Multivariate Bayesian Analysis with Incomplete Data: Application to Local Ancestry Effects on Admixed Transcriptome

Tae Kim* and Dan Nicolae**

Department of Statistics, University of Chicago, 5747 S. Ellis Avenue, Chicago, IL.

**email:* tk382@uchicago.edu

***email:* nicolae@uchicago.edu

Summary: Modern biomedical data sets present statistical challenges by often providing measurements of multiple phenotypes of each individual. One of the challenges is that the multidimensional nature of the data requires the estimation of complex dependence structure, and another is that these data sets are often incomplete due to various technical reasons. In this paper, we propose a Bayesian hierarchical model and variable selection method that tackle both challenges. We use multivariate normal regression model to detect associations between a covariate and correlated phenotypes and to simultaneously estimate the variance-covariance. Also, we provide a sensible way to circumvent the issues related to the missing values. Simulations show that the proposed method works well even with low sample size, weak signals, and high missing rates. We apply this method to Genotype-Tissue Expression (GTEx) data to identify genes where either local or global ancestry is associated with the expression level with any of the tissues, and we find multiple genes with cross-tissue signals.

Key words: Bayesian variable selection; Markov chain Monte Carlo method; Multivariate regression analysis; Local ancestry; Multi-tissue gene expression.

This paper has been submitted for consideration for publication in *Biometrics*

1. Introduction

The recent development of sequencing technologies has allowed scientists to study human biology in a more comprehensive manner. In particular, the Genotype-Tissue Expression (GTEx) Project has led to valuable research about the relationship between genotype and gene expression levels across different tissues by providing both genomic data and multiple-tissue transcriptomic data . While the past works that utilize GTEx data have mostly focused on eQTL analysis (Consortium et al., 2015; Li et al., 2017; Fagny et al., 2017), none has attempted to find the relationship between the gene expression level and genetic ancestry. The genotype data provided in GTEx allows accurate estimation of the genetic ancestry of each individual, and we aim to study the relationship between the gene expression level and the genetic ancestry in admixed population while accounting for the heterogeneity across the multiple tissues.

The analysis of GTEx data presents two important statistical challenges. First, it is difficult to fully account for the complex dependence structure across the tissues because the covariance matrix introduces a large number of parameters. Second, the tissue accessibility is different for each sample, introducing high proportion of missing values Incomplete data not only hinders common tasks such as computing the likelihood of the data but also reduces the amount of information. Given the low sample size of admixed population, missing values make inference particularly demanding.

We propose a Bayesian hierarchical modeling and inference method to analyze the effect of local ancestry of African American genome on incomplete, correlated gene expression level data. Our method not only accounts for the covariance structure of the response variable but also provides a flexible interpretation of the effect size β and circumvents the issues

introduced by missing values. Simulations show that including the variance structure in the model allows us to borrow information across the observations from multiple tissues to improve statistical power. The model is fitted through Markov chain Monte Carlo (MCMC) algorithm.

Past research has investigated Bayesian variable selection methods in various contexts, and many works use spike and slab prior as a widely accepted method. This prior models the regression coefficients as a mixture of a point mass at 0 and a normal distribution (George & McCulloch, 1997; Hernández-Lobato et al., 2013; Narisetty et al., 2014). There have been extensions to a multivariate linear regression as well (Brown et al., 1998; Lee et al., 2017). However, they tend to put a restrictive prior on the effect size. For example, some assume that the effect sizes follow the same variance structure as the response variable or that they come from an independent normal distribution with an arbitrarily fixed amount of variance. Meanwhile, Guan & Stephens (2011) under the context of univariate Bayesian variable selection, parametrizes the effect size using the concept of proportion of variance explained (PVE) so that the interpretation is direct and intuitive. Our method extends this prior into a multivariate version to allow more flexibility and better interpretation.

