subjects with single observation. Moreover, it is expected that longitudinal phenotypes of the same individual usually exhibit positive correlation and the strength of the correlation decreases with the time separation, we proposed to use random time coefficient to account for the temporal correlation. For the age effect, we used a penalized spline of third degree with 11 knots based on the percentiles of standardized age. We also included ICV, gender, education, and handedness as covariates in W.

Table 4.1: Empirical comparison of L2R2 and G-SMuRFS under Cases 1-4 based on the six selection criteria for moderate and extreme sparsity of B. The means and standard deviations of these criteria are also calculated and their standard deviations are presented in parentheses.

Case	Sparsity	Method	MEN(B)	MEN(Γ)	PEN	Joint R^2	Average R^2	AIC	BIC
$p = 50$	Moderate	L2R2	8.21 (2.71)	3.68 (1.20)	2.30 (0.07)	95.29 (0.10)	91.45 (0.17)	9.90 (0.03)	10.37 (0.03)
$d = 50$		G-SMuRFS	25.40 (5.19)	11.12 (2.34)	4.89 (0.07)	95.00 (0.07)	90.35 (0.12)	10.73 (0.02)	11.67 (0.02)
	Extreme	L2R2	1.53 (0.51)	1.79 (0.61)	1.69 (0.07)	96.46 (0.09)	91.10 (0.24)	9.90 (0.04)	10.37 (0.04)
		G-SMuRFS	9.02 (1.83)	11.14 (2.32)	3.54 (0.05)	96.33 (0.05)	89.76 (0.14)	10.73 (0.02)	11.67 (0.02)
$p = 100$	Moderate	L2R2	11.44 (6.83)	4.27 (2.59)	2.38 (0.07)	94.96 (0.10)	90.87 (0.17)	10.58 (0.03)	11.09 (0.03)
$d = 100$		G-SMuRFS	32.46 (2.78)	11.04 (1.04)	4.53 (0.07)	95.09 (0.07)	90.51 (0.14)	11.50 (0.02)	13.55 (0.02)
	Extreme	L2R2	1.02 (0.39)	2.22 (1.09)	1.16 (0.07)	97.55 (0.07)	90.72 (0.22)	10.70 (0.06)	11.21 (0.06)
		G-SMuRFS	4.36 (0.37)	11.06 (1.04)	1.86 (0.03)	97.93 (0.03)	90.55 (0.14)	11.50 (0.02)	13.55 (0.02)
$p = 200$	Moderate	L2R2	4.52 (4.45)	3.15 (3.08)	2.05 (0.07)	95.60 (0.11)	91.77 (0.20)	10.60 (0.04)	11.21 (0.04)
$d = 100$		G-SMuRFS	23.48 (1.15)	13.54 (0.77)	2.40 (0.05)	96.66 (0.06)	93.23 (0.11)	11.49 (0.02)	15.59 (0.02)
	Extreme	L2R2	1.41 (2.19)	3.02 (2.34)	1.20 (1.40)	97.90 (1.57)	90.49 (0.82)	11.11 (0.45)	11.73 (0.45)
		G-SMuRFS	3.19 (0.16)	13.72 (0.78)	0.67 (0.01)	99.03 (0.02)	92.77 (0.12)	11.50 (0.02)	15.59 (0.02)
$p = 400$	Moderate	L2R2	4.33 (4.33)	6.94 (6.94)	1.35 (1.35)	97.01 (97.01)	92.63 (92.63)	10.65 (10.65)	11.47 (11.47)
$d = 100$		G-SMuRFS	11.32 (11.32)	14.08 (14.08)	1.11 (1.11)	98.31 (98.31)	95.25 (95.25)	12.12 (12.12)	20.32 (20.32)
	Extreme	L2R2	1.23 (2.70)	6.21 (10.85)	0.95 (1.71)	98.52 (2.11)	90.77 (0.98)	11.65 (0.87)	12.47 (0.87)
		G-SMuRFS	1.96 (0.10)	16.16 (1.00)	0.27 (0.01)	99.56 (0.01)	94.53 (0.11)	12.12 (0.03)	20.32 (0.03)

We consider two sets of top genes, which may be associated with AD. First, we chose the top 10 Genes listed in the AlzGene (www.alzgene.org) database and found 114 SNPs on those genes from the ADNI database. After the quality control (removal of SNPs with more than 5% missing, minor allele frequency larger than 10%, and Hardy-Weinberg equilibrium testing), 87 SNPs, APOE-ϵ4 and their interaction with age were included in model (4.1). Second, we chose the top 40 genes with 1,224 SNPs used by Wang et al. (2012). After the standard quality control, we were able to get 1,072 SNP and their interactions with age. Fig. 4.5(a) presents the map of linkage disequilibrium (LD) among the 1,071 selected SNPs, in which the first one is APOE-ϵ4 and the next 87 are from the top 10 AlzGene genes. Inspecting Figure 4.5 (a) reveals a clear clustering pattern of SNPs by gene. Specifically, SNPs within a gene have large LD correlations, whereas SNPs between different genes have almost zero LD correlations.

After determining X, Z, and W, we fitted L2R2 model (4.1) to ADNI data as follows. To determine the rank of B, L2R2 was run for up to $r = 10$ layers. By comparing the five different selection criteria, we chose $r = 3$ layers as the optimal rank for the final data analysis. We ran the Gibbs sampler for $20,000$ iterations after $20,000$ burn-in iterations. For G-SMuRFS, we used the same Y matrix, combined W and X matrices into a single predictor matrix, and then used the 5-fold cross validation to choose the optimal penalty.

4.5.1 Longitudinal Age Effect

Based on the MCMC samples, we calculated the fitted spline functions of standardized ROIs. See Figure 4.4 for details. Some selected ROIs, which tend to decline in volume with age, include left and middle temporal gyri, left and superior temporal gyri, and left and right amygdala, among others. There are other ROIs, including mostly hollow areas and white matters, that increase in volume. These rising volume regions

include left and right lateral ventricles, left and right frontal lobe white matter, and left and right temporal lobe, among others. Not surprisingly, the trends in most regions show structural symmetry.

4.5.2 Regions of Interest

Based on the MCMC samples, we calculated the posterior median and median absolute deviations (MAD) of U and V, and B, and then we used the standard normal approximation to calculate the p-values of each component of U, V, and B. We used $1.426 \times \text{MAD}$ to compute robust standard errors from the posterior median based MAD for each element of B; then used normal approximation to compute $p-$values for thresholding B. Specifically, we created two new matrices based on the estimated B in order to detect important ROIs and SNPs. We first applied this thresholding method to B in order to compute a new matrix B_{bin}, in which β_{jk} was set at zero if its negative $\log_{10}(p)$ is less than 6, and set to 1 otherwise. Figure 4.5 (d) presents the estimated posterior median map of B, in which the elements with their negative $\log_{10}(p)$ values less than 6 were set to zero. Inspecting Figure 4.5 (d) reveals sparsely distributed points along the horizontal and vertical directions in the estimated B, which indicates that the low-rank model would fit the ADNI data reasonably well. Then, we calculated a 93×93 matrix $B_{bin}^T B_{bin}$ and a 176×176 matrix $B_{bin} B_{bin}^T$ and presented them in the first row of Figure 4.5. Similar approach was taken for U and V.

