Bayesian Group-Sparse Multi-Task Regression for NeuroImaging Genetics

(I) Keelin Greenlaw[1], Elena Szefer[2], Jinko Graham[2], Mary Lesperance[1], Farouk Nathoo[1]

(II) Yin Song[1], Shufei Ge[2], Yunlong Nie[2], Jiguo Cao[2], Liangliang Wang[2], Farouk Nathoo[1]

[1]University of Victoria, [2]Simon Fraser University

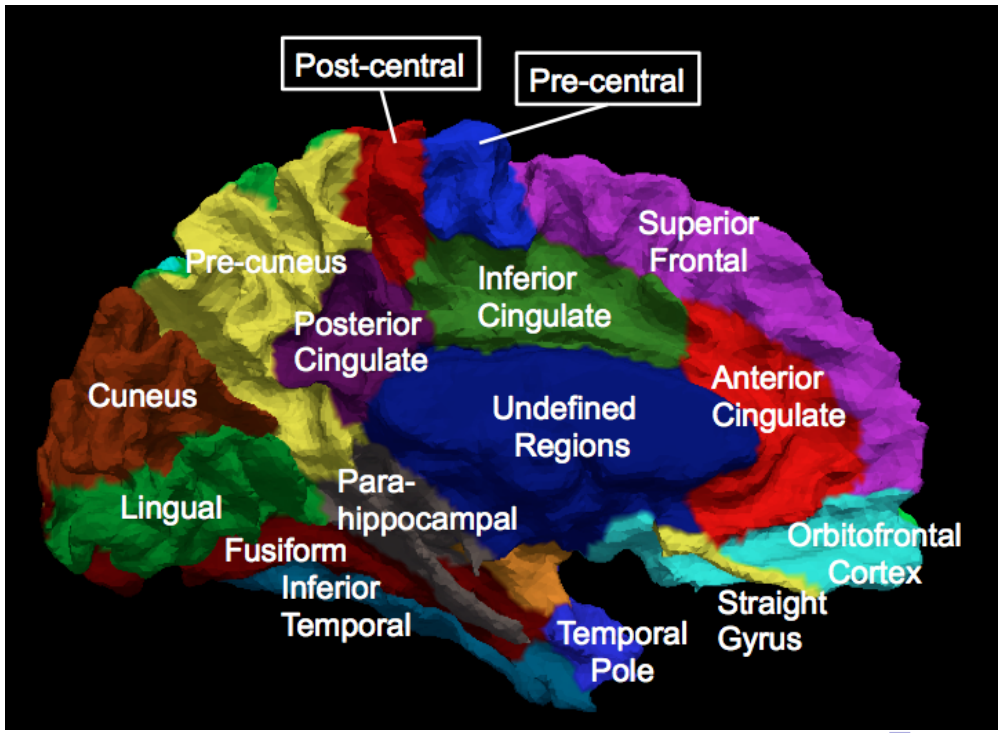
Outline

Introduction: Imaging Genomics

- Imaging genetics: interest in associations between genetic variations and neuroimaging measures as quantitative traits (QTs).
- Compared to case-control status, the QTs derived through neuroimaging may have have increased statistical power, may be closer to the underlying biological etiology of disease, perhaps making it easier to identify underlying genes.

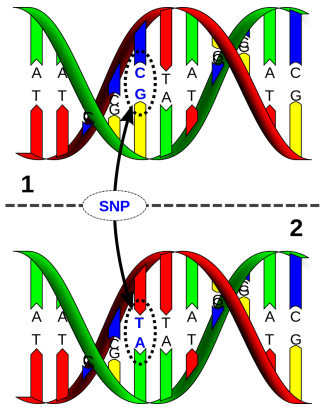
Introduction: Imaging Genomics

- Statistically, interested in a multivariate regression analysis, where the response vector comprises potentially interlinked brain imaging phenotypes that we relate to high-throughput single nucleotide polymorphism (SNP) data.
- We focus here on multivariate phenotypes (volumetric and cortical thickness values) of not very high dimension (up to 100 or so) derived from MRI for certain ROIs.



Introduction: Imaging Genomics

- The SNPs are naturally grouped by their belonging genes, and multiple SNPs from a given gene may jointly carry out genetic functionalities. Would like to account for this group structure in the regression analysis.



Introduction: Imaging Genomics

- We develop a Bayesian approach based on a continuous shrinkage prior that encourages sparsity and induces dependence in the regression coefficients corresponding to SNPs within the same gene, and across different components of the response.
- Our approach is related to the Bayesian group lasso (Park and Casella, 2008; Kyung et al., 2010) but adapted for multivariate phenotypes.
- Primarily motivated by the Group-Sparse Multi-task regression and feature selection estimator (somewhat) recently proposed by Wang et al. [2012].

Outline

Wang et al. Estimator: Data set up

Imaging data

- $\mathbf{y}_\ell = (\mathbf{y}_{\ell 1}, \ldots, \mathbf{y}_{\ell c})^T, \ \ \ell = 1, \ldots, n$
- n subjects; c response variables (QTs)

Genetic data

- $\mathbf{x}_\ell = (\mathbf{x}_{\ell 1}, \ldots, \mathbf{x}_{\ell d})^T, \ \ \ell = 1, \ldots, n$
- $\mathbf{x}_{\ell j} \in \{0, 1, 2\}$ is the number of minor allele for j^{th} SNP.
- d SNPs, which can be grouped into K genes: π_k for $k = 1, 2, \ldots, K$.

Regression coefficients

- $E(\mathbf{y}_\ell) = \mathbf{W}^T \mathbf{x}_\ell, \ \ \ell = 1, \ldots, n$
- $\mathbf{W}$ is a d x c matrix; each w_{ij} is a coefficient.

Wang et al. Estimator: First major component

$$\hat{\mathbf{W}} = \arg\min_{\mathbf{W}} \left\{ \sum_{\ell=1}^{n} ||\mathbf{W}^T \mathbf{x}_\ell - \mathbf{y}_\ell||_2^2 + \gamma_1 \sum_{k=1}^{K} \sqrt{\sum_{i \in \pi_k} \sum_{j=1}^{c} w_{ij}^2} + \gamma_2 \sum_{i=1}^{d} \sqrt{\sum_{j=1}^{c} w_{ij}^2} \right\}$$

- Residual sum of squares; element w_{ij} of $\mathbf{W}$ measures the relative importance of the i^{th} SNP to the j^{th} phenotype.

Wang et al. Estimator: Second major component

$$\hat{\mathbf{W}} = \arg\min_{\mathbf{W}} \left\{ \sum_{\ell=1}^{n} ||\mathbf{W}^T \mathbf{x}_\ell - \mathbf{y}_\ell||_2^2 + \gamma_1 \sum_{k=1}^{K} \sqrt{\sum_{i \in \pi_k} \sum_{j=1}^{c} w_{ij}^2} + \gamma_2 \sum_{i=1}^{d} \sqrt{\sum_{j=1}^{c} w_{ij}^2} \right\}$$

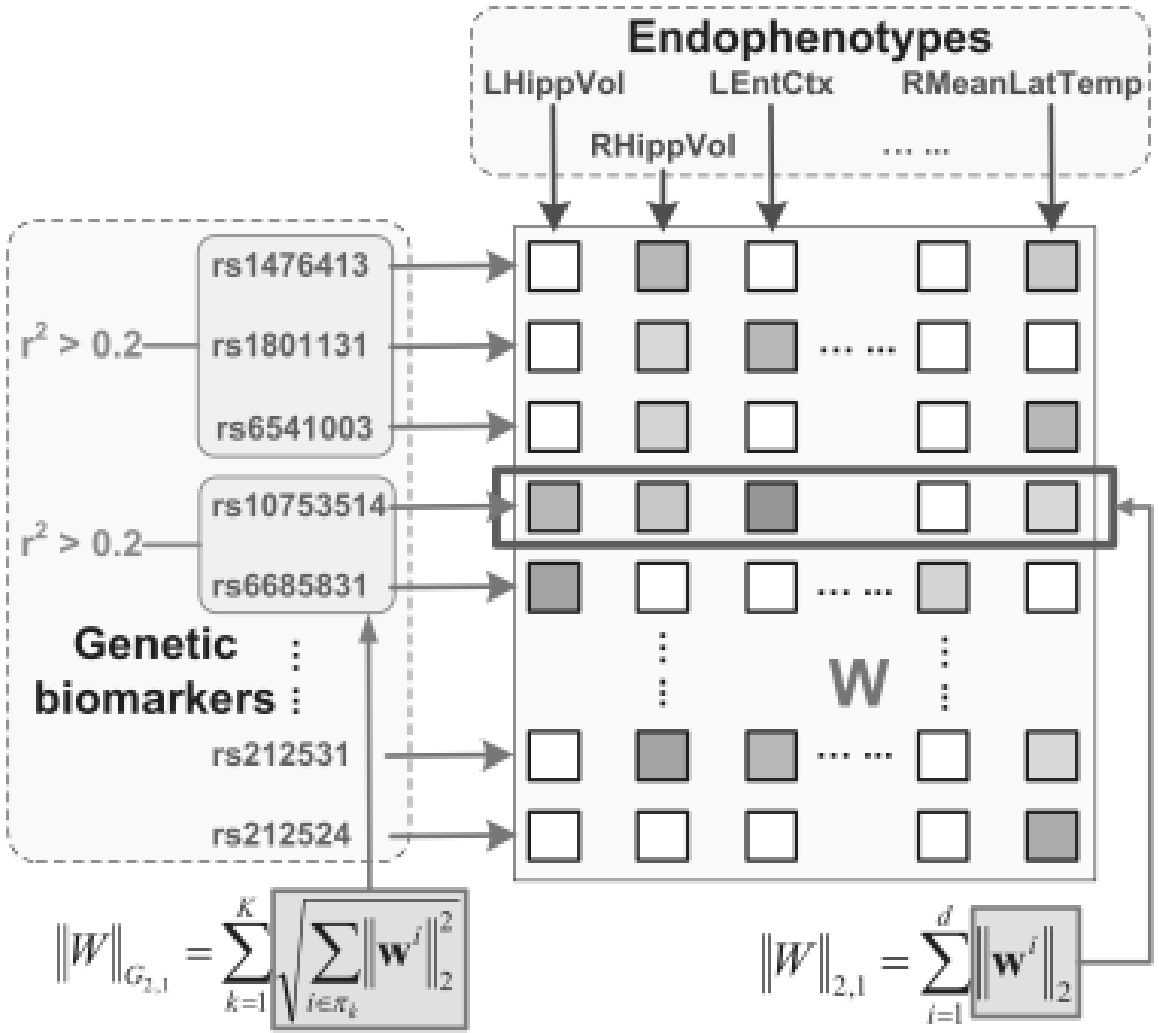
- Inspired by group lasso [Yuan and Lin, 2006], Wang et al. introduce a new form of regularization ($G_{2,1} - norm$) to address group-wise association among SNPs.
- Coefficients within a group, across all QTs, are penalized together via $\ell_2 - norm$ while $\ell_1 - norm$ is used to sum up group-wise penalties to enforce sparsity between groups.
- $G_{2,1} - norm$ regularization differs from group lasso as it penalizes regression coefficients for a group of SNPs across all responses jointly.