Also, most past works that attempt to analyze incomplete multivariate data have focused on imputations. There are methods to effectively impute the gene expression level matrix, one specifically for the missing values driven by tissue accessibility in GTEx, but not many methods attempt to analyze only the available data (Wang et al., 2016; Oba et al., 2003; Li et al., 2017). In this work, we propose a way to sensibly circumvent the missing value issue altogether by making an "Missing at Random" (MAR) assumption (Little & Rubin, 2014; Rubin, 1976). Sometimes, certain data-generating process drives the correlation between

the missing pattern and the unobserved data. For example, some gene expression levels are recorded as missing in RNA-seq because not enough reads have been mapped, in which case the missingness suggests that the underlying true value is low and the missing pattern holds relevant information about the data. However, the missing pattern in our case only depends on the tissue availability of the patients. This is mostly technical rather than biological. Possible factors include whether the sample is post-mortem or surgical, which surgery the subject went through, and which tissue is difficult to maintain fresh samples. Therefore, we assume MAR, meaning tissue availability holds no information regarding the underlying true gene expression level. This assumption simplifies the inference procedure even in the presence of missing values. The algorithm successfully estimates the covariance structure even with different numbers of observations for each phenotype.

The remainder of this article is organized as follows. In Section 2, we describe the Bayesian framework and our choice of priors. We discuss the computation and inference, including Markov chain Monte Carlo algorithm that works around the missing values. In Section 3, we present the simulation results that examine the effectiveness of our proposed method under various settings. We prove the algorithm's superiority to traditional linear analysis and show the algorithm's robustness to hyperparameter mis-specifications. In Section 4, we share the real data application of local ancestry and multi-tissue expression level. We conclude the article with a discussion in Section 5.

2. Bayesian Framework for Multivariate Data

This section introduces the notation and the details of the Bayesian normal regression model. It also explains our choices of prior distributions and hyperparameter specifications. The MCMC algorithm is presented along with the adjustments that account for the missing values.

2.1 *Model Formulation*

We model the effect of local ancestry $\mathbf{x}$ on gene expression $\mathbf{y}$, one gene at a time, assuming all other covariates have been accounted for. Each subject $i = 1, \cdots, n$ has a length T response vector $\mathbf{y_i} = (y_{i1}, \cdots, y_{iT})$ that contains more than 1 missing value. T represents the number of tissues. We consider a linear regression model with multivariate response $\mathbf{y_i}$ with a predictor vector $\mathbf{x} = (x_1, \cdots x_n)^T$, the coefficient parameter $\boldsymbol{\beta} = (\beta_1, \cdots, \beta_T)^T$, and the mean parameter $\boldsymbol{\mu} = (\mu_1, \cdots, \mu_T)^T$. We assume $\mathbf{y_i}$, including the unobserved values, follows a multivariate normal distribution, $\mathbf{y_i}|\boldsymbol{\beta}, \boldsymbol{\Sigma} \sim \mathcal{N}_T(\boldsymbol{\mu} + x_i\boldsymbol{\beta}, \boldsymbol{\Sigma})$, with a dense, unstructured covariance matrix $\Sigma \in \mathbb{R}^{T\times T}$. When $\mathbf{y_i}$ and $\mathbf{x}$ are centered, the calculation is simplified while posterior for $\boldsymbol{\beta}$ is unaffected, so we assume $\mathbf{y_i}$ and $\mathbf{x}$ are centered henceforth. The mean term disappears, and our final model is

$$\mathbf{y_i}|\boldsymbol{\beta}, \boldsymbol{\Sigma} \sim \mathcal{N}_T(x_i\boldsymbol{\beta}, \boldsymbol{\Sigma}). \tag{1}$$

2.2 *Prior Distributions*

We use "spike-and-slab" prior for the coefficients β as suggested by past literature for Bayesian variable selection (Mitchell & Beauchamp, 1988; George & McCulloch, 1993, 1997). The latent binary variable γ_t for $t = 1, \cdots, T$ indicates whether the variable t is included in the model. If $\gamma_t = 1$, β_t is non-zero and comes from the distribution $\mathcal{N}(0, \sigma_\beta^2)$, and if $\gamma_t = 0$, β_t comes from a point mass at 0.

$$\beta_t \sim \gamma_t\mathcal{N}(0, \sigma_\beta^2) + (1 - \gamma_t)\delta_0. \tag{2}$$

The latent variables γ_t independently follow Bernoulli distribution $\gamma_t \sim Ber(\pi)$, and $\pi \sim$ Be(a, b) where a and b are hyperparameters to be specified. For the algorithm's simplicity, we obtain the marginal prior of $\boldsymbol{\gamma}$ by integrating $p(\boldsymbol{\gamma}, \pi)$ over $p(\pi)$. The marginal prior only depends on the size of the vector: $|\boldsymbol{\gamma}| = \sum_{t=1}^{T} \gamma_t$.

$$p(\boldsymbol{\gamma}) = \frac{\Gamma(a + b)\Gamma(|\boldsymbol{\gamma}| + a)\Gamma(T + b - |\boldsymbol{\gamma}|)}{\Gamma(T + a + b)\Gamma(a)\Gamma(b)}. \tag{3}$$

We explain our choice of the hyperparameters a and b in the next section.