Since at the presence of interaction inference on main affect is misleading and our primary focus is on change of SNP's effect over age, for netwrok building we only used interaction parts of B, U, and V. We adopted several approaches to select the top ROIs. First, we built column sums of $B_{bin}^T B_{bin}$ based on the age by SNP interaction coefficients; then ranked the top ROIs based on the column sums of $B_{bin}^T B_{bin}$ giving us the ROIs with maximum number of significant coefficients. Next, we calculated p-values for each layer

of V similar to the p-value calculation of B, then computed negative vlues of $log_{10}(p)$ at each layer V; then ranked the ROIs based on the $log_{10}(p)$. Finally, we ranked the ROIs based on the sum of the abolute values of sparse B. The results are reported in table 4.2 and 4.3; gyris are apparently predominantly affected by SNP-age interaction.

4.5.3 SNPs

Using the interaction parts of B, U, and V we and adopted several approaches to select the top SNPs. First, we built column sums of $B_{bin}B_{bin}^{\mathrm{T}}$ based on the age by SNP interaction coefficients; then ranked the top SNPs based on the column sums of $B_{bin}B_{bin}^{\mathrm{T}}$ giving us the ROIs with maximum number of significant coefficients. Next, we calculated p-values for each layer of U similar to the p-value calculation of B, then computed negative vlues of $log_{10}(p)$ at each layer U; then ranked the ROIs based on the $log_{10}(p)$. Finally, we ranked the ROIs based on the sum of the abolute values of sparse B. The results are reported in table 4.2 and 4.3. APOE-ϵ4 is among the top 20 SNPs from the model with SNPs from top 10 genes.

Among the top SNPs rs880436(BIN1) has positive age interaction with perirhinal cortex left, perirhinal cortex right, uncus left, temporal pole right, amygdala left, uncus right, amygdala right, temporal pole left, hippocampal formation left, hippocampal formation right, entorhinal cortex left, entorhinal cortex right, inferior temporal gyrus right, parahippocampal gyrus right, middle temporal gyrus right, middle temporal gyrus left, superior temporal gyrus left, inferior temporal gyrus left, and parahippocampal gyrus left. SNP rs3752237(ABCA7) has positive age interaction with lateral ventricle right, and lateral ventricle left. It has negative age interaction with inferior temporal gyrus left, lateral occipitotemporal gyrus right, insula right, amygdala left, uncus right, inferior temporal gyrus right, superior temporal gyrus left, superior temporal gyrus right, middle temporal gyrus left, hippocampal formation right, hippocampal formation left,

middle temporal gyrus right, and amygdala right. SNP rs3752240(ABCA7) has negative age interaction with inferior temporal gyrus left, inferior temporal gyrus right, superior temporal gyrus right, superior temporal gyrus left, middle temporal gyrus left, middle temporal gyrus right, hippocampal formation right, hippocampal formation left, and amygdala right. SNP rs33978622(CD33) has positive age interaction with amygdala right, hippocampal formation left, hippocampal formation right, middle temporal gyrus left, superior temporal gyrus right, and superior temporal gyrus left. SNP rs10501608(PICAL) has negative age interaction with , superior temporal gyrus right, lateral occipitotemporal gyrus left, middle temporal gyrus right, superior temporal gyrus left, lateral occipitotemporal gyrus right, and middle temporal gyrus left.

4.6 Discussion

We have developed a Bayesian analysis L2R2 to model the association between repeatedly measured high-dimensional responses and high-dimensional covariates with a novel application in imaging genetic data. We have introduced a low rank regression model to approximate the large association matrix through the standard SVD. We combined a sparse latent factor model and random effects to more flexibly capture the complex spatiotemporal correlation structure. We have incorporated splines to capture the effect of aging and combined traditional coefficient estimation with low rank approach. L2R2 dramatically reduces the number of parameters to be sampled and tested leading to a remarkably faster sampling scheme and efficient inference. We have shown good finite-sample performance of L2R2 in both the simulation studies and ADNI data analysis. Our data analysis results have confirmed the important role of well-known genes such as APOE-ϵ4 in the pathology of ADNI, while highlighting other potential candidates that warrant further investigation.

Many issues still merit further research. First, it is important to consider the joint

of genetic markers and environmental factors on high-dimensional imaging phenotypes (Thomas, 2010). Second, it will be interesting to incorporate common variant and rare variant genetic markers in L2R2 (Bansal et al., 2010). Third, the key features of GLRR can be adapted to more complex data structures (e.g., twin and family) and other parametric and semiparametric models. Fourth, the method can be extended to combine different imaging phenotypes calculated from other imaging modalities, such as diffusion tensor imaging, functional magnetic resonance imaging (fMRI), and electroencephalography (EEG), in imaging genetic studies. Fifth, one could incorporate group structure among SNPs, as apparent gene-based grouping from the LD correlation plots, by choosing group priors for U.

Table 4.2: Top ROIs based on $B_{bin}^{\mathrm{T}} B_{bin}$, p-values of U, and magnitude of coefficients for model using SNPs from top 10 genes.

From $B_{bin}^{\mathrm{T}} B_{bin}$	From p-values of U	From Magnitude of Coefficients
hippocampal formation right	caudate nucleus right	middle temporal gyrus left
middle temporal gyrus left	angular gyrus right	hippocampal formation right
hippocampal formation left	temporal lobe WM right	hippocampal formation left
amygdala right	postcentral gyrus left	superior temporal gyrus right
superior temporal gyrus left	lateral occipitotemporal gyrus left	amygdala right
middle temporal gyrus right	superior parietal lobule left	superior temporal gyrus left
uncus right	inferior occipital gyrus right	middle temporal gyrus right
inferior temporal gyrus left	nucleus accumbens left	perirhinal cortex right
perirhinal cortex right	supramarginal gyrus right	perirhinal cortex left
amygdala left	lateral occipitotemporal gyrus right	temporal pole left
inferior temporal gyrus right	precuneus right	inferior temporal gyrus left
perirhinal cortex left	amygdala left	uncus right
temporal pole left	frontal lobe WM left	uncus left
entorhinal cortex left	middle frontal gyrus right	inferior temporal gyrus right
uncus left	corpus callosum	entorhinal cortex left
superior temporal gyrus right	inferior temporal gyrus right	temporal pole right
temporal pole right	middle occipital gyrus left	amygdala left
entorhinal cortex right	middle temporal gyrus left	lateral occipitotemporal gyrus right
lateral occipitotemporal gyrus right	supramarginal gyrus left	lateral occipitotemporal gyrus left
parahippocampal gyrus left	medial occipitotemporal gyrus left	entorhinal cortex right

Table 4.3: Top ROIs based on $B_{bin}B_{bin}^{\mathrm{T}}$, p-values of U, and magnitude of coefficients for model using SNPs from top 45 genes.