Wang et al. Estimator: Third major component

$$\hat{\mathbf{W}} = \arg\min_{\mathbf{W}} \left\{ \sum_{\ell=1}^{n} ||\mathbf{W}^{T}\mathbf{x}_{\ell}\text{-}\mathbf{y}_{\ell}||_2^2 + \gamma_1 \sum_{k=1}^{K} \sqrt{\sum_{i \in \pi_k} \sum_{j=1}^{c} w_{ij}^2} + \gamma_2 \sum_{i=1}^{d} \sqrt{\sum_{j=1}^{c} w_{ij}^2} \right\}$$

- As an important group may contain irrelevant individual SNPs, or a less important group may contain individually significant SNPs, an additional penalty term is added for individual structured sparsity.
- The second penalty term enforces $\ell_{2,1} - norm$ regularization for individual SNPs.

Wang et al. Estimator: 'G-SMuRFS'



Wang et al. Estimator: 'G-SMuRFS'

Group-sparse multitask regression and feature selection

$$\hat{\mathbf{W}} = \arg\min_{\mathbf{W}} \left\{ \sum_{\ell=1}^{n} ||\mathbf{W}^T \mathbf{x}_\ell - \mathbf{y}_\ell||_2^2 + \gamma_1 \sum_{k=1}^{K} \sqrt{\sum_{i \in \pi_k} \sum_{j=1}^{c} w_{ij}^2} + \gamma_2 \sum_{i=1}^{d} \sqrt{\sum_{j=1}^{c} w_{ij}^2} \right\}$$

- The combination of both penalty terms make up the novel method for SNP selection, dubbed 'G-SMuRFS' by the authors.
- Computation of $\hat{\mathbf{W}}$ is based on a simple iterative algorithm that converges to the global optimum.
- Tuning parameters, γ_1 and γ_2, are chosen by standard 5-fold cross-validation in the range of $(10^{-5}, 10^{-4}, \ldots, 10^4, 10^5)$.

Wang et al. Estimator: Limitations

- The proposed method only provides a point estimate of the regression coefficients. A method for computing standard errors is lacking.
- By noting the connection between penalized regression methods and Bayesian models, [Kyung et al., 2010, Park and Casella, 2008] we develop an equivalent hierarchical Bayesian model.
- This allows for inference based on the posterior distributions. As we can validly summarize the spread of the posterior, we have valid measures of variability. Interval estimates can then guide SNP selection.

Outline

Bayesian Model: Priors $\mathbf{W}$

$$\hat{\mathbf{W}} = \arg\min_{\mathbf{W}} \left\{ \sum_{\ell=1}^{n} ||\mathbf{W}^{\mathsf{T}}\mathbf{x}_\ell - \mathbf{y}_\ell||_2^2 + \gamma_1 \sum_{k=1}^{K} \sqrt{\sum_{i \in \pi_k} \sum_{j=1}^{c} w_{ij}^2} + \gamma_2 \sum_{i=1}^{d} \sqrt{\sum_{j=1}^{c} w_{ij}^2} \right\} \quad (1)$$

We specify a model hierarchy such that the posterior mode is identical to $\hat{\mathbf{W}}$ in (1).
First level: quantitative imaging traits, conditional on $\mathbf{W}$ and σ^2, are independently distributed as multivariate normal.

$$\mathbf{y}_\ell \,|\mathbf{W}, \sigma^2 \overset{ind}{\sim} MVN_c(\mathbf{W}^{\mathsf{T}}\mathbf{x}_\ell \,,\, \sigma^2 I_c) \quad \ell = 1, \ldots, n$$

Bayesian Model

Let $\mathbf{W}^{(k)} = \{w_{ij} | i \in \pi_k, j = 1, \ldots, c\}$ be submatrix with rows corresponding to the k^{th} gene, $k = 1, \ldots, K$.

We assign conditionally independent priors to each $\mathbf{W}^{(k)}$ to coincide with the penalty terms in (1) as follows:

$$\mathbf{W}^{(k)} | \lambda_1, \lambda_2, \sigma^2 \overset{ind}{\sim} p(\mathbf{W}^{(k)} | \lambda_1, \lambda_2, \sigma^2) \quad k = 1, \ldots, K \tag{2}$$

$$p(\mathbf{W}^{(k)} | \lambda_1, \lambda_2, \sigma^2) \propto \exp\left\{ -\frac{\lambda_1}{\sigma} \sqrt{\sum_{i \in \pi_k} \sum_{j=1}^{c} w_{ij}^2} \right\} \prod_{i \in \pi_k} \exp\left\{ -\frac{\lambda_2}{\sigma} \sqrt{\sum_{j=1}^{c} w_{ij}^2} \right\}. \tag{3}$$

Proposition 1. (Prior Propriety) The prior for $\mathbf{W}$ based on (2) and (3) is proper.

- Density of a product multivariate Laplace distribution induces dependence in coefficients across imaging phenotypes at both the SNP and gene level.
- Given the likelihood and prior the posterior mode is by construction the estimator of Wang et al. [2012].

Proposition 2. (Scale mixture representation) For each $i \in \{1, \ldots, d\}$ let $k(i) \in \{1, \ldots, K\}$ denote the gene associated with the i^{th} SNP. The prior (3) can be obtained through the following scale mixture representation:

$$w_{ij} \mid \sigma^2,\ \tau_1^2, \ldots, \tau_K^2,\ \omega_1^2, \ldots, \omega_d^2 \overset{ind}{\sim} N\left(0,\ \sigma^2\left(\frac{1}{\tau_{k(i)}^2} + \frac{1}{\omega_i^2}\right)^{-1}\right), \tag{4}$$

with continuous scale mixing variables $\tau^{\mathbf{2}} = (\tau_1^2, \ldots, \tau_K^2)'$ and $\omega^{\mathbf{2}} = (\omega_1^2, \ldots, \omega_d^2)'$ distributed according to the density

$$p(\tau^{\mathbf{2}}, \omega^{\mathbf{2}} | \lambda_1^2, \lambda_2^2) \propto \prod_{k=1}^{K} \left(\frac{\lambda_1^2}{2}\right)^{\left(\frac{m_k c+1}{2}\right)} (\tau_k^2)^{\left(\frac{m_k c+1}{2}\right)-1} \exp\left\{-\left(\frac{\lambda_1^2}{2}\right)\tau_k^2\right\}$$
$$\times \left[\prod_{i \in \pi_k} \left(\frac{\lambda_2^2}{2}\right)^{\left(\frac{c+1}{2}\right)} (\omega_i^2)^{\left(\frac{c+1}{2}\right)-1} \exp\left\{-\left(\frac{\lambda_2^2}{2}\right)\omega_i^2\right\} (\tau_k^2 + \omega_i^2)^{-\frac{c}{2}}\right] \tag{5}$$

Outline

Model Fitting: Full Conditionals

The proposed hierarchical model results in standard full conditional distributions (Gaussian, Inverse-Gaussian, Inverse-Gamma).

- $[\text{vec}(\mathbf{W}^{(k)T}) | \mathbf{Y}, \mathbf{W}^{(-k)}, \underset{\sim}{\tau^2}, \underset{\sim}{\omega^2}, \sigma^2, \lambda_1^2, \lambda_2^2] \sim MVN_{m_k c} \quad k = 1, \ldots, K$
- $[\nu_k = \frac{1}{\tau_k^2} \mid \mathbf{Y}, \mathbf{W}, \tau^2_{(-k)}, \underset{\sim}{\omega^2}, \sigma^2, \lambda_1^2, \lambda_2^2] \sim \textit{Inverse-Gaussian} \quad \text{for } k = 1, \ldots, K$
- $[\eta_i = \frac{1}{\omega_i^2} \mid \mathbf{Y}, \mathbf{W}, \underset{\sim}{\tau^2}, \omega^2_{(-i)}, \sigma^2, \lambda_1^2, \lambda_2^2] \sim \textit{Inverse-Gaussian} \quad \text{for } i = 1, \ldots, d$
- $[\sigma^2 | \mathbf{Y}, \mathbf{W}, \underset{\sim}{\tau^2}, \underset{\sim}{\omega^2}, \lambda_1^2, \lambda_2^2] \sim \textit{Inv} - \textit{Gamma}$

Past work on Bayesian lassos [Park and Casella, 2008, Kyung et al., 2010] have discussed two methods for estimation of tuning parameters $(\lambda_1^2, \lambda_2^2)$.

Model Fitting: Estimation of λ_1^2 and λ_2^2

Fully Bayesian model

Assign conditionally conjugate gamma priors for λ_1^2 and λ_2^2.

$$\lambda_1^2 \sim Gamma(r_1, \delta_1)$$

$$\lambda_2^2 \sim Gamma(r_2, \delta_2)$$

λ_1^2 and λ_2^2 can be included as unknown parameters in the Gibbs Sampling algorithm; we can accommodate the uncertainty associated with their values.

Model Fitting: Estimation of λ_1^2 and λ_2^2

Empirical Bayes framework

Motivated by potential sensitivity to priors an alternative suggested approach is to estimate the tuning parameters by maximizing the marginal likelihood.

$$\hat{\lambda}_1^2,\ \hat{\lambda}_2^2 \ = \ \underset{\lambda_1^2,\lambda_2^2}{\arg\max} \int_{\Theta} p\left(\mathbf{Y}, \Theta \mid \lambda_1^2,\ \lambda_2^2\right) d\Theta$$

$$\textit{where } \Theta = (\mathbf{W}, \underset{\sim}{\tau^2}, \underset{\sim}{\omega^2}, \sigma^2)$$

This using a Monte Carlo EM algorithm.

Model Fitting: Issues with λ_1^2 and λ_2^2 Estimation

We begin investigating the behaviour of our MCMC algorithm by simulating data from the model, where the underlying true **W** is known.