It is possible to model $\boldsymbol{\gamma}$ to have a non-trivial correlation structure. One natural way is to define a latent variable following a multivariate normal distribution with mean 0 and variance same as that of Y, and make an arbitrary threshold for each dimension to obtain correlated binary variables. However, this requires the computation of cumulative distribution of multivariate normal which does not have a closed form. Moreover, even when γ_t is independent a priori, the correlation structure inferred from the data will decide the selection of γ_t. Using an independent prior for γ_t is equivalent to only using the correlation information that comes from the data, and therefore we believe it is a valid choice.

Next, we model the covariance matrix Σ to follow the inverse Wishart prior distribution

$$\Sigma \sim W^{-1}(\nu, \nu\Phi). \tag{4}$$

which is conjugate to the multivariate normal variance. The posterior mean is a weighted average of the hyperparameter Φ and the empirical covariance matrix, and the weights are decided by the degrees of freedom ν and sample size n.

The set-up so far is quite standard and is backed up by past literature (George & McCulloch, 1993; Mitchell & Beauchamp, 1988). There have been various suggestions for the specification of σ_β^2 without a general consensus on a natural choice. Some use an arbitrary fixed number or an estimate of the coefficients from the data (Brown et al., 1998; Lee et al., 2017). Some attempt to put a prior on γ to make the model more flexible (Liang et al., 2008). Here, we choose an option that best aids the interpretation. Guan & Stephens (2011) focuses on what the prior implies about the proportion of variance in $\mathbf{y}$ explained by $\mathbf{x}$ (PVE) under a univariate setting for GWAS. Other priors have assumed the independence of $\boldsymbol{\gamma}$ and σ_β^2,

which implies that more complex models are expected to have higher PVE. However, both in GWAS and in gene expression level analysis, it seems plausible a priori that a simple model has a higher PVE and a complex model has a low PVE. For example, local ancestry can mainly drive the variation of a gene's expression level in one tissue but has no effect in other tissues. In this case, a large proportion of variance of $\mathbf{y}$ is explained by $\mathbf{x}$ although $|\boldsymbol{\gamma}|$ is only 1.

So we expand this to a multivariate version with correlated response variables. In univariate linear regression, PVE is defined as R^2, but it does not have a natural extension to the multivariate setting because there is no scalar representation of the covariance matrix. As an approximation, we use the trace of the covariance matrix. It has an intuitive interpretation of the amount of variance explained because the trace of a matrix is equal to the sum of its eigenvalues. For example, in the principal component analysis, a normalized eigenvalue is the proportion of variance explained by each principal component.

To formalize the multivariate PVE, let $V(\boldsymbol{\beta}) = \frac{1}{n} tr[(\mathbf{x}\boldsymbol{\beta}^T)^T(\mathbf{x}\boldsymbol{\beta}^T)]$ denote the trace of the empirical variance of $\mathbf{x}\boldsymbol{\beta}^T$. In the beginning of the paper, we centered both response and the covariate, so there is no need to consider the means. Then $\text{PVE}(\boldsymbol{\beta}) = \frac{V(\beta)}{V(\beta)+\text{tr}(\Sigma)}$. We define h as the approximation of the expectation of PVE.