From $B_{bin}B_{bin}^{\mathrm{T}}$	From p-values of U	From Magnitude of Coefficients
medial front-orbital gyrus right	occipital lobe WM left	supramarginal gyrus right
superior frontal gyrus right	parietal lobe WM left	superior temporal gyrus right
globus palladus right	hippocampal formation right	middle temporal gyrus left
globus palladus left	lateral occipitotemporal gyrus right	precentral gyrus left
putamen left	parahippocampal gyrus left	inferior occipital gyrus left
inferior frontal gyrus left	middle occipital gyrus right	middle frontal gyrus left
putamen right	temporal pole left	superior temporal gyrus left
frontal lobe WM right	uncus left	superior frontal gyrus left
parahippocampal gyrus left	perirhinal cortex left	supramarginal gyrus left
angular gyrus right	superior frontal gyrus right	cingulate region left
middle frontal gyrus right	corpus callosum	angular gyrus right
subthalamic nucleus right	superior parietal lobule right	middle frontal gyrus right
nucleus accumbens right	cuneus left	inferior frontal gyrus left
uncus right	1e20erior temporal gyrus right	inferior frontal gyrus right
cingulate region left	insula right	lateral occipitotemporal gyrus right
fornix left	putamen left	postcentral gyrus right
frontal lobe WM left	uncus right	precentral gyrus right
precuneus right	precuneus right	postcentral gyrus left
subthalamic nucleus left	angular gyrus right	lateral front-orbital gyrus left
posterior limb of internal capsule left	lateral occipitotemporal gyrus left	cingulate region right

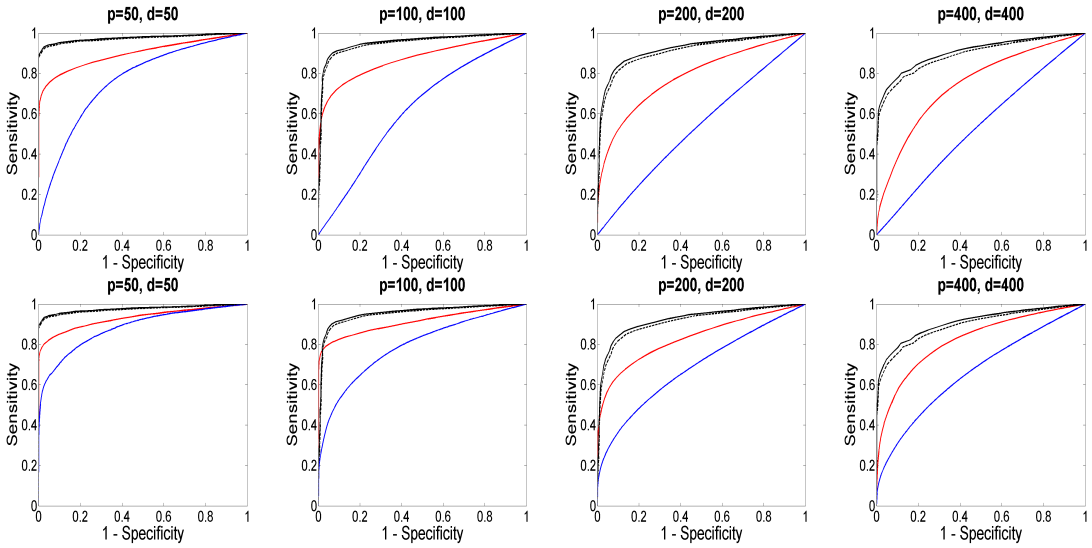

Figure 4.1: Simulation results: Mean ROC curves from L2R2 (red line for B, black line for Γ), and G-SMuRFS (blue line for B, black dashed line for Γ) based on 100 samples of size $n = 100$ each. Top row for moderately sparse B and bottom row for extremely sparse B, while Γ remains the same in both scinarios.

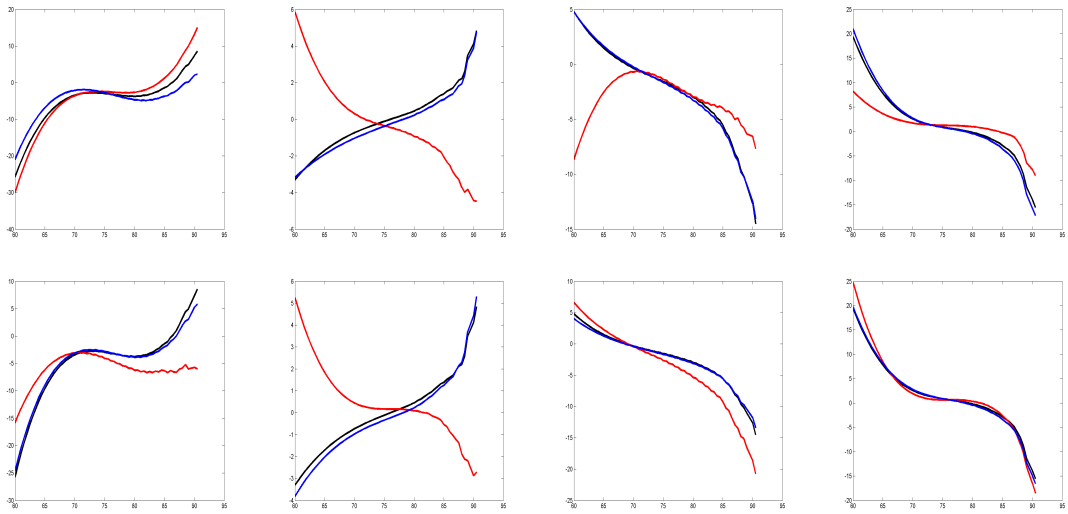

Figure 4.2: Simulation results: Splines for standardized volumes of selected ROIs (from left to right, respectively, ROIs 1, 4, 7 and 8) from single sample. Black lines are generated by true G, red lines by estimates from G-SMuRFS, and blue by estimates from LGLRR. Top row is based on G when B is moderately sparse and bottom row is based on G when B is extremely sparse. L2R2 did a decent job in estimating the true splines while G-SMuRFS can be off for some ROIs.