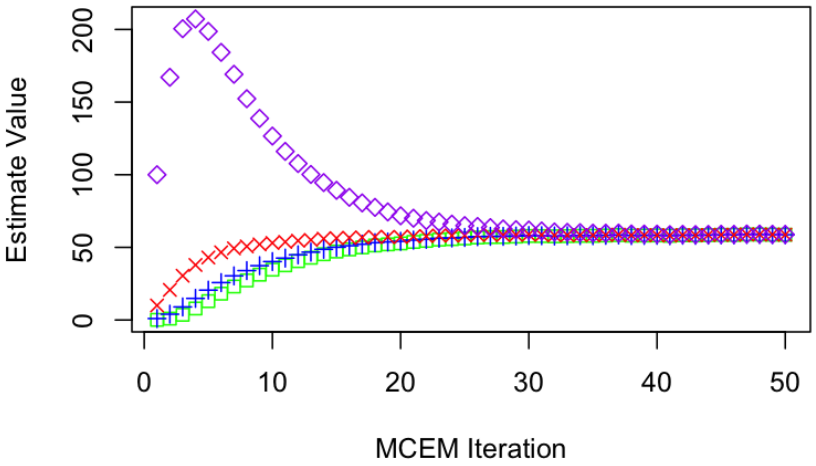
The behaviour changes drastically in two different settings.

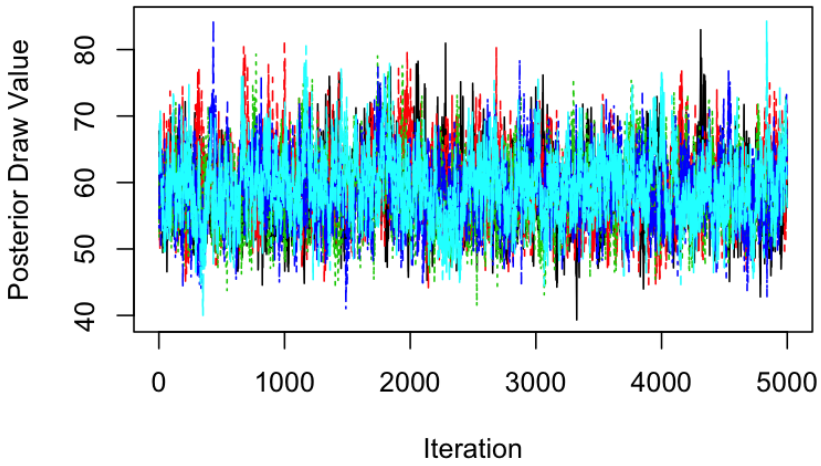
Case 1

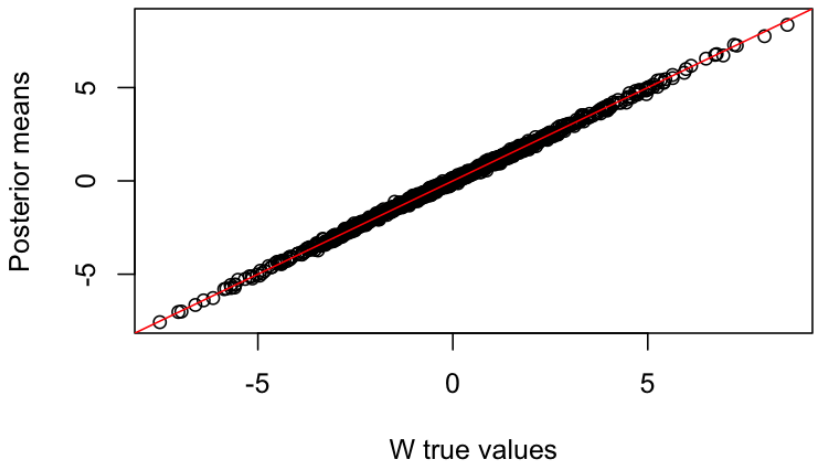
number of SNPs (d) $\ll$ number of simulated observations (n)

Behaviour:
Everything works fine!
Gibbs sampling λ_1^2 and λ_2^2 estimates converge to reasonable values.
MCEM converges.
Resulting Bayes estimate of **W** looks very accurate.

Simulation ($d = 200, n = 500$): Empirical/Full Bayes







Model Fitting: Issues with λ_1^2 and λ_2^2 Estimation

We begin investigating the behaviour of our MCMC algorithm by simulating data from the model, where the underlying true $\mathbf{W}$ is known.

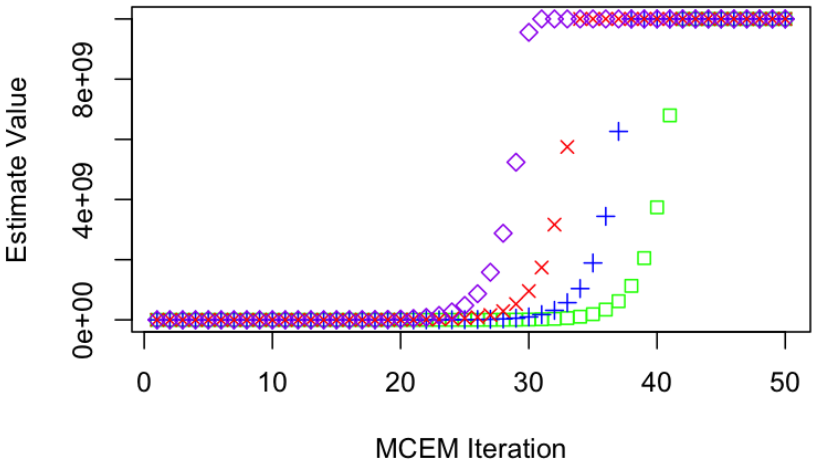
The behaviour changes drastically in two different settings.

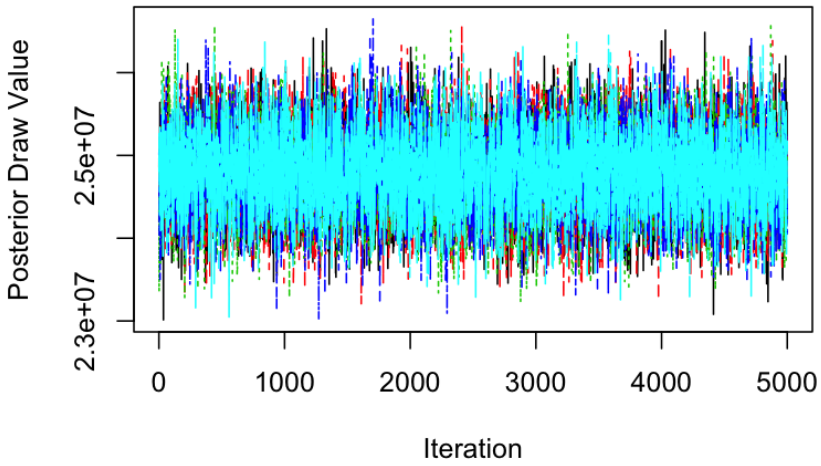
Case 2

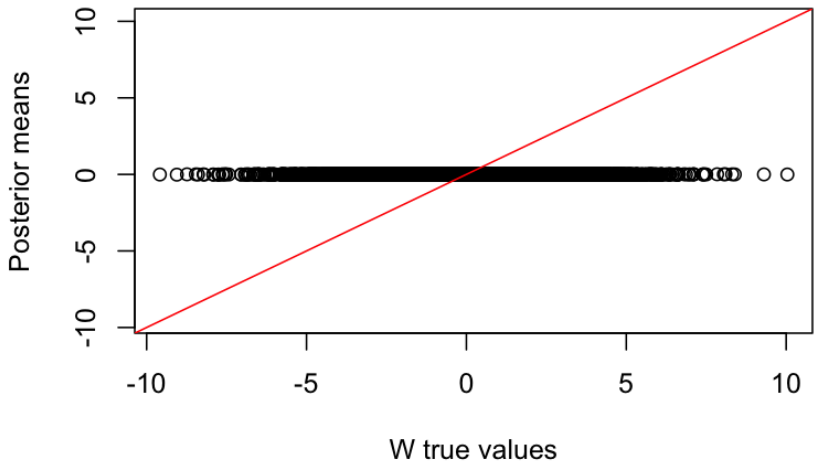
number of SNPs (d) $>$ number of simulated observations (n)

Behaviour:
Fully Bayes: Gibbs sampling λ_1^2 and λ_2^2 chains converge to very large values.
Empirical Bayes: MCEM diverges.
Resulting Bayes estimate of $\mathbf{W}$ exhibits severe overshrinkage.

Simulation ($d = 1500, n = 500$): Empirical/Full Bayes

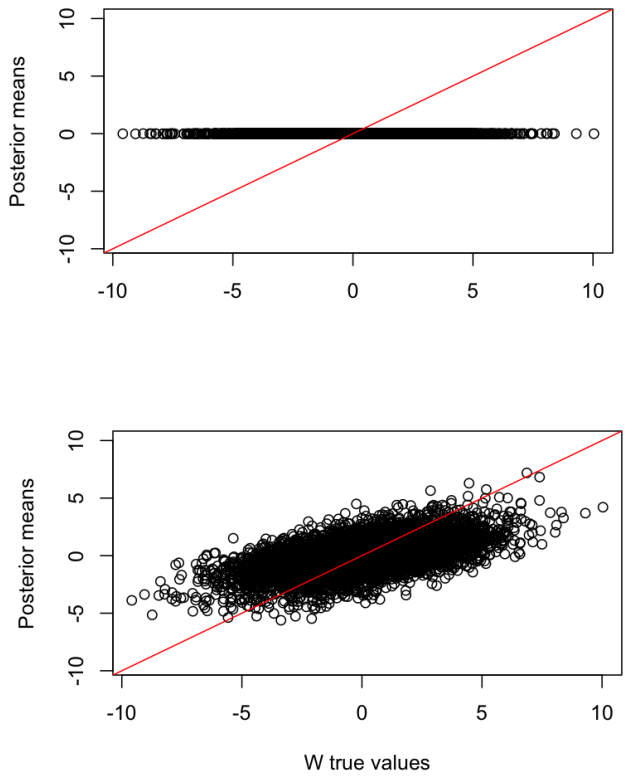






Simulation ($d = 1500, n = 500$): Fixed λ_1^2 and λ_2^2 Results

Empirical/Fully Bayes: Tuning parameter estimation is causing the problem. When λ_1^2 and λ_2^2 are fixed at their data generating values, the Bayes estimate looks better in cases where $d > n$.