$$h := \frac{E(V(\boldsymbol{\beta}))}{E(V(\boldsymbol{\beta})) + \text{tr}(\Sigma)} \approx E(PVE(\boldsymbol{\beta}) \mid \Sigma, \gamma, \sigma_\beta^2)$$

where

$$E(V(\boldsymbol{\beta}) \mid \Sigma, \sigma_\beta^2, \gamma) = \sum_{t:\gamma_t=1} \sigma_\beta^2 \sum_{i=1}^{n} \frac{x_i^2}{n}.$$

with expectation being taken over $\boldsymbol{\beta}$. Note that we approximate the expectation of a ratio as a ratio of expectations. This approximation is equivalent to approximating $1-\text{tr}(\Sigma)E\left(\frac{1}{V(\beta)+\text{tr}(\Sigma)}\right)$ as $1 - \text{tr}(\Sigma)\left(\frac{1}{E(V(\beta))+\text{tr}(\Sigma)}\right)$, and Jensen's inequality tells us that this formulation of h systematically overestimates the PVE. The error is 0 when $\sigma_\beta^2 = 0$ and the error grows

as σ_β^2 becomes larger. However, we don't expect σ_β^2 to be very large in our application, and we believe that this approximation works well as a proxy to PVE. Therefore, h can be represented in terms of σ_β^2.

$$h = \frac{\sum_{t:\gamma_t=1} \sigma_\beta^2 \sum_{i=1}^n \frac{x_i^2}{n}}{\sum_{t:\gamma_t=1} \sigma_\beta^2 \sum_{i=1}^n \frac{x_i^2}{n} + \mathrm{tr}(\Sigma)}.$$

Then we can parametrize σ_β^2 in terms of h, $\sigma_\beta^2(h, \boldsymbol{\gamma}, \Sigma) = \frac{h \cdot \mathrm{tr}(\Sigma)}{(\sum_t \gamma_t)(1-h)\sum_{i=1}^n \frac{x_i^2}{n}}$. When no variables are selected, $\sum_t \gamma_t$ is 0, and $\sigma_\beta^2(h, \mathbf{0}, \Sigma)$ becomes non-finite, so we add a "pseudo-count" in the denominator. Our final specification of σ_β^2 given h, $\boldsymbol{\gamma}$, and Σ is

$$\sigma_\beta^2(h, \boldsymbol{\gamma}, \Sigma) = \frac{h \cdot \mathrm{tr}(\Sigma)}{(\sum_t \gamma_t + 1)(1-h)\sum_{i=1}^n \frac{x_i^2}{n}}. \tag{5}$$

With these adjustments, we can parametrize the model with h instead of with σ_β^2 according to an almost non-informative prior that helps interpretation. We use a uniform prior on h, and the specification of its hyperparameters is discussed in the next section.

2.3 *Specifications of Hyperparameters*

The hyperparameters ν and Φ are easy to interpret because the posterior mean is a weighted average of the empirical covariance matrix $\hat{\Sigma}$ and Φ, and n and ν respectively decides each weight. Φ can be estimated empirically for the application of GTEx data. We average the empirical covariance matrices computed from the available gene expression measurements from $\sim 10,000$ genes to use as Φ, effectively leveraging information across many genes. This is from an assumption that tissue-tissue correlation is similar across many genes. For example, for any gene, the expression level from sun-exposed skin tissue will be more correlated to not-sun-exposed skin tissue than to a brain tissue. For other applications, standard choices such as $T \times T$ identity matrix are acceptable as well, as long as it is symmetric and positive definite.

Since we have a reasonable choice of Φ, we set the degrees of freedom ν to n, giving the

same weight to the prior and to the data. Although it is possible to use smaller ν to give minimal weight to the prior and allow flexibility in the choice of Φ, in the particular case of gene expression level of African Americans where n is small and data is highly incomplete, it is difficult to get a good estimate of the covariance structure, so we decide to give more weight to the prior.

Next, we fix a and b for the prior distribution of π that reflects the sparsity of the model. For a well-justified prior, we look at past eQTL analyses with GTEx data that studies multi-tissue gene expression level. Consortium et al. (2015) tested the SNPs for their effects on gene expression level in various tissues, and the result showed that much more SNPs were related to only 1 or all of the tissues than to a few tissues, showing a U-shaped pattern with respect to the number of tissues. Although the profiles involving only a few tissues have many more possible combinatorial patterns, eQTLs show high tissue specificity and tissue ubiquity. We expect similar behavior from the effects of local ancestry. We first expect that most of the genes will show signal in no tissues. For the rest of the genes, we expect many of them to show signals on either 1 or all T tissues.