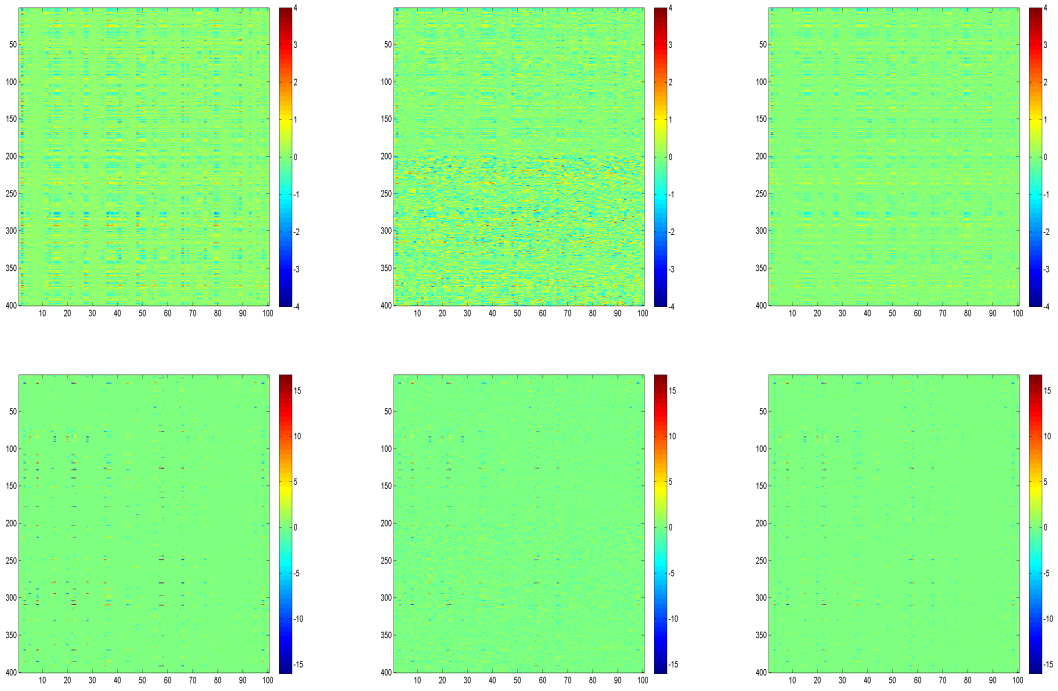

Figure 4.3: Simulation results: Image plots of the low-rank component B from single sample. True B on the left, G-SMuRFS in the middle, and L2R2 on the right. Top row is moderately sparse B and bottom row is extremely sparse B. For moderatly sparse B G-SMuRFS may pick up too much noise.

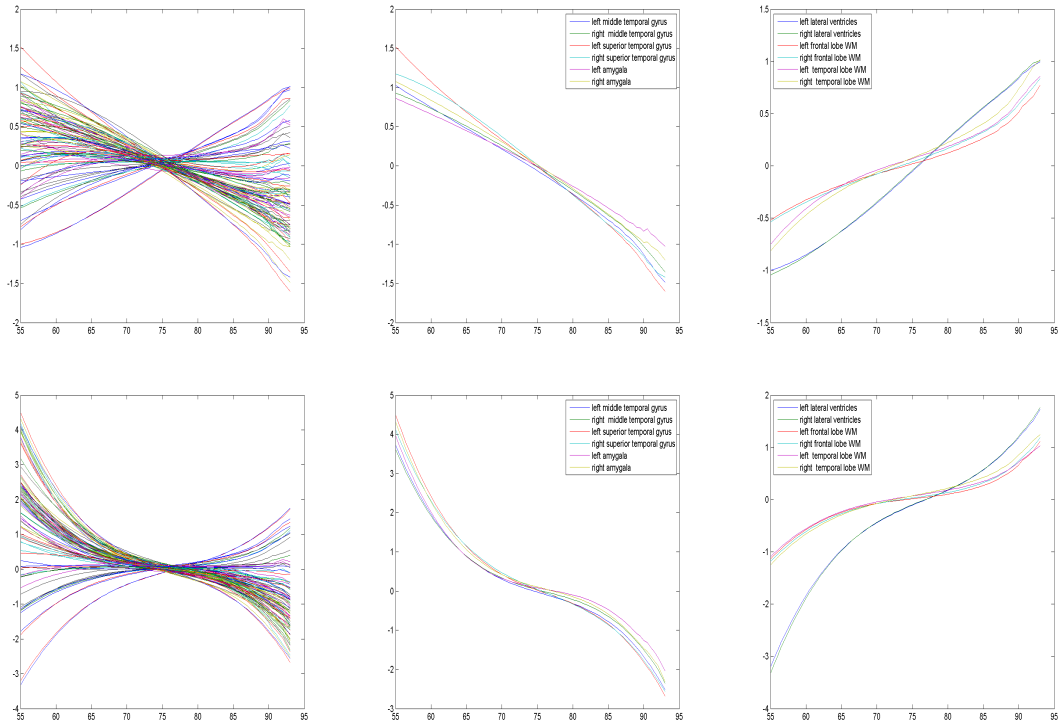

Figure 4.4: Splines functions: all the ROIs on the left, selected ROIs with declining volumes in the middle, selected ROIs with increasing volumes on the right. Top row from the model using SNPs from top 10 genes, bottom row from the model using SNPs from top 45 genes.

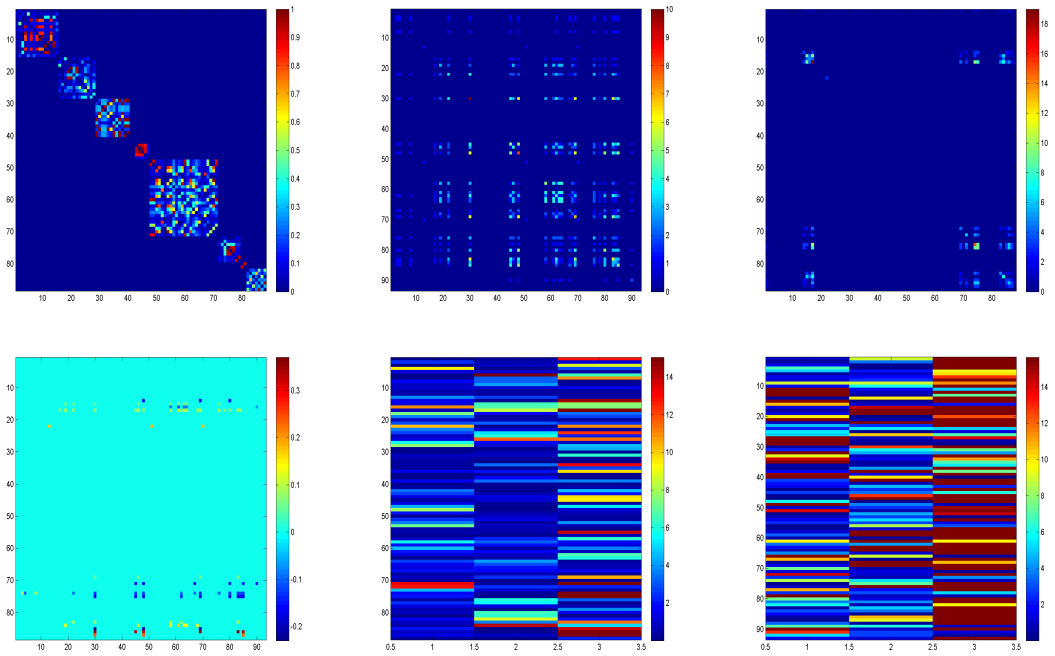

Figure 4.5: Data analysis results from SNPs in the top 10 genes: Top panel (a) left- LD correlation of selected SNPs from top 10 genes in AlzGene database (b) middle- ROI network from binary B (c) right- SNP network from binary B. Bottom panel (d) left- age by SNP interaction part of sparse B after thresholding with negative $\log_{10}(p) > 10$, (e) middle- negative $log_{10}(p)$ of U (f) right- negative $\log_{10}(p)$ of V.