Model Fitting: λ_1^2, λ_2^2 Estimation Discussion

Problem with choosing tuning parameters:

- With a large number of SNPs, choosing the tuning parameters based on the marginal likelihood/posterior leads to over shrinkage.
- Study the shape of the marginal likelihood, $p(\mathbf{Y}|\,\lambda_1^2, \lambda_2^2)$.

Model Fitting: Studying the Marginal Likelihood

Marginal likelihood:

$$p(\mathbf{Y}|\,\lambda_1^2,\lambda_2^2) = \int p\,(\mathbf{Y},\mathbf{W},\sigma^2,\underset{\sim}{\tau}^2,\underset{\sim}{\omega}^2\,|\,\lambda_1^2,\lambda_2^2)\;d\,\mathbf{W}\,d\,\sigma^2\,d\,\underset{\sim}{\tau}^2\,d\,\underset{\sim}{\omega}^2$$

W is marginalized out of the expression by using the basic properties of the Gaussian distribution.

$$\mathbf{Y}\,|\,\underset{\sim}{\tau}^2,\underset{\sim}{\omega}^2,\sigma^2 \;\sim\; MVN(0,\;(I_c\otimes\mathbf{X})\Sigma_w(I_c\otimes\mathbf{X}^T)+\sigma^2\,I_{cn}\,)$$

$$where\;\;\Sigma_w=\sigma^2\,I_c\otimes Diag\left\{\left(\frac{1}{\omega_i^2}+\frac{1}{\tau_{k(i)}^2}\right)^{-1},\;i=1,\ldots,d\right\}$$

Model Fitting: Studying the Marginal Likelihood

$$p(\mathbf{Y}|\,\lambda_1^2,\lambda_2^2) = \int \left[\int_0^\infty p(\mathbf{Y},|\,\sigma^2,\underset{\sim}{\tau^2},\underset{\sim}{\omega^2})p(\sigma^2)d\,\sigma^2\right] p(\underset{\sim}{\tau^2},\underset{\sim}{\omega^2}\,|\,\lambda_1^2,\lambda_2^2)\;d\,\underset{\sim}{\tau^2}\;d\,\underset{\sim}{\omega^2}$$

- Using properties of the Inv-Gamma distribution, σ^2 is analytically integrated out of the expression.
- The remaining integration is analytically intractable. Require some approximations to obtain a closed form expression.
- First approximate the scale mixing distribution as

$$\begin{aligned} p(\underset{\sim}{\tau^2},\underset{\sim}{\omega^2}\,|\lambda_1^2,\lambda_2^2) \approx & \prod_{k=1}^{K} \text{Gamma}\left(\tau_k^2\middle|\left(\frac{m_k c+1}{2}\right),\left(\frac{\lambda_1^2}{2}\right)\right) \\ & \times \prod_{i=1}^{d} \text{Gamma}\left(\omega_i^2\middle|\left(\frac{c+1}{2}\right),\left(\frac{\lambda_2^2}{2}\right)\right) \end{aligned} \tag{6}$$

where terms of the form $(\tau_k^2+\omega_i^2)^{-\frac{c}{2}}$ are omitted in the product.

Model Fitting: Studying the Marginal Likelihood

We then use a plug-in approximation.

$$p(\mathbf{Y}|\,\lambda_1^2,\ \lambda_2^2)\ =\ E_{\underset{\sim}{\tau^2},\,\underset{\sim}{\omega^2}}\left[p(\mathbf{Y}|\,\underset{\sim}{\tau^2},\underset{\sim}{\omega^2})\right]\ \approx p(\mathbf{Y}|\,E[\underset{\sim}{\tau^2}]\,,E[\underset{\sim}{\omega^2}]\,)$$

where

$$E[\tau_k^2] = \frac{m_k c + 1}{\lambda_1^2}\ \ ;\ \ E[\omega_i^2] = \frac{c+1}{\lambda_2^2}$$

under the Gamma approximation.

Model Fitting: Studying the Marginal Likelihood

Marginal Likelihood Approximation

$$p(\mathbf{Y}|\,\lambda_1^2,\ \lambda_2^2) \approx$$

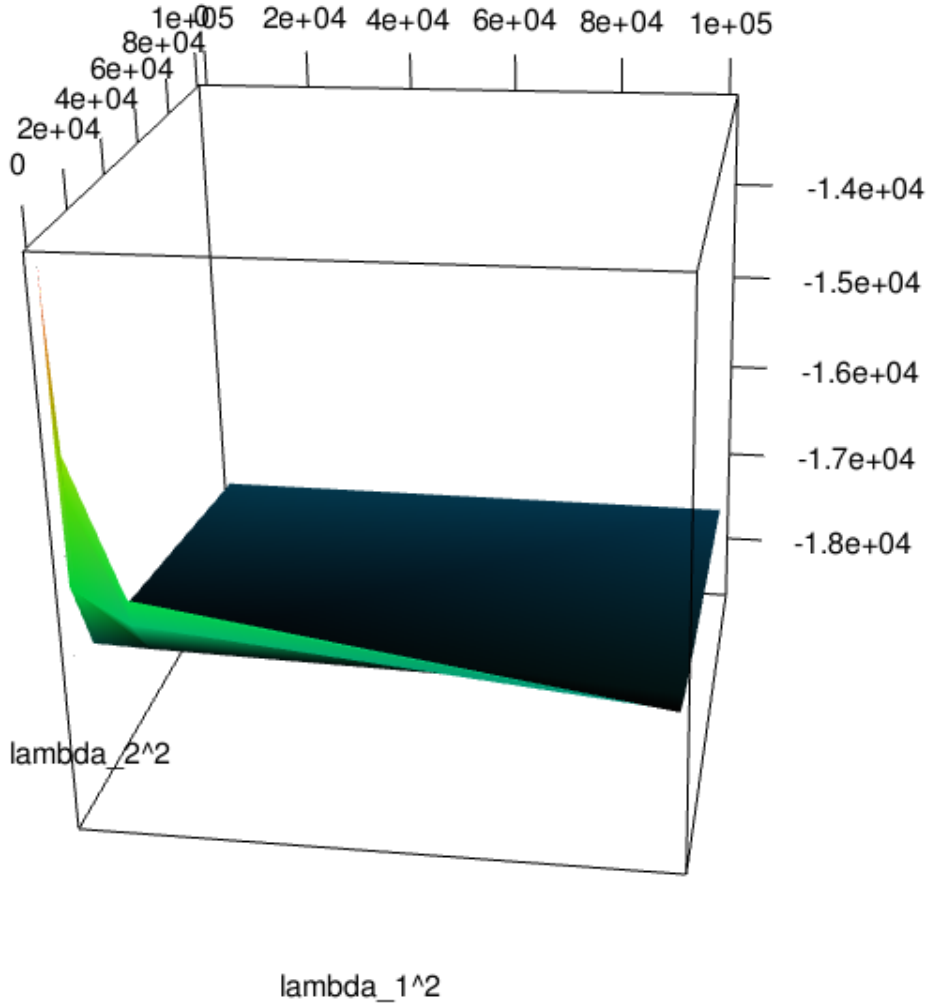
$$(2\pi)^{-\frac{nc}{2}}\, a_\sigma^{b\sigma}\, \frac{\Gamma(\frac{nc}{2}+a_\sigma)}{\Gamma(a_\sigma)} \times \left|(\mathrm{I_c}\otimes\mathbf{X})\left(\mathrm{I_c}\otimes \mathrm{Diag}\left\{\left(\frac{\lambda_2^2}{\mathrm{c}+1}+\frac{\lambda_1^2}{\mathrm{m_{k(i)}c}+1}\right)^{-1}\right\}\right)(\mathrm{I_c}\otimes\mathbf{X}^{\mathrm{T}})+\mathrm{I_{cn}}\ )\right|^{-\frac{1}{2}}\times$$

$$\left(b_\sigma+\frac{1}{2}\mathbf{Y}^T\left[(\mathrm{I_c}\otimes\mathbf{X})\left(\mathrm{I_c}\otimes \mathrm{Diag}\left\{\left(\frac{\lambda_2^2}{\mathrm{c}+1}+\frac{\lambda_1^2}{\mathrm{m_{k(i)}c}+1}\right)^{-1}\right\}\right)(\mathrm{I_c}\otimes\mathbf{X}^{\mathrm{T}})+\mathrm{I_{cn}}\ )\right]^{-1}\mathbf{Y}\right)^{-(\frac{nc}{2}+a_\sigma)}$$

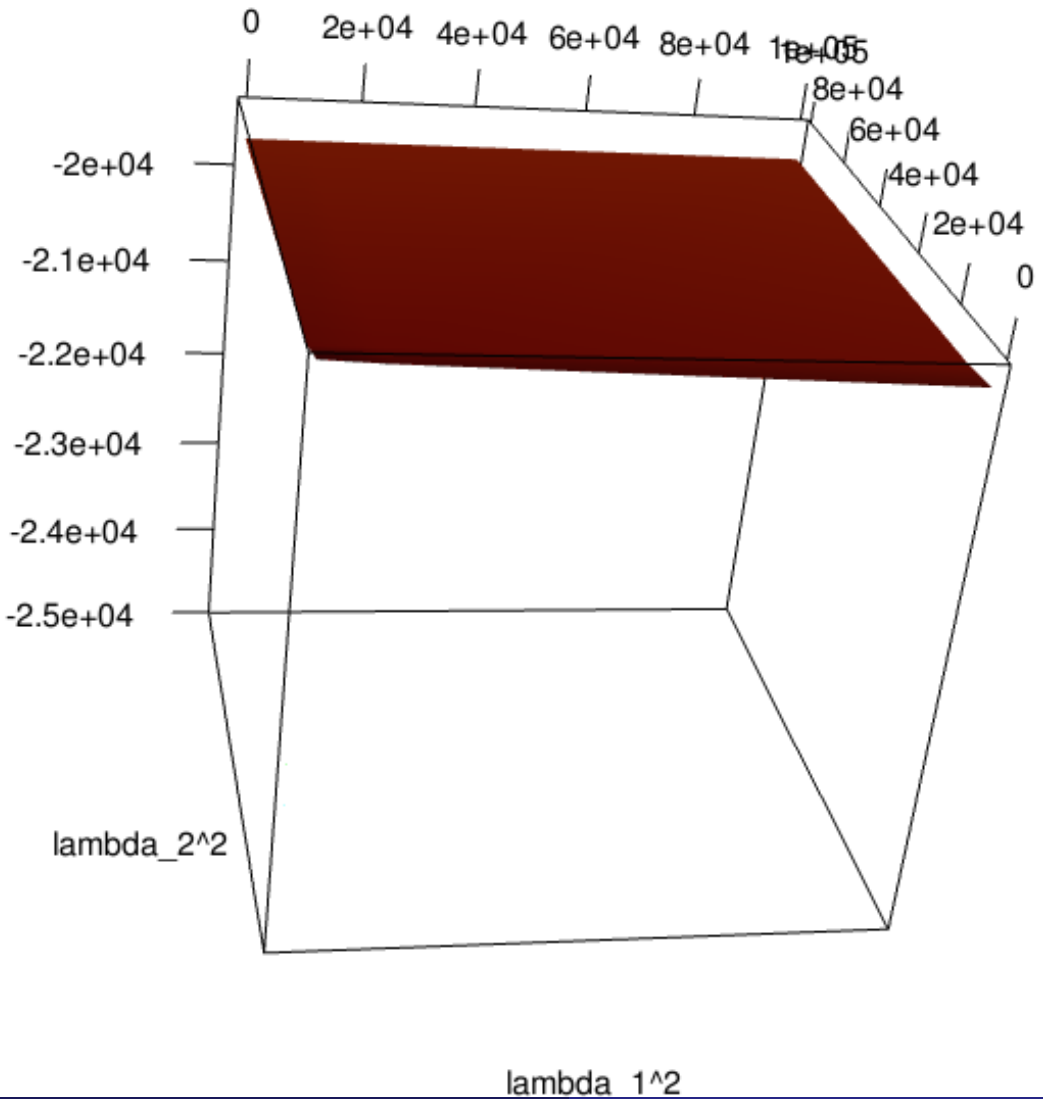
- The approximation is evaluated over a grid of $(\lambda_1^2, \lambda_2^2)$ values for different sets of simulated data.