Figure 1 shows the marginal prior of $\boldsymbol{\gamma}$ for different values of (a, b). When $a = b = 0.1$, $p(\boldsymbol{\gamma})$ is symmetrically U-shaped with the highest density at $|\boldsymbol{\gamma}| = 0$ and $|\boldsymbol{\gamma}| = T$, but it doesn't give particularly large weight to the null case. When $a = 0.1$ and $b = 5$, more weight is given to $|\boldsymbol{\gamma}| = 0$, but the graph is not U-shaped. When $a = 0.01$ and $b = 0.5$, the expected $|\boldsymbol{\gamma}|$ is same as before, around 90% of the weight is given to $|\boldsymbol{\gamma}| = 0$, and it also keeps the U-shape among the rest of the cases. We believe this reflects our prior belief about the effect of local ancestry on multi-tissue gene expression level, so we use this setting for the algorithm. Although this may seem very restrictive compared to a non-informative prior of $\pi \sim \text{Be}(1, 1)$,

given that we are testing the 30,000 genes separately, we believe a more conservative prior is appropriate for a reasonable error control.

We next choose the prior for h. Guan & Stephens (2011) uses non-informative uniform prior Unif$(0, 1)$ on h, but h near the boundary of this support can be problematic in the multivariate context. For example, when $h = 0$, σ_β^2 becomes 0, and the algorithm would no longer add any variables. Also, when h is small and σ_β^2 is too close to zero, the normal distribution $N(0, \sigma_\beta^2)$ does not have much discriminating power from the point mass at 0, disabling the spike-and-slab prior of $\boldsymbol{\beta}$ (George & McCulloch, 1997). On the other hand, when h becomes close to 1, the denominator becomes close to 0. Moreover, PVE value close to 1 is unrealistic in most biological applications. To account for these boundary cases, we use Unif$(0.1, 0.9)$ as the prior for h. Having a lower bound on h effectively puts an appropriate lower bound on σ_β^2, and restricting the range of h by putting an upper bound can decrease the search space and can expedite the algorithm while reflecting our belief that local ancestry explains less than 90% of the variance of gene expression level. In other applications, the upper bound can be extended to higher values, as long as it is strictly less than 1.

[Figure 1 about here.]

2.4 *MCMC algorithm*

The Markov-chain Monte Carlo algorithm is based on the following factorization of the joint model, $p(y|\beta, \Sigma)p(\beta|\sigma_\beta, \gamma)p(\gamma|\pi)p(\pi)p(\sigma_\beta^2)p(\Sigma)$. Replacing $p(\sigma_\beta)$ with $p(h)$ and integrating out γ leads to the following form

$$\prod_{i=1}^{n} p(\mathbf{y_i}|\boldsymbol{\beta}, \Sigma)p(\boldsymbol{\beta}|h, \boldsymbol{\gamma})p(\boldsymbol{\gamma})p(h)p(\Sigma). \tag{6}$$

This is equivalent to the product of the likelihood, prior for $\boldsymbol{\beta}$, prior for $\boldsymbol{\gamma}$, prior for h, and prior for Σ. This serves as the target distribution in our MCMC algorithm. We initialize $\boldsymbol{\gamma}^{(0)}$ and $\boldsymbol{\beta}^{(0)}$ as $\mathbf{0}$, $\Sigma^{(0)}$ as Φ, and the $\sigma_\beta^{2(0)}$ as 1. At each iteration j, we repeat the following steps

of updating $\boldsymbol{\gamma}$ and $\boldsymbol{\beta}$, updating Σ, and updating h. The details below are for jth iteration.

We first update $\boldsymbol{\beta}^{(j)}$ and $\boldsymbol{\gamma}^{(j)}$ simultaneously using the Metropolis-Hastings algorithm (MH) given fixed $\Sigma^{(j-1)}$ and $\sigma_\beta^{2(j-1)}$ (Robert & Casella, 2013; Hastings, 1970). We assign with $\boldsymbol{\beta}^{(j)} = \boldsymbol{\beta}^{(j-1)}$ and $\boldsymbol{\gamma}^{(j)} = \boldsymbol{\gamma}^{(j-1)}$, propose $\boldsymbol{\beta}^*$ and $\boldsymbol{\gamma}^*$ 500 times, and accept the proposals whenver appropriate. We propose $\boldsymbol{\gamma}^*$ by changing the status of one variable from the values of $\boldsymbol{\gamma}^{(j)}$. Each variable $t = 1, \cdots, T$ has an equal chance $(1/T)$ of being selected. If the picked variable t is already in the model $\left(\gamma_t^{(j)} = 1\right)$, then proposed value γ_t^* is 0. If the variable t is not in the model $\left(\gamma_t^{(j)} = 0\right)$, then proposed value γ_t^* is 1.