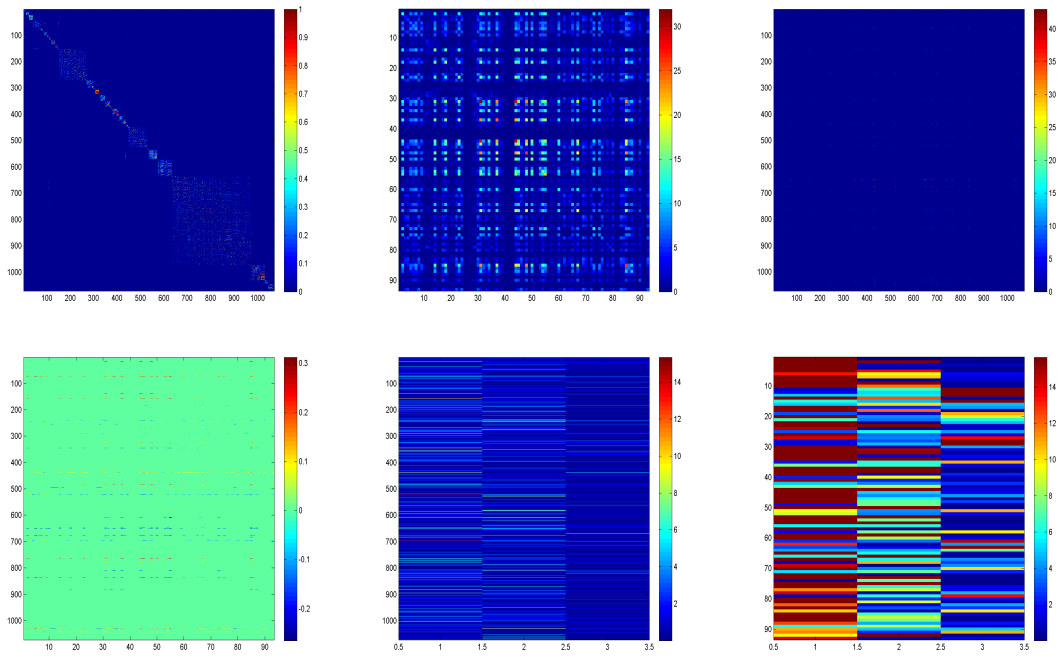

Figure 4.6: Data analysis results from SNPs in the top 45 genes: Top panel (a) left- LD correlation of selected SNPs from top 45 genes in AlzGene database (b) middle- ROI network from binary B (c) right- SNP network from binary B. Bottom panel (d) left- age by SNP interaction part of sparse B after thresholding with negative $log_{10}(p) > 10$, (e) middle- negative $log_{10}(p)$ of U (f) right- negative $log_{10}(p)$ of V.

Table 4.4: Top SNPs based on $B_{bin}B'_{bin}$, p-values of U, and magnitude of coefficients for model using SNPs from top 10 genes.

From $B_{bin}B'_{bin}$	From p-values of U	From Magnitude of Coefficients
rs880436(BIN1)	rs880436(BIN1)	rs33978622(CD33)
rs3752237(ABCA7)	rs1354106(CD33)	rs3752237(ABCA7)
rs3752240(ABCA7)	rs3752237(ABCA7)	rs880436(BIN1)
rs1408077(CR1)	rs3752240(ABCA7)	rs3752240(ABCA7)
rs33978622(CD33)	rs33978622(CD33)	rs3865444(CD33)
rs3826656(CD33)	rs10501608(PICAL)	rs10501608(PICAL)
rs677909(PICAL)	rs12734030(CR1)	rs1354106(CD33)
rs10501608(PICAL)	rs10501604(PICAL)	rs988337(CD33)
rs3865444(CD33)	rs988337(CD33)	rs10194375(BIN1)
rs988337(CD33)	rs6458573(CD2AP)	rs1408077(CR1)
rs1354106(CD33)	APOE34(APOE)	rs12734030(CR1)
rs12734030(CR1)	rs10200967(BIN1)	rs3826656(CD33)
rs10779339(CR1)	rs3826656(CD33)	rs10779339(CR1)
APOE34(APOE)	rs6709337(BIN1)	rs677909(PICAL)
rs10194375(BIN1)	rs10194375(BIN1)	APOE34(APOE)
rs1571344(CR1)	rs650877(CR1)	rs1571344(CR1)
rs2025935(CR1)	rs677909(PICAL)	rs2025935(CR1)
rs4310446(CR1)	rs9395285(CD2AP)	rs4310446(CR1)
rs11117959(CR1)	rs610932(MS4A6)	rs11117959(CR1)
rs10127904(CR1)	rs662196(MS4A6)	rs10127904(CR1)

Table 4.5: Top SNPs based on $B_{bin}B'_{bin}$, p-values of U, and magnitude of coefficients for model using SNPs from top 45 genes.

From $B_{bin}B'_{bin}$	From p-values of U	From Magnitude of Coefficients
rs3752237(ABCA7)	rs1057490(ENTPD)	rs7905923(SORCS)
rs472664(SORCS)	rs6088662(PRNP)	rs1997660(PGBD1)
rs1997660(PGBD1)	rs17177040(SORCS)	rs1358024(TF)
rs7905923(SORCS)	rs4513489(CCR2)	rs2900712(SORCS)
rs2239942(GAPDH)	rs1473180(DAPK1)	rs3752237(ABCA7)
rs2327389(NEDD9)	rs10787011(SORCS)	rs472664(SORCS)
rs2900712(SORCS)	rs1336269(LOC65)	rs1057490(ENTPD)
rs6608762(OTC)	rs1358024(TF)	rs2239942(GAPDH)
rs1473180(DAPK1)	rs6608762(OTC)	rs1330001(SORCS)
rs1358024(TF)	rs6441961(CCR2)	rs6608762(OTC)
rs1330001(SORCS)	rs6584307(ENTPD)	rs6088662(PRNP)
rs1336269(LOC65)	rs2273684(PRNP)	rs2327389(NEDD9)
rs7870463(DAPK1)	rs583791(MS4A6)	rs1336269(LOC65)
rs4935775(SORL1)	rs1360246(SORCS)	rs17177040(SORCS)
rs10787011(SORCS)	rs10779339(CR1)	rs4513489(CCR2)
rs17602572(MS4A6)	rs1699105(SORL1)	rs4935775(SORL1)
rs1057490(ENTPD)	rs597668(EXOC3)	rs1473180(DAPK1)
rs11194016(SORCS)	rs4309(ACE)	rs10787011(SORCS)
rs6088662(PRNP)	rs11193377(SORCS)	rs7870463(DAPK1)
rs10779339(CR1)	rs17496723(NEDD9)	rs11117959(CR1)

CHAPTER 5

CONCLUSION

First, we focused on penalized covariance estimation and 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 Θ, 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 on prevision matrix, the method works well for non-full rank data, and performs shrinkage and estimation simultaneously.