'Nicely Behaved' Marginal Likelihood Approximation

- simulated data:
  $d = 200; c = 5; n = 500$
- Clearly identified mode and no over-shrinkage in estimate of $\mathbf{W}$.



'Poorly Behaved' Marginal Likelihood Approximation



- simulated data: $d = 1500; c = 5; n = 500$
- Grid search: majority of simulated surfaces have maximum point at: $\lambda_1^2 \geq 10^4; \lambda_2^2 \geq 10^4$
- So we have some better intuition now, but still not entirely clear.

Model Fitting: Studying the Marginal Likelihood

- Fix $n < \infty$, $c < \infty$, assume $m_k = m, \forall k$ and $\mathbf{X}\mathbf{X}^T = \mathbf{I}$, and examine $\lim_{d\to\infty} \tilde{p}(\mathbf{Y}|\, \lambda_1^2,\ \lambda_2^2)$ where now $d = mK$.
  1. Case: $d \to \infty$, $m \to \infty$, $K < \infty$
  2. Case: $d \to \infty$, $m < \infty$, $K \to \infty$
  3. Case: $d \to \infty$, $m \to \infty$, $K \to \infty$
- Can show that $\lim_{d\to\infty} \tilde{p}(\mathbf{Y}|\, \lambda_1^2,\ \lambda_2^2)$ does not depend on λ_1^2 in cases (1) and (3).
- For these cases empirical Bayes (in limit of large d) is completely uninformative about λ_1^2; The same is also true for fully Bayes if λ_1^2 and λ_2^2 are assumed independent a priori.

Model Fitting: Cross-Validation for λ_1^2, λ_2^2?

- Abandon the model and use simpler prior?
- We found empirically that cross-validation avoids some of the observed problems with FB and EB choice of the tuning parameters when $d > n$. I don't know why, but do have some intuition.
- Combining Gibbs sampling with data-splitting over a 2-D grid of tuning parameters is computationally intensive.

Model Fitting: WAIC

- We use WAIC (Watanabee, 2010) which does not require any data splitting for its computation and can be viewed as an approximation to leave-one-out cross-validation (Gelman, Hwang and Vehtari, 2013).

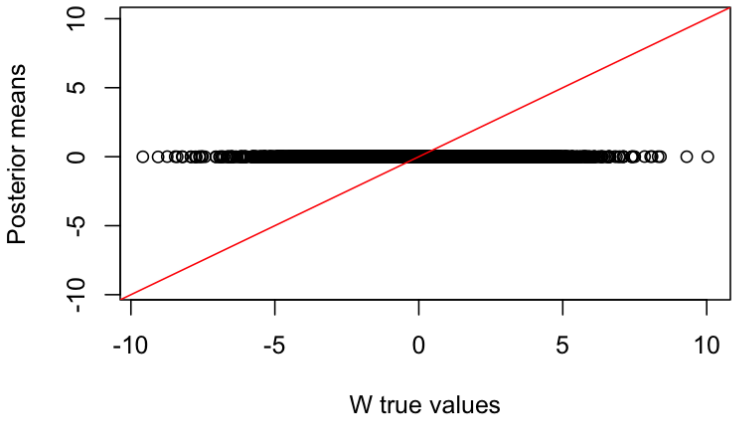
$$WAIC = -2\sum_{l=1}^{n} \log E_{\mathbf{W},\sigma^2}[p(\mathbf{y}_\ell \mid \mathbf{W}, \sigma^2)|\mathbf{y}_1, \ldots, \mathbf{y}_n]$$

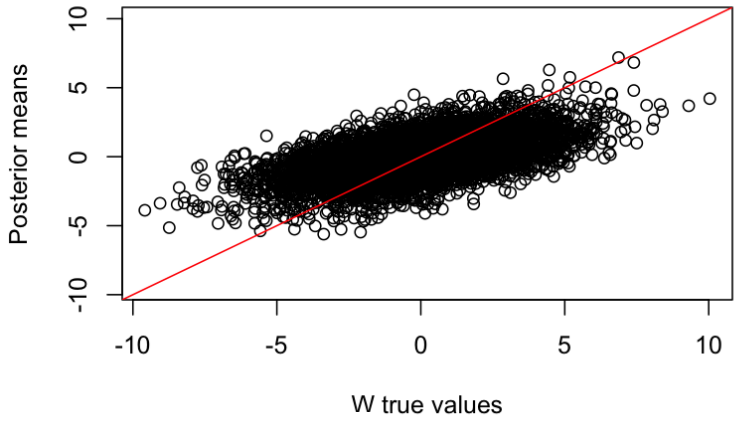
$$+2\sum_{l=1}^{n} V_{\mathbf{W},\sigma^2}[\log p(\mathbf{y}_\ell \mid \mathbf{W}, \sigma^2)|\mathbf{y}_1, \ldots, \mathbf{y}_n]$$

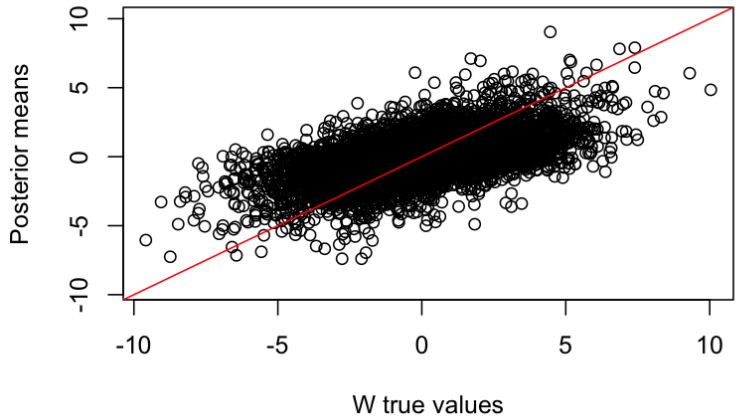
- We run Gibbs samplers in parallel over a 2D grid for λ_1^2, λ_2^2 and choose the tuning parameters minimizing WAIC.

Simulation $d = 1500, n = 500$:

Tuning parameters - Fully Bayes with gamma hyper-priors; plugin true values, lowest WAIC:







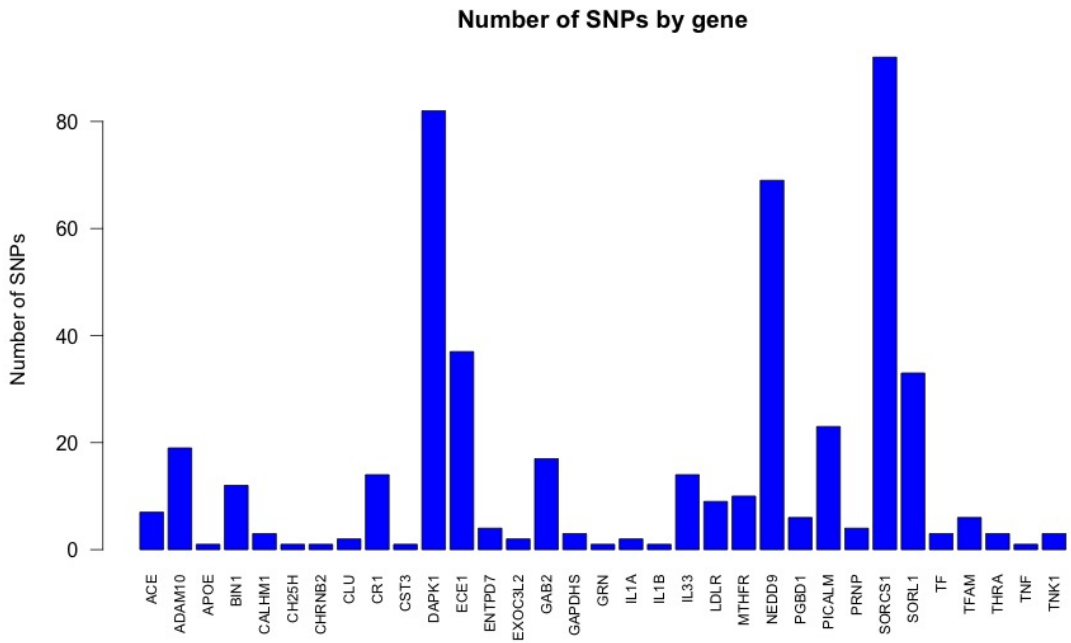
Outline

Simulation Study: The Data

Genetic Data

The SNP covariates used for data simulation come from the Alzheimer's Disease Neuroimaging Initiative (ADNI) database.

We include genetic data on 632 subjects over 486 SNPs belonging to 33 different genes.



Simulation Study: The Data

True **W** Structure

A **W** matrix is simulated from its prior distribution with the following settings.

- number of SNPs $(d) = 486$
- SNPs are partitioned into 33 (K) genes
- number of phenotypes $(c) = 12$
- $\sigma^2 = \lambda_1^2 = \lambda_2^2 = 2$

Sparsity is introduced to **W** by setting all but 50 rows to zero.
Only the following rows are left at their simulated values.