Based on $\boldsymbol{\gamma}^*$, we draw β_t^* from for all t such that $\gamma_t^* = 1$. We need to pre-fix u_β^2 and w_β^2 as prespecified variances of the proposal density Lee et al. (2017). The β_t of the newly added variable t is drawn from $N(0, u_\beta^2)$. The variables that are already in the model and are not removed are drawn from $N(\boldsymbol{\beta}^{(j)}_{t:\gamma_t^{(j)}=1,\gamma_t^*=1}, w_\beta^2)$. The reason we use two separate variances for the update is to expedite the mixing and convergence of the algorithm. u_β^2 reflects an approximate size of σ_β^2 as it determines the initial guess of the effect size of a variable. Empirically, we use the variance of linear regression coefficients for the variables with p-values less than 0.05. One might suggest using $\sigma_\beta^{2(j-1)}$ for u_β^2, but this unnecessarily slows down the convergence because the posterior of σ_β^2 has high variance. w_β^2 reflects the size of perturbation that fine-tunes the coefficients that are already in the model, and around 1/100 of u_β^2 works well.

With the proposed $\boldsymbol{\gamma}^*$ and $\boldsymbol{\beta}^*$, compute the acceptance probability as the product of the likelihood ratio, prior ratio for $\boldsymbol{\beta}$, prior ratio for $\boldsymbol{\gamma}$, and proposal ratio. Note that the prior

for Σ is canceled out. First, the likelihood ratio is,

$$\begin{aligned} L &= \left(\frac{\prod_{i=1}^{n} f(\mathbf{y_i} \mid \boldsymbol{\beta}^*, \Sigma^{(j-1)})}{\prod_{i=1}^{n} f(\mathbf{y_i} \mid \boldsymbol{\beta}^{(j)}, \Sigma^{(j-1)})} \right) \\ &= \frac{\prod_i \exp\left(-\frac{1}{2}(\mathbf{y_i} - x_i\boldsymbol{\beta}^*)^T \Sigma^{-1(j-1)} (\mathbf{y_i} - x_i\boldsymbol{\beta}^*)\right)}{\prod_i \exp\left(-\frac{1}{2}(\mathbf{y_i} - x_i\boldsymbol{\beta}^{(j)})^T \Sigma^{-1(j-1)} (\mathbf{y_i} - x_i\boldsymbol{\beta}^{(j)})\right)} \end{aligned} \tag{7}$$

Note that since $\mathbf{y_i}$ has missing values, we need some adjustments in computing L. We discuss this in detail in the next section. The prior ratio for $\boldsymbol{\beta}$ is

$$B = \frac{p(\boldsymbol{\beta}^*)}{p(\boldsymbol{\beta}^{(j)})} = \frac{\prod_{t:\gamma_t^*=1} \exp\left(-\frac{\beta_t^{2*}}{2\sigma_\beta^{2(j-1)}}\right)}{\prod_{t:\gamma_t^{(j-1)}=1} \exp\left(-\frac{\beta_t^{2(j)}}{2\sigma_\beta^{2(j-1)}}\right)}$$

The prior ratio for $\boldsymbol{\gamma}$ is

$$G = \frac{p(\boldsymbol{\gamma}^*)}{p(\boldsymbol{\gamma}^{(j)})} = \frac{\Gamma(a + |\boldsymbol{\gamma}^*|)\Gamma(T + b - |\boldsymbol{\gamma}^*|)}{\Gamma(a + |\boldsymbol{\gamma}^{(j)}|)\Gamma(T + b - |\boldsymbol{\gamma}^{(j)}|)}.$$