Second, we developed a Bayesian GLRR to model the association between high-dimensional responses and high-dimensional covariates with a novel application in imaging genetic data. We have introduced a low rank regression model to approximate the large association matrix through the standard SVD. We have used a sparse latent factor model to more flexibly capture the complex spatial correlation structure among high-dimensional responses. We have proposed Bayesian local hypothesis testing to identify

significant effects of genetic markers on imaging phenotypes, while controlling for multiple comparisons. GLRR dramatically reduces the number of parameters to be sampled and tested leading to a remarkably faster sampling scheme and efficient inference. We have shown good finite-sample performance of GLRR in both the simulation studies and ADNI data analysis. Our data analysis results have confirmed the important role of well-known genes such as APOE-ϵ4 in the pathology of ADNI, while highlighting other potential candidates that warrant further investigation.

Finally, we developed a Bayesian analysis L2R2 to model the association between repeatedly measured high-dimensional responses and high-dimensional covariates with a novel application in imaging genetic data. We have introduced a low rank regression model to approximate the large association matrix through the standard SVD. We combined a sparse latent factor model and random effects to more flexibly capture the complex spatiotemporal correlation structure. We have incorporated splines to capture the effect of aging and combined traditional coefficient estimation with low rank approach. L2R2 dramatically reduces the number of parameters to be sampled and tested leading to a remarkably faster sampling scheme and efficient inference. We have shown good finite-sample performance of L2R2 in both the simulation studies and ADNI data analysis. Our data analysis results have confirmed the important role of well-known genes such as APOE-ϵ4 in the pathology of ADNI, while highlighting other potential candidates that warrant further investigation.

APPENDIX: DERIVATION OF CONDITIONALS

Full Conditional of LGLRR

Conditionals for Δ: from equation (4.6) we can write,

$$\begin{aligned} p(\delta_l|-) &\propto \text{etr}\left(-\frac{1}{2}\tau_\delta\delta_l^2 + (Y_{A,l} - \delta_l X\boldsymbol{u}_l\varepsilon_l)\Theta(Y_{A,l} - \delta_l X\boldsymbol{u}_l\varepsilon_l)^{\mathrm{T}}\right) \\ &\propto \text{etr}\left(\delta_l\boldsymbol{u}_l^{\mathrm{T}}X^{\mathrm{T}}Y_{A,l}\Theta\varepsilon_l - \frac{1}{2}\delta_l^2(\tau_\delta + \varepsilon_l\Theta^{\mathrm{T}}\varepsilon_l)(\boldsymbol{u}_l^{\mathrm{T}}X^{\mathrm{T}}X\boldsymbol{u}_l)\right) \\ p(\tau_\delta|-) &\propto \tau_\delta^{a_o+\frac{1}{2}r-1}\text{etr}\left(-\frac{1}{2}(b_0 + \sum_{l=1}^{r}\tau_\delta\delta_l^2)\right). \end{aligned}$$

This implies $p(\delta_l|-) \sim N(\sigma_{\delta_l}^2\boldsymbol{u}_l^{\mathrm{T}}X^{\mathrm{T}}Y_{A,l}\Theta\varepsilon_l, \sigma_{\delta_l}^2)$,

with $p(\tau_\delta|-) \sim \text{Ga}(a_0 + \frac{1}{2}r, b_0 + \frac{1}{2}\sum_{l=1}^{r}\delta_l^2)$, where

$\sigma_{\delta_l}^2 = \left\{\tau_\delta + (\varepsilon_l\Theta^{\mathrm{T}}\varepsilon_l)(\boldsymbol{u}_l^{\mathrm{T}}X^{\mathrm{T}}X\boldsymbol{u}_l)\right\}^{-1}$ and $Y_{A,l} = Y - W\Gamma - Z\Gamma - \sum_{l'\neq l}\delta_l X\boldsymbol{u}_l\varepsilon_l^{\mathrm{T}}$

Conditionals for U: from equation (4.6) we can write,

$$\begin{aligned} p(\boldsymbol{u}_l|-) &\propto \text{etr}\left(-\frac{1}{2}\boldsymbol{u}_l^{\mathrm{T}}(pI_p)\boldsymbol{u}_l + (Y_{A,l} - \delta_l X\boldsymbol{u}_l\varepsilon_l)\Theta(Y_{A,l} - \delta_l X\boldsymbol{u}_l\varepsilon_l)^{\mathrm{T}}\right) \\ &\propto \text{etr}\left(\boldsymbol{u}_l^{\mathrm{T}}\delta_l X^{\mathrm{T}}Y_{A,l} - \frac{1}{2}\boldsymbol{u}_l^{\mathrm{T}}(pI_p + \delta_l^2\varepsilon_l^{\mathrm{T}}\Theta\varepsilon_l X^{\mathrm{T}}X)\boldsymbol{u}_l\right). \end{aligned}$$

This implies, $p(\boldsymbol{u}_l|-) \sim N_p\left(\delta_l\Sigma_{\boldsymbol{u}_l}X^{\mathrm{T}}Y_{A,l}\Theta\varepsilon_l, \Sigma_{\boldsymbol{u}_l}\right)$, where

$\Sigma_{\boldsymbol{u}_l} = \left\{pI_p + \delta_l^2(\varepsilon_l^{\mathrm{T}}\Theta\varepsilon_l)X^{\mathrm{T}}X\right\}^{-1}$.

Conditionals for V: from equation (4.6) we can write,

$$\begin{aligned} p(\varepsilon_l|-) &\propto \text{etr}\left(-\frac{1}{2}\varepsilon_l^{\mathrm{T}}(dI_d)\varepsilon_l + (Y_{A,l} - \delta_l X\boldsymbol{u}_l\varepsilon_l)\Theta(Y_{A,l} - \delta_l X\boldsymbol{u}_l\varepsilon_l)^{\mathrm{T}}\right) \\ &\propto \text{etr}\left(\varepsilon_l^{\mathrm{T}}\delta_l\Theta Y_{A,l}^{\mathrm{T}}X\boldsymbol{u}_l - \frac{1}{2}\varepsilon_l^{\mathrm{T}}(pI_p + \delta_l^2\boldsymbol{u}_l^{\mathrm{T}}X^{\mathrm{T}}X\boldsymbol{u}_l\varepsilon_l X^{\mathrm{T}}X)\varepsilon_l\right). \end{aligned}$$

This gives, $p(\varepsilon_l|-) \sim N_d(\delta_l \Sigma_{\varepsilon_l} \Theta Y_{A,l}^{\mathrm{T}} X \boldsymbol{u}_l, \Sigma_{\varepsilon_l})$, where
$\Sigma_{\varepsilon_l} = \left\{ dI_d + \delta_l^2 (\boldsymbol{u}_l^{\mathrm{T}} X^{\mathrm{T}} X \boldsymbol{u}_l) \Theta \right\}^{-1}$.