- rows corresponding to 5 genes of SNP sizes 14, 10, 6, 4, 1 (35 SNPs)
- rows corresponding to 15 other SNPs

1. Study I: $n = 632$ $(n > d)$, (Study III has MVT_4 errors)
2. Study II: $n = 250$ $(n < d)$, (Study IV has MVT_4 errors)

Simulation Study: Methodology

Simulate 100 datasets. We apply the Wang et al. estimator and our Gibbs-WAIC Bayesian method to each.

Wang et al. model fitting

Interval estimates based on non-parametric bootstrap.

Bayesian model fitting

Interval estimates are 95% equal-tail credible intervals from MCMC sampler.

Simulation Study: Coverage Probability

Table: Simulation studies - interval estimation. The coverage probability of each approximate 95% credible/confidence interval is estimated based on 100 simulation replicates and then averaged (MCP) overall.

Study I ($n > d$)	
Method	MCP (overall)
Bayesian Model	0.95
Nonparametric Bootstrap	0.85
Study II ($n < d$)	
Method	MCP (overall)
Bayesian Model	0.94
Nonparametric Bootstrap	0.85
Study III ($n > d$, t_4-errors)	
Method	MCP (overall)
Bayesian Model	0.97
Nonparametric Bootstrap	0.86
Study IV ($n < d$, t_4-errors)	
Method	MCP (overall)
Bayesian Model	0.95
Nonparametric Bootstrap	0.84

Simulation Study: Coverage Probability

Table: Simulation studies - interval estimation. The coverage probability of each approximate 95% credible/confidence interval is estimated based on 100 simulation replicates and then averaged (MCP) overall and also separately over the parameters that correspond to active SNPs.

	Study I	
Method	MCP (overall)	MCP ($w_{ij} \neq 0$)
Bayesian Model	0.95	0.83
Nonparametric Bootstrap	0.85	0.45
	Study II	
Method	MCP (overall)	MCP ($w_{ij} \neq 0$)
Bayesian Model	0.94	0.72
Nonparametric Bootstrap	0.85	0.42
	Study III	
Method	MCP (overall)	MCP ($w_{ij} \neq 0$)
Bayesian Model	0.97	0.77
Nonparametric Bootstrap	0.86	0.49
	Study IV	
Method	MCP (overall)	MCP ($w_{ij} \neq 0$)
Bayesian Model	0.95	0.73
Nonparametric Bootstrap	0.84	0.41

ADNI Data Application: The Data

- Both genetic and structural MRI data used in this project were obtained from the Alzheimer's Disease Neuroimaging Initiative 1 (ADNI-1) database.
- We include genetic and brain measurement data on 632 subjects.

ADNI Data Application: Genetic Data

- Among all SNPs, only include SNPs belonging to the top 40 Alzheimer's Disease (AD) candidate genes listed on the AlzGene database as of June 10, 2010.
- Data presented here are queried from the most recent genome build as of December 2014, from ADNI-1 genomic data.
- After quality control and imputation steps, the genetic data used in this study includes 486 SNPs from 33 genes.

ADNI Data Application: MRI Data

Brain data: volumetric and cortical thickness values (extracted from MRI using FreeSurfer) for 56 regions of interest are selected to be associated with 486 SNPs from 33 genes with 632 subjects.

ID	Measurement	Region of interest
AmygVol	Volume	Amygdala
CerebCtx	Volume	Cerebral cortex
CerebWM	Volume	Cerebral white matter
HippVol	Volume	Hippocampus
InfLatVent	Volume	Inferior lateral ventricle
LatVent	Volume	Lateral ventricle
EntCtx	Thickness	Entorhinal cortex
Fusiform	Thickness	Fusiform gyrus
InfParietal	Thickness	Inferior parietal gyrus
InfTemporal	Thickness	Inferior temporal gyrus
MidTemporal	Thickness	Middle temporal gyrus
Parahipp	Thickness	Parahippocampal gyrus
PostCing	Thickness	Posterior cingulate
Postcentral	Thickness	Postcentral gyrus
Precentral	Thickness	Precentral gyurs
Precuneus	Thickness	Precuneus
SupFrontal	Thickness	Superior frontal gyrus
SupParietal	Thickness	Superior parietal gyrus
SupTemporal	Thickness	Superior temporal gyrus
Supramarg	Thickness	Supramarginal gyrus
TemporalPole	Thickness	Temporal pole
MeanCing	Mean thickness	Caudal anterior cingulate, isthmus cingulate,posterior cingulate,rostral anterior cingulate
MeanFront	Mean thickness	Caudal midfrontal, rostral midfrontal, superior frontal, lateral orbitofrontal, and medial orbitofrontal gyri, frontal pole
MeanLatTemp	Mean thickness	Inferior temporal, middle temporal, and superior temporal gyri
MeanMedTemp	Mean thickness	Fusiform, parahippocampal, and lingual gyri, temporal pole and transverse temporal pole
MeanPar	Mean thickness	Inferior and superior parietal gyri, supramarginal gyrus, and precuneus
MeanSensMotor	Mean thickness	Precentral and postcentral gyri,
MeanTemp	Mean thickness	Inferior temporal, middle temporal, superior temporal, fusiform, parahippocampal, lingual gyri, temporal pole transverse temporal pole

ADNI Data Application: Methodology

We apply the Wang et al. method and our Gibbs-WAIC Bayesian method to the data.

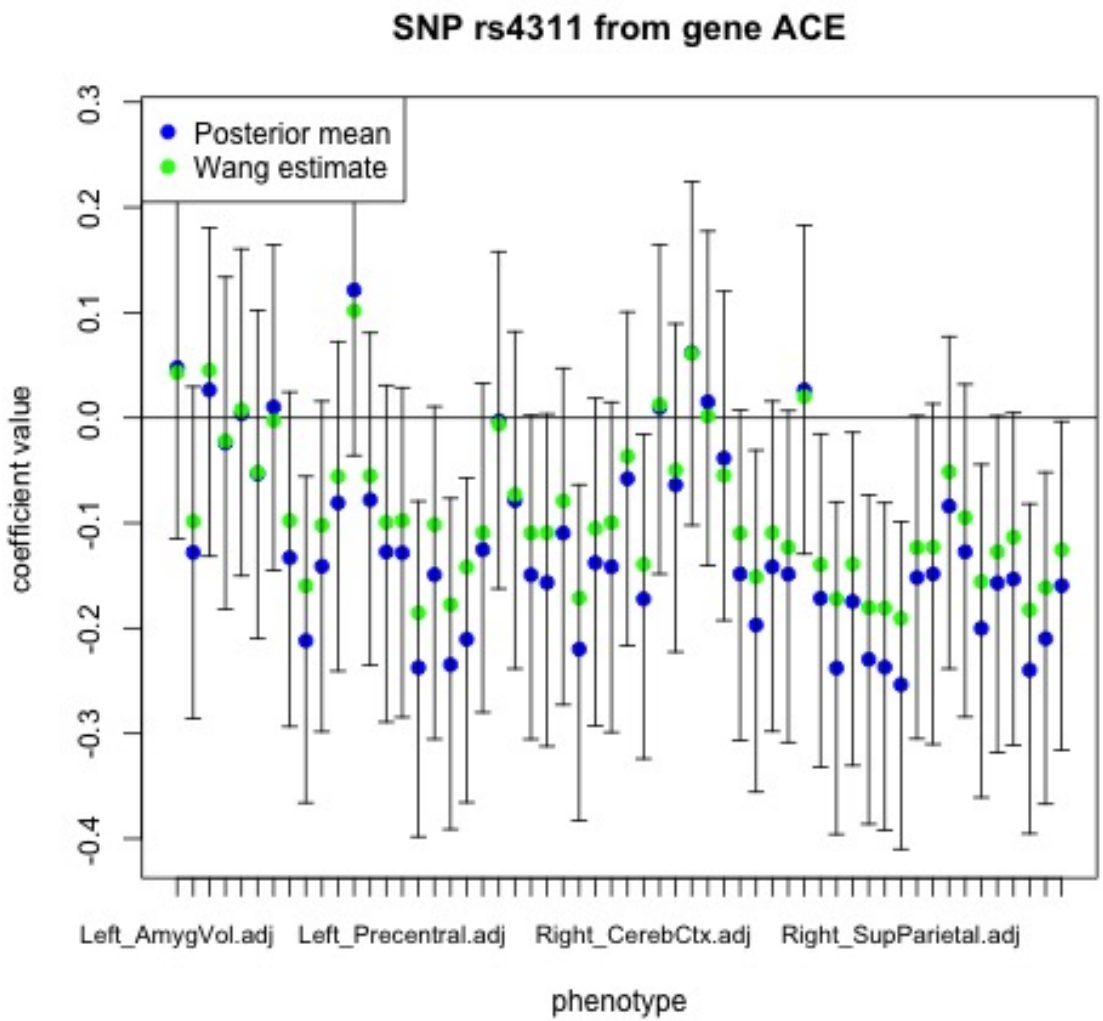
Wang et al. model fitting and SNP selection

- Wang et al. assign weights to each SNP by summing the absolute values of the estimated coefficients of a single SNP over all phenotypes.
- SNPs are ranked based on their weights.