The proposal ratio when adding a variable is like below. First, we define g as the number of non-zero values in $\gamma^{(j-1)}$, and t as the index of the variable that we are adding ($\gamma_t^{(j)} = 0, \gamma_t^* = 1$). q_1 is the proposal distribution of $\boldsymbol{\beta}$ (Hastings, 1970).

$$\begin{aligned} P_{\text{add}} &= \frac{1/(g+1)}{1/(T-g)} \times \frac{q_1(\boldsymbol{\beta}^* \to \boldsymbol{\beta}^{(j)} \mid \boldsymbol{\gamma}^*)}{q_1(\boldsymbol{\beta}^{(j)}) \to \boldsymbol{\beta}^* \mid \boldsymbol{\gamma}^{(j-1)})} \\ &= \frac{1/(g+1)}{1/(T-g)} \times \frac{1}{\frac{1}{\sqrt{2\pi u_\beta^2}} \exp(-\frac{\beta_t^{*2}}{2u_\beta^2})} \end{aligned}$$

Similarly, the proposal ratio when deleting a variable t is ($\gamma_t^{(j)} = 1$, $\gamma_t^* = 0$)

$$P_{\text{delete}} = \frac{1/(T-g+1)}{1/g} \times \frac{\frac{1}{\sqrt{2\pi u_\beta^2}} \exp\left(-\frac{\beta_t^{(j)2}}{2u_\beta^2}\right)}{1}$$

To summarize, the acceptance probability is $\min(A, 1)$ where A is either $L \times B \times P_{\text{add}}$ or $L \times B \times P_{\text{delete}}$. With probability of $\min(A, 1)$, we make an update $\beta^{(j)} = \beta^*$ and $\gamma^{(j)} = \gamma^*$. Otherwise, we propose another set of $\boldsymbol{\beta}^*$ and $\boldsymbol{\gamma}^*$ values without changing $\boldsymbol{\beta}^{(j)}$ or $\boldsymbol{\gamma}^{(j)}$. We repeat the process until we exhaust 500 different proposals.

Then we update $\Sigma^{(j)}$ given $\boldsymbol{\beta}^{(j)}$ and $\boldsymbol{\gamma}^{(j)}$, and $h^{(j-1)}$ by drawing from the posterior distribution. Since inverse Wishart distribution is the conjugate prior for multivariate normal

variance, we can easily get the closed form posterior for $\Sigma^{(j)}$:

$$p(\Sigma^{(j)}|\boldsymbol{\beta}^{(j)}, h^{(j-1)}, \boldsymbol{\gamma}^{(j)}, \mathbf{y}) \\ = W^{-1}\left(n+\nu, \sum_{i=1}^{n}(\mathbf{y_i} - x_i\boldsymbol{\beta}^{(j)})(\mathbf{y_i} - x_i\boldsymbol{\beta}^{(j)})^T + \nu\Phi\right) \quad (8)$$

where $\frac{1}{n}\sum_i(\mathbf{y_i} - x_i\boldsymbol{\beta}^{(j)})(\mathbf{y_i} - x_i\boldsymbol{\beta}^{(j)})^T$ is the MLE of the covariance matrix of the data. This adds complication because $\mathbf{y_i}$ is not a complete vector. Next section about missing data explains the adjustment to this posterior in detail.

Next, we update $h^{(j)}$ through Metropolis-Hastings algorithm. First specify $h^{(j)} = h^{(j-1)}$ and test for 100 proposals $h^* = h^{(j)} + \delta$ where δ is randomly drawn from Unif$(-0.1, 0.1)$. h^* is reflected around the boundary of the support $(0.05, 0.9)$. We compute σ^*_β from the proposed h^* and calculate thacceptance probability. The updates are symmetric, so proposal probability is ignored. Then the acceptance probability is $min(C, 1)$ where

$$C = \frac{p(\boldsymbol{\beta}^{(j)} \mid \sigma^{*2}_\beta)}{p(\boldsymbol{\beta}^{(j)} \mid \sigma^{2(j)}_\beta)} = \frac{\prod_{t:\gamma_t^{(j)}=1}\exp\left(-\frac{\beta_t^{2(j)}}{2\sigma^{2*}_\beta}\right)}{\prod_{t:\gamma_t^{(j)}=1}\exp\left(-\frac{\beta_t^{2(j)}}{2\sigma^{2(j)}_\beta}\right)}$$

With probability of $min(C, 1)$ update $h^{(j)} = h^*$ and $\sigma^{(j)}_\beta = \sigma^*_\beta$. Otherwise, we propose new h^* without changing $h^{(j)}$. We repeat the process until we exhaust 100 proposals.