Sampling Γ by Columns

One could sample the coefficient matrix one element at a time, which will be time consuming and less attractive in high-dimensional setting. Another approach will be to convert the whole mtrix into a vector and sample at once; this requires a covariance matrix with dimension *qd timesqd*, which can be quite large an require huge memory making it infeasible. We choose a middle path motivated by (Khondker et al., 2013) in covariance estimation setting; this columnwise sampling scheme allows computionally efficient sampling with a feasible dimension of the conditinal covariance matrix. Let $Y_\Gamma = Y - XU\Delta V - Z\Gamma$, then we can write $Y_\Gamma = W\Gamma + E$. For $k = 1, \ldots, d$ we can partition $Y_\Gamma = (y_{\Gamma,k} \;\; Y_{\Gamma,-k})$, $\Gamma = (\gamma_k \;\; \Gamma_{-k})$, and Θ as

$$\Theta = \begin{pmatrix} \theta_{kk} & \theta_k^{\mathrm{T}} \\ \theta_k & \Theta_{-kk} \end{pmatrix}.$$

In the above partition $y_{\Gamma,k}$ is the kth column of Y_Γ, $Y_{\Gamma,-k}$ is the matrix after dropping the kth column of Y_Γ , θ_{kk} is the element at kth row and kth column of Θ, θ_k is the kth column of Θ after dropping θ_{kk}, Θ_{-kk} is the matrix after dropping kth row and kth column of Θ. We can write

$$\begin{aligned}
p(\gamma_k|-) &\propto \mathrm{etr}\left(-\frac{1}{2}\left[(Y_\Gamma - W\Gamma)^{\mathrm{T}}(Y_\Gamma - W\Gamma)\Theta + \gamma_k^{\mathrm{T}}(\tau_\gamma I_q)\gamma_k \right] \right) \\
&\propto \mathrm{etr}\left(-\frac{1}{2}(y_{\Gamma,k} - W\gamma_k)^{\mathrm{T}}(y_{\Gamma,k} - W\gamma_k)\theta_{kk} \right) \\
&\quad \times \mathrm{etr}\left((y_{\Gamma,k} - W\gamma_k)^{\mathrm{T}}(Y_{\Gamma,-k} - W\Gamma_{-k})\theta_k - \frac{1}{2}\gamma_k^{\mathrm{T}}(\tau_\gamma I_q)\gamma_k \right) \\
&\propto \mathrm{etr}\left(\gamma_k^{\mathrm{T}} W^{\mathrm{T}} \{ y_{\Gamma,k} - (Y_{\Gamma,-k} - W\Gamma_{-k})\theta_k \} - \frac{1}{2}\gamma_k^{\mathrm{T}}(\theta_{kk} W^{\mathrm{T}} W + (\tau_\gamma I_q))\gamma_k \right).
\end{aligned}$$

This gives us $p(\gamma_k|-) \sim N_d(W^{\mathrm{T}}\{y_{\Gamma,k} - (Y_{\Gamma,-k} - W\Gamma_{-k})\theta_k, \Sigma_{\gamma_k})$,

where $\Sigma_{\gamma_k} = \left\{\theta_{kk}W^{\mathrm{T}}W + (\tau_\gamma I_q)\right\}^{-1}$. The conditionals for b can be derived in a similar fashion. Conditionals for all other parameters are straightforward.

BIBLIOGRAPHY

Armagan, A., Dunson, D., and Lee, J. (2011), "Generalized double Pareto shrinkage," *submitted*.

Banerjee, O., Ghaoui, L. E., and d'Aspremont, A. (2007), "Model selection through sparse maximum likelihood estimation," *Journal of Machine Learning Research*, 9, 485–516.

Bansal, V., Libiger, O., Torkamani, A., and Schork, N. J. (2010), "Statistical analysis strategies for association studies involving rare variants," *Nature Reviews Genetics*, 11, 773–785.

Barnard, J., McCulloch, R., and Meng, X. (2000), "Modeling covariance matrices in terms of standard deviations and correlations with application to shrinkage," *Statistica Sinica*, 10, 1281–1311.

Benjamini, Y. and Hochberg, Y. (1995), "Controlling the false discovery rate: a practical and powerful approach to multiple testing," *Journal of the Royal Statistical Society, Ser. B*, 57, 289–300.

Bhattacharya, A. and Dunson, D. B. (2011), "Sparse Bayesian infinite factor models," *Biometrika*, 98, 291–306.

Breiman, L. (1996), "Heuristics of instability and stabilization in model selection," *Annals of Statistics*, 24, 2350–2383.

Breiman, L. and Friedman, J. (1997), "Predicting multivariate responses in multiple linear regression," *Journal of the Royal Statistical Society*, 59, 3–54.

Brown, P., Vannucci, M., and Fearn, T. (2002), "Bayes Model Averaging With Selection of Regressors," *Journal of the Royal Statistical Society (Series B)*, 64, 519–536.

Candés, E. J., Li, X., Ma, Y., and Wright, J. (2009), "Robust principal component analysis," *Submitted*.

Cannon, T. D. and Keller, M. (2006), "Endophenotypes in the genetic analyses of mental disorders," *Annu Rev Clin Psychol.*, 40, 267–290.

Carvalho, C. M. and Scott, J. G. (2009a), "The horseshoe estimator for sparse signals," *Biometrika*, 97, 497–512.

— (2009b), "Objective Bayesian model selection in Gaussian graphical models," *Biometrika*, 96, 785–801.

Chen, H. and Wang, Y. (2011), "A penalized spline approach to functional mixed effects models analysis," *Biometrics*, 67, 861–870.

Chen, K., Chan, K. S., and Stenseth, N. R. (2012), "Reduced-rank stochastic regression with a sparse singular value decomposition," *Journal of the Royal Statistical Society (Series B)*, 74, 203–221.

Chen, Z. and Dunson, D. (2003), "Random effects selection in linear mixed models," *Biometrics*, 59, 762–769.

Chiang, M. C., Barysheva, M., Toga, A. W., Medland, S. E., Hansell, N. K., James, M. R., McMahon, K. L., de Zubicaray, G. I., Martin, N. G., Wright, M. J., and Thompson, P. M. (2011a), "BDNF gene effects on brain circuitry replicated in 455 twins," *NeuroImage*, 55, 448–454.

Chiang, M. C., McMahon, K. L., de Zubicaray, G. I., Martin, N. G., Hickie, I., Toga, A. W., Wright, M. J., and Thompson, P. M. (2011b), "Genetics of white matter development: A DTI study of 705 twins and their siblings aged 12 to 29," *NeuroImage*, 54, 2308–2317.

Davidson, E. and Levin, M. (2005), "Gene regulatory networks," *Proceedings of the National Academy of Sciience*, 102, 4935.

Dawid, A. P. and Lauritzen, S. L. (1993), "Hyper-Markov laws in the statistical analysis of decomposable graphical models," *The Annals of Statistics*, 21, 1272–1317.

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

Ding, X., He, L., and Carin, L. (2011), "Bayesian robust principal component analysis," *IEEE Transaction on Imaging*, PP, 1–1.

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

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

Escoufer, Y. (1973), "Le traitement des variables vectorielles," *Biometrics*, 29, 751–760.

Fan, J., Feng, Y., and Wu, Y. (2009), "Network Exploration Via the Adaptive LASSO and SCAD Penalties," *Annals of Applied Statistics*, 3, 521–541.

Fan, J. and Li, R. (2001), "Variable selection via nonconcave penalized likelihood and its oracle properties," *Journal of the American Statistical Association*, 96.