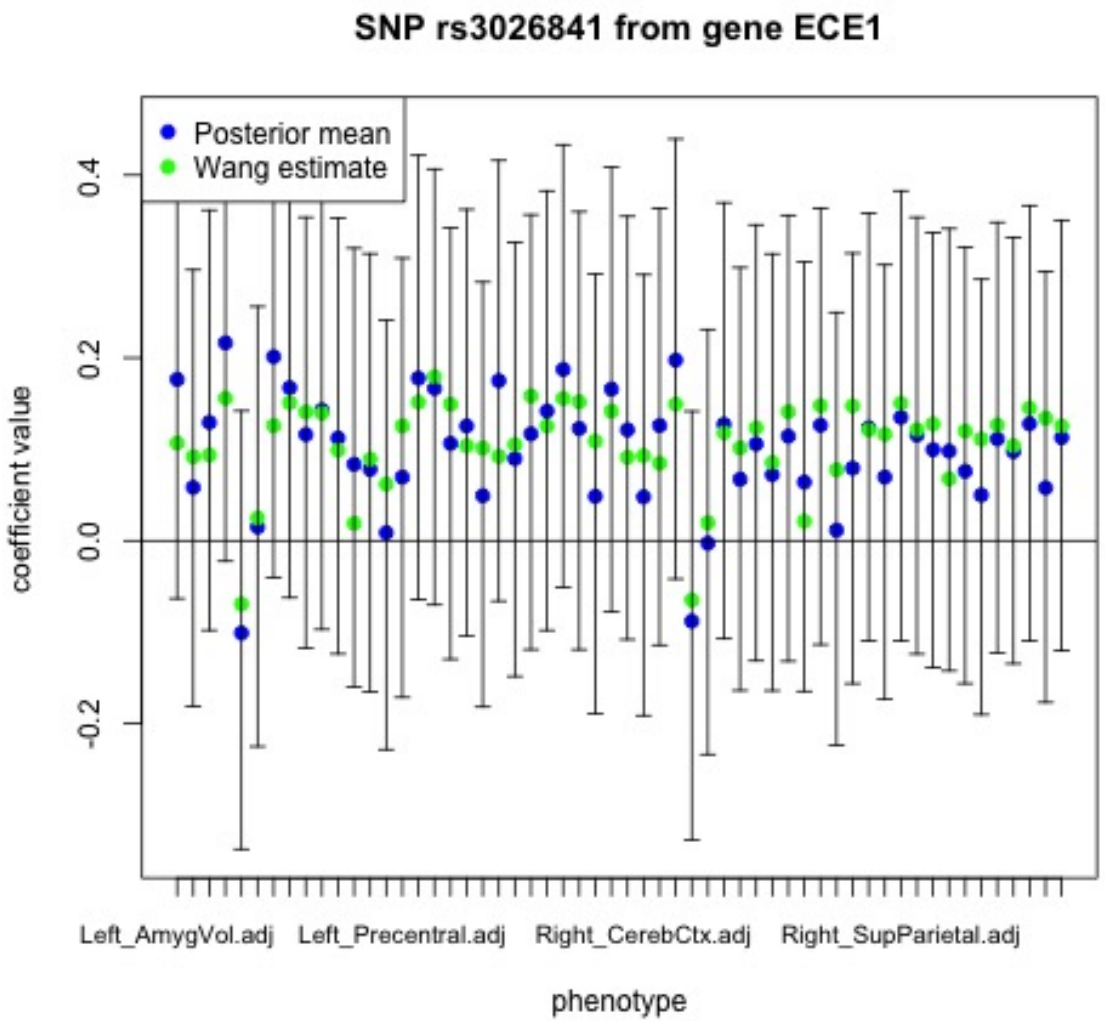
Bayesian model fitting and SNP selection

- Choose SNPs where an associated 95% equal-tail credible interval (ETCI) excludes zero for any brain measure.
- There are a total of 45 SNPs selected with total 152 elements of $\mathbf{W}$ that have 95% ETCI's that do not contain zero.
- Most important SNPs: rs405509 from APOE gene, and rs4311 from ACE gene, both potentially associated with a large number of brain measures.

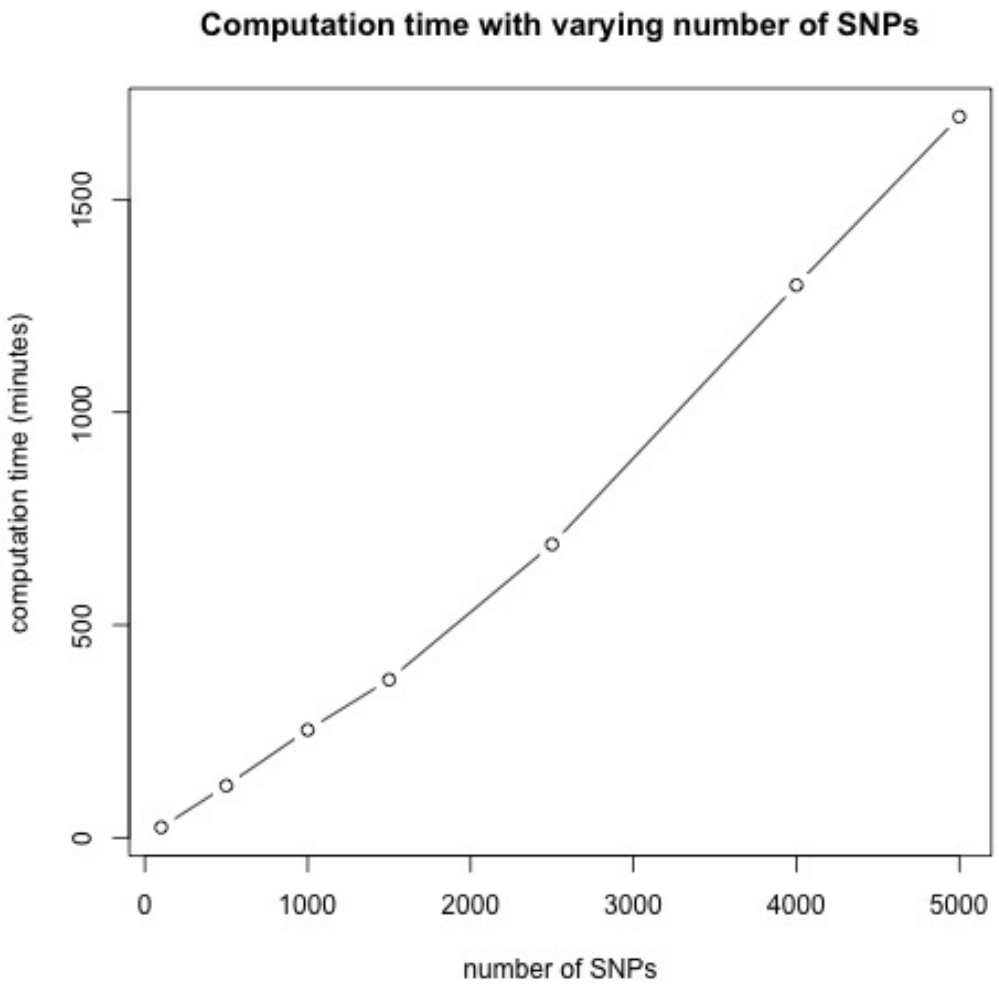
ADNI Data Application: Bayesian Model Results
SNP rs4311 from ACE Gene



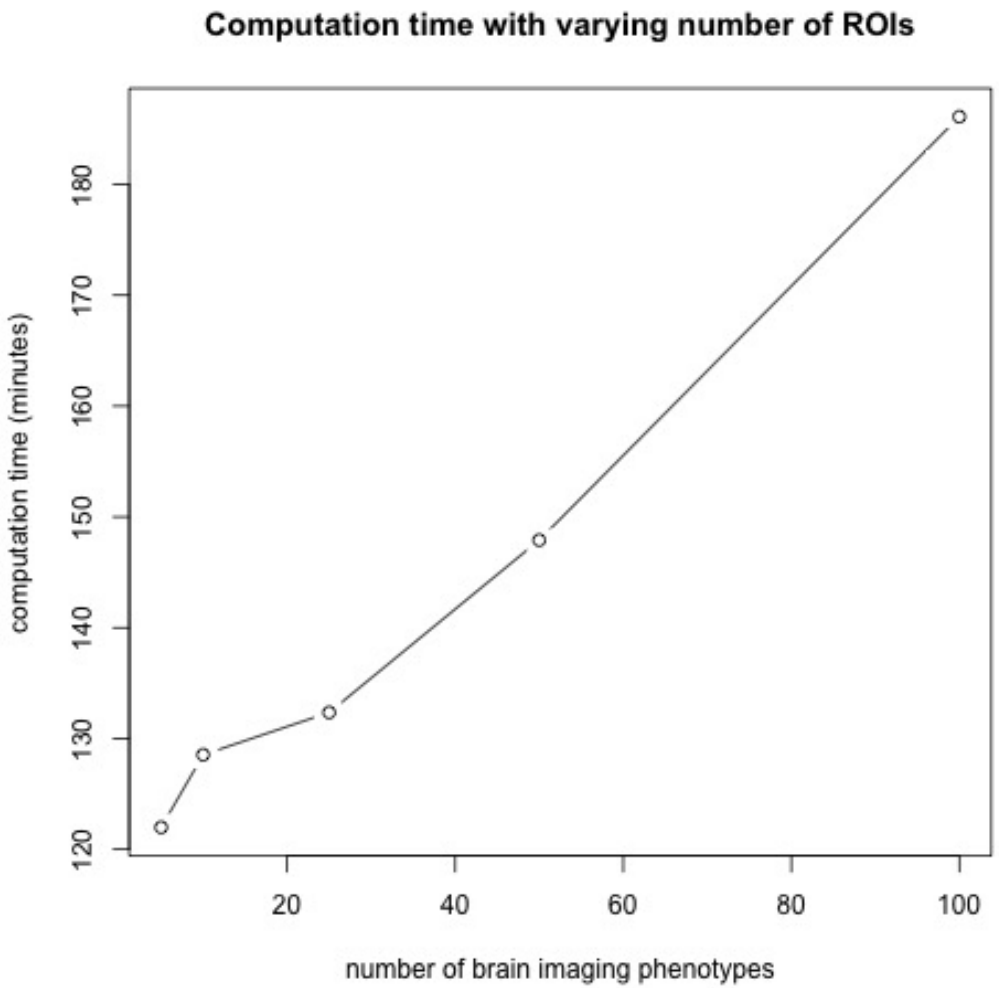
ADNI Data Application: Wang et al. Top Ranked SNP rs3026841 from gene ECE1