2.5 *Accounting for Missing Data*

The multi-tissue expression level data from GTEx has many missing values, and the proposed computational algorithm is not feasible if the data is incomplete. The two main challenges are computing the acceptance probability in the Metropolis-Hastings algorithm and computing the empirical covariance matrix for updating Σ. In this section, we make certain assumptions about the missing pattern of the data and propose a way to work around the missing values.

Past works that attempt to analyze GTEx's multi-tissue gene expression level matrix

have focused on imputations. Many of them also focused only on some of the tissues that have plenty of observations (Li et al., 2017; Consortium et al., 2015). Here, we adopt a classical approach by modeling M (Rubin, 1976), a binary random variable indicating data availability, and use all the available tissues even with less than 10 observations.

We define $M = (M_1, \cdots, M_n)$ sa a matrix random variable of missing data indicator. Each M_i is a length T vector with values of 0 or 1, indicating tissue availability for individual i. The probability that M takes the value $\boldsymbol{m} = (\boldsymbol{m_1}, \cdots, \boldsymbol{m_n})$ given $\mathbf{y} = (\mathbf{y_1}, \cdots, \mathbf{y_n})$ is $g(\boldsymbol{m}|\mathbf{y})$. As mentioned in the introduction, we assume that the tissue availability holds no information regarding the gene expression level, either observed or unobserved. This condition where the missing pattern is independent of the underlying true values is called missing at random (MAR). Under the MAR assumption, $g(\boldsymbol{m_i} \mid \mathbf{y_i}) = g(\boldsymbol{m_i})$ takes the same value for all $\mathbf{y_i}$, and this allows simpler analysis of incomplete data (Rubin, 1976). For notational convenience, consider a separation of $\mathbf{y_i}$ into the observed part $\mathbf{y_{io}}$ and the missing part $\mathbf{y_{im}}$.

One challenge of the current version of the algorithm is the computation of the acceptance probability, especially the likelihood ratio (7), when we update $\boldsymbol{\beta}$ and $\boldsymbol{\gamma}$ with conditioning on fixed Σ. The likelihood $\prod_{i=1}^{n} L(\boldsymbol{\beta}|\mathbf{y_i}, \Sigma) = \prod_{i=1}^{n} f(\mathbf{y_i} \mid \boldsymbol{\beta}, \Sigma)$ cannot be computed when $\mathbf{y_i}$ is not a complete vector. The full posterior distribution of the parameter $\boldsymbol{\beta}$ accounting for $\boldsymbol{m}$ is proportional to

$$p(\boldsymbol{\beta}) \prod_{i=1}^{n} \int f(\mathbf{y_i} \mid \boldsymbol{\beta}, \Sigma) g(\boldsymbol{m_i}) d\mathbf{y_{im}}$$

and, under MAR, this is equivalent to

$$c \cdot p(\boldsymbol{\beta}) \prod_{i=1}^{n} \int f(\mathbf{y_i} \mid \boldsymbol{\beta}, \Sigma) d\mathbf{y_{im}}$$

where c is some constant that is canceled out in likelihood ratio. In our context, $f(\mathbf{y_i} \mid \boldsymbol{\beta}, \Sigma)$ is multivariate normal $N_T(\mathbf{y_i}; \boldsymbol{\beta}, \Sigma)$, and so $\int f(\mathbf{y_i} \mid \boldsymbol{\beta}) d\mathbf{y_{im}}$ is equivalent to the marginal

density of multivariate normal $N_T(\mathbf{y_{io}}; \boldsymbol{\beta_{i,o}}, \Sigma_{i,o})$. $\boldsymbol{\beta}_{i,o}$ is the subvector of coefficient $\boldsymbol{\beta}$ only at the observed index of individual i, and $\Sigma_{i,o}$ is similarly the submatrix of covariance matrix Σ with rows and columns indexed at the observed part of individual i. This shows that replacing the full joint likelihood with the marginal likelihood does not influence the posterior of