Fan, J. and Lv, J. (2010), "A selective overview of variable selection in high dimensional feature space," *Statistica Sinica*, 20, 101–148.

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

— (2008b), "Sparse inverse covariance estimation with the graphical lasso," *Biostatistics*, 9, 432–441.

Frühwirth-Schnatter, S. and Tüchler, R. (2008), "Bayesian parsimonious covariance estimation for hierarchical linear mixed models," *Statistical Computing*, 18, 1–13.

Gao, W., Zhu, H., Giovanellom, K., Smith, J., Shen, D., Gilmoref, J., and Lin, W. (2009), "Evidence on the emergence of the brain default network from 2-week-old to 2-year-old healthy pediatric subjects," *PNAS*, 106, 6790–6795.

Geweke, J. and Zhou, G. (1996), "Measuring the pricing error of the arbitrage pricing theory," *The Review of Financial Studies,*, 9, 557–587.

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

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

Izenman, A. J. (1975), "Reduced-rank regression for the multivariate linear model," *Journal of Multivariate Analysis*, 5.

Jolliffe, I. T. (2002), *Principal Component Analysis*, New York: Springer.

Khondker, Z., Zhu, H., Chu, H., Lin, W., and Ibrahim, J. (2013), "The Bayesian covariance lasso," *Statistics and its Interface,*, 6(2), 243–259.

Khondker, Z. S., Zhu, H., Chu, H., Lin, W., and Ibrahim, J. G. (2011), "The Bayesian covariance lasso," *Submitted*.

Kyung, M., Gill, J., and Ghosh, M. (2010), "Penalized Regression, Standard Errors, and Bayesian Lassos," *Bayesian Analysis*, 5, 369–412.

Liu, J. S., Liang, F., and Wong, W. H. (2000), "The multiple-try method and local optimization in metropolis sampling," *Journal of the American Statistical Association*, 95.

Meinshausen, N. and Buhlmann, P. (2010), "Stability selection (with discussion)." *Journal of the Royal Statistical Society: Series B,*, 72, 417–473.

Müller, P., Parmigiani, G., Robert, C., and Rousseau, J. (2004), "Optimal sample size for multiple testing: The case of gene expression microarrays," *Journal of the American Statistical Association*, 99, 990–1001.

Park, T. and Casella, G. (2008), "The Bayesian Lasso," *Journal of the American Statistical Association*, 103, 681–686.

Paus, T. (2010), "Population neuroscience: Why and how," *Human Brain Mapping*, 31, 891–903.

Peng, J., Zhu, J., Bergamaschi, A., Han, W., Noh, D. Y., Pollack, J. R., and Wang, P. (2010), "Regularized multivariate regression for identifying master predictors with application to integrative genomics study of breast cancer," *Annals of Applied Statistics*, 4, 53–77.

Peper, J. S., Brouwer, R. M., Boomsma, D. I., Kahn, R. S., and Pol, H. E. H. (2007), "Genetic influences on human brain structure: A review of brain imaging studies in twins," *Human Brain Mapping*, 28, 464–473.

Reinsel, G. C. and Velu, P. (1998), *Multivariate reduced-rank regression: theory and applications*, New York: Springer.

Rothman, A. J., Levina, E., and Zhu, J. (2010), "Sparse multivariate regression with covariance estimation," *Journal of Computational and Graphical Statistics*, 19, 947–962.

Sachs, K., Perez, O., Pe'er, D., Lauffenburger, D., and Nolan, G. (2003), "Causal protein-signaling networks derived from multiparameter single cell data," *Science*, 308, 523–529.

Schäfer, J. and Strimmer, K. (2005), "A shrinkage approach to large-scale covariance matrix estimation and implications for functional genomics," *Genetics and Molecular Biology*, 4, 1–32.

Scharinger, C., Rabl, U., Sitte, H. H., and Pezawas, L. (2010), "Imaging genetics of mood disorders," *NeuroImage*, 53, 810–821.

Schur, I. (1909), "n the characteristic roots of a linear substitution with an application to the theory of integral equations," *Mathematische Annalen*, 66, 488–510.

Shen, D. G. and Davatzikos, C. (2004), "Measuring temporal morphological changes robustly in brain MR images via 4-dimensional template warping," *NeuroImage*, 21, 1508–1517.

Silver, M., Janousova, E., Hue, X., Thompson, P., and Montana, G. (2012), "Identification of gene pathways implicated in Alzheimer's disease using longitudinal imaging phenotypes with sparse regression," *Neuroimage*, 63, 1681–1694.

Smith, M. and Kohn, R. (2002), "Bayesian parsimonious covariance matrix estimation for longitudinal data," *Journal of the American Statistical Association*, 87, 1141–1153.

Thomas, D. (2010), "Gene–environment-wide association studies: emerging approaches," *Nature Reviews Genetics*, 11, 259–272.

Tibshirani, R. (1996), "Regression shrinkage and selection via the lasso," *Journal of the Royal Statistical Society (Series B)*, 58.

Turlach, B. A., Venables, W. N., and Wright, S. J. (2005), "Simultaneous variable selection." *Technometrics*, 47, 349–363.

Turner, J. A., Smyth, P., Macciardi, F., Fallon, J., Kennedy, J., and Potkin, S. (2006), "Imaging phenotypes and genotypes in schizophrenia." *Neuroinformatics*, 40, 21–49.

Vounou, M., Janousova, E., Wolz, R., Stein, J. L., Thompson, P. M., Rueckert, D., Montana, G., and ADNI (2011), "Sparse reduced-rank regression detects genetic associations with voxel-wise longitudinal phenotypes in Alzheimer's disease," *NeuroImage*, 60, 700–716.

Vounou, M., Nichols, T. E., Montana, G., and ADNI (2010), "Discovering genetic associations with high-dimensional neuroimaging phenotypes: A sparse reduced-rank regression approach," *NeuroImage*, 53, 1147–1159.

Wang, H., Nie, F., Huang, H., Kim, S., Nho, K., Risacher, S. L., Saykin, A. J., and Shen, L. (2012), "Identifying quantitative trait loci via group-sparse multitask regression and feature selection: an imaging genetics study of the ADNI cohort," *Bioinformatics*, 28, 229–237.

Wang, L. and Dunson, D. (2010), "Semiparametric Bayes multiple testing: Applications to tumor data," *Biometrics*, 66, 493–501.

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

Yang, R. and Berger, I. (2007), "Estimation of covariance matrix using the reference prior," *Unpublished paper*.

Yin, J. and Li, H. (2011), "A sparse conditional gaussian graphical model for analysis of genetical genomics data," *Annals of Applied Statistics*, 5, 2630–2650.

Yuan, M., Ekici, A., Lu, Z., and Monterio, R. (2007), "Dimension Reduction and Coefficient Estimation in Multivariate Linear Regression," *Journal of the Royal Statistical Society, Ser. B*, 69, 329–346.

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

Zou, H. (2006), "The Adaptive Lasso and Its Oracle Properties," *Journal of the American Statistical Association*, 101.