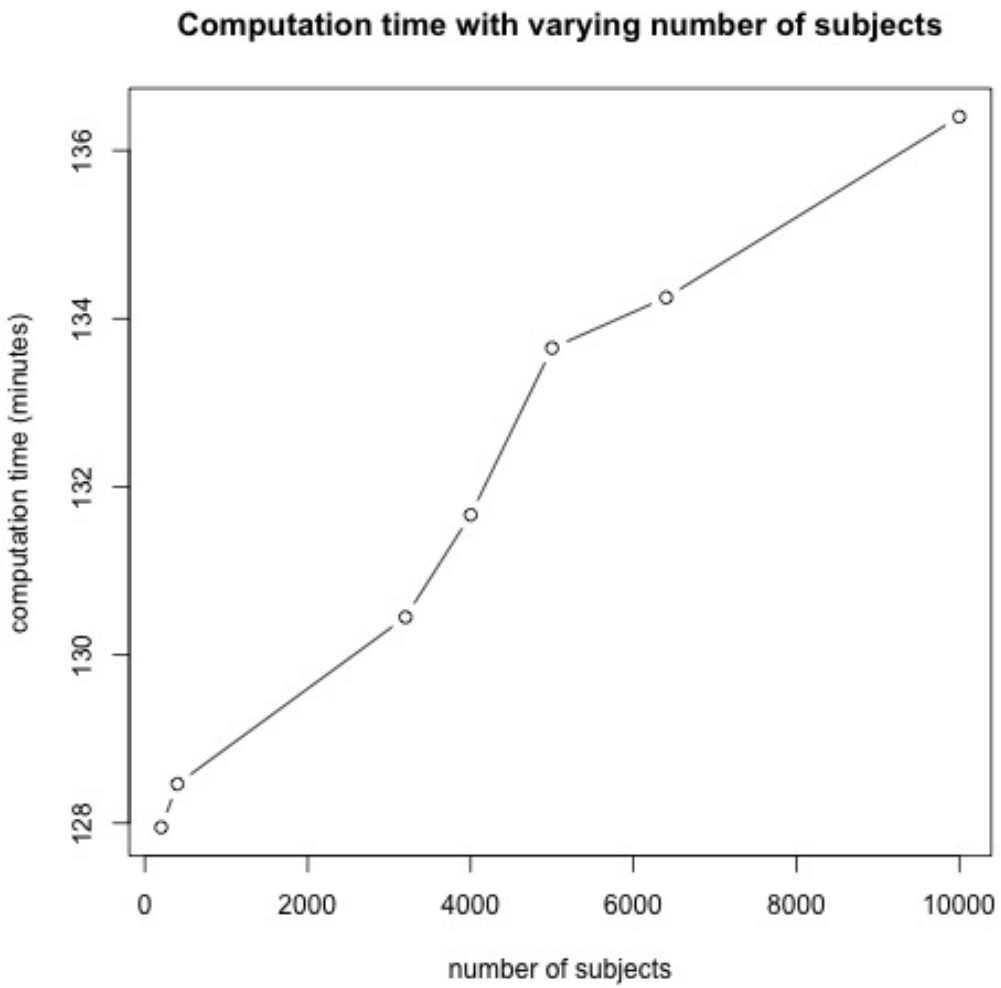
Computation Scaling - Number of SNPs ($c = 12$, $n = 600$)



Scaling - Dimension of Phenotype ($d = 500$, $n = 600$)



Scaling - Number of Subjects ($c = 12$, $d = 500$)



Extending the Model - Spatial Correlation

- First level of the model is currently quite simplistic:

$$\mathbf{y}_\ell \,|\mathbf{W}, \sigma^2 \overset{ind}{\sim} MVN_c(\mathbf{W}^\mathsf{T}\mathbf{x}_\ell \,,\, \sigma^2 I_c) \quad \ell = 1, \ldots, n$$

Extending the Model - Spatial Correlation

- First level of the model is currently quite simplistic:

$$\mathbf{y}_\ell \,|\mathbf{W}, \sigma^2 \overset{ind}{\sim} MVN_c(\mathbf{W}^\mathsf{T}\mathbf{x}_\ell \,,\, \sigma^2 I_c) \quad \ell = 1, \ldots, n$$

1. Chosen originally to obtain a correspondence between our model and the group sparse estimator of Wang et al (2012).

Extending the Model - Spatial Correlation

- First level of the model is currently quite simplistic:

$$\mathbf{y}_\ell \,|\mathbf{W}, \sigma^2 \overset{ind}{\sim} MVN_c(\mathbf{W}^\mathsf{T}\mathbf{x}_\ell \,,\, \sigma^2 I_c) \quad \ell = 1, \ldots, n$$

1. Chosen originally to obtain a correspondence between our model and the group sparse estimator of Wang et al (2012).
2. Scale each component of the response to make the assumption of a single variance component σ^2 more tenable.

Extending the Model - Spatial Correlation

- First level of the model is currently quite simplistic:

$$\mathbf{y}_\ell \,|\mathbf{W}, \sigma^2 \overset{ind}{\sim} MVN_c(\mathbf{W}^\mathsf{T}\mathbf{x}_\ell \,,\, \sigma^2 I_c) \quad \ell = 1, \ldots, n$$

1. Chosen originally to obtain a correspondence between our model and the group sparse estimator of Wang et al (2012).
2. Scale each component of the response to make the assumption of a single variance component σ^2 more tenable.
3. Leads to sparse precision matrices in the full conditional distributions of $\mathbf{W}^{(k)}$; using associated numerical methods then leads to a reduction in computation time of more than three orders of magnitude for large datasets. Don't want to give that up.

Extending the Model - Spatial Correlation

- The MRI (FreeSurfer) response comprise brain summary measures at the ROI's on both hemispheres of the brain. Data exhibit two types of correlation:

Extending the Model - Spatial Correlation

- The MRI (FreeSurfer) response comprise brain summary measures at the ROI's on both hemispheres of the brain. Data exhibit two types of correlation:
  1. Correlation between corresponding measures on opposite hemispheres (e.g. volume of left hippocampus correlated with volume of right hippocampus).

Extending the Model - Spatial Correlation

- The MRI (FreeSurfer) response comprise brain summary measures at the ROI's on both hemispheres of the brain. Data exhibit two types of correlation:
  1. Correlation between corresponding measures on opposite hemispheres (e.g. volume of left hippocampus correlated with volume of right hippocampus).
  2. Spatial correlation between measures collected on same hemisphere.

Extending the Model - Spatial Correlation

- The MRI (FreeSurfer) response comprise brain summary measures at the ROI's on both hemispheres of the brain. Data exhibit two types of correlation:
  1. Correlation between corresponding measures on opposite hemispheres (e.g. volume of left hippocampus correlated with volume of right hippocampus).
  2. Spatial correlation between measures collected on same hemisphere.
- Let $\mathbf{y}_{\ell i} = (y_{li}^{(L)}, y_{li}^{(R)})'$ be the brain summary measures obtained at the i^{th} ROI for both hemispheres; $\mathbf{y}_\ell = (\mathbf{y}_{\ell 1}', \ldots, \mathbf{y}_{\ell c/2}')'$ is the imaging data for subject l.

$$\mathbf{y}_\ell = \mathbf{W}^\mathsf{T}\mathbf{x}_\ell + \boldsymbol{\epsilon}_l$$

- Spatial model for $\boldsymbol{\epsilon}_l$ is based on a bivariate conditional autoregressive model (Gelfand and Vounatsou, 2013).

Extending the Model - Spatial Correlation

- Assume $\mathbf{A}$ is an adjacency matrix $A_{ij} \in \{0, 1\}$ representing spatial neighbourhood structure of ROIs on each hemisphere, with $\mathbf{D_A} = \text{diag}\{A_{i\cdot},\ i = 1, ..., c/2\}$.
- Conditional specification:

$$\boldsymbol{\epsilon}_{Ii} | \boldsymbol{\epsilon}_{I\{-i\}}, \rho, \Sigma \sim \text{BVN}\left(\frac{\rho}{A_{i\cdot}} \sum_{j=1}^{c/2} A_{ij} \boldsymbol{\epsilon}_{Ij},\ \frac{1}{A_{i\cdot}} \Sigma\right),\ i = 1, \ldots, c/2$$

1. $\rho \in [0, 1)$ characterizes spatial dependence
2. $\Sigma_{12} / \sqrt{\Sigma_{11} \Sigma_{22}} \in [-1, 1]$ characterizes dependence in brain measures across opposite hemispheres

Extending the Model - Spatial Correlation

- The first level of the spatial model is then

$$\mathbf{y}_\ell \,|\mathbf{W}, \rho, \Sigma \overset{ind}{\sim} MVN_c(\mathbf{W}^\mathsf{T}\mathbf{x}_\ell \,,\, (\mathbf{D_A} - \rho\mathbf{A})^{-1} \otimes \Sigma) \quad \ell = 1, \ldots, n$$

  with other levels of the model as before.
- So long as the adjacency matrix $\mathbf{A}$ is sparse this model still results in sparse precision matrices where required for faster computation
- Two additional parameters $\rho \sim \text{Unif}(0, 1)$ and $\Sigma \sim \text{inv-Wishart}(\mathbf{S}, \nu)$ are easily added to existing Gibbs sampling algorithm.

Outline

Discussion

- We use a hierarchical Bayes representation of the estimator proposed by Wang et al. [2012] and develop a Gibbs sampling approach for uncertainty quantification.
- Interval estimates seem to have reasonable coverage probabilities for the settings considered; much better than applying non-parametric bootstrap to original estimator.
- There are a number of model extensions to be considered: spatial model, improved scaling, other shrinkage priors.
- Tuning parameters: comparison of hierarchical Bayes, empirical Bayes, and cross-validation yields unexpected results.

Discussion

Papers:

1. Keelin Greenlaw*, Elena Szefer, Jinko Graham, Mary Lesperance, **Farouk S. Nathoo**. "A Bayesian Group Sparse Multi-Task Regression Model for Imaging Genetics." (2017). *Bioinformatics*, DOI: 10.1093.
2. **Nathoo, Farouk S.**, Keelin Greenlaw*, and Mary Lesperance. "Regularization Parameter Selection for a Bayesian Multi-Level Group Lasso Regression Model with Application to Imaging Genomics." (2016). *Pattern Recognition in Neuroimaging, PRNI 2016, IEEE*, DOI: 10.1109/PRNI.2016.7552328.
3. Elena Szefer, Donghuan Lu, **Farouk S. Nathoo**, Mirza Faisal Beg, Jinko Graham. "Multivariate association between single-nucleotide polymorphisms in Alzgene linkage regions and structural changes in the brain: discovery, refinement and validation." (2017). *Statistical Applications in Genetics and Molecular Biology* Under Revision.

Software: R Package 'bgsmtr' available on CRAN.

Funding Acknowledgement:

- NSERC (Discovery Grant, Tier II CRC)
- CANSSI (CRT on Neuroimaging Data Analysis)

References

Minjung Kyung, Jeff Gill, Malay Ghosh, and George Casella. Penalized regression, standard errors, and bayesian lassos. *Bayesian Analysis*, 5(2):369–411, 2010.

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

Hua Wang, Feiping Nie, Heng Huang, Sungeun Kim, Kwangsik Nho, Shannon L Risacher, Andrew J Saykin, Li Shen, et al. Identifying quantitative trait loci via group-sparse multitask regression and feature selection: an imaging genetics study of the adni cohort. *Bioinformatics*, 28(2):229–237, 2012.

Ming Yuan and Yi Lin. Model selection and estimation in regression with grouped variables. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 68(1):49–67, 2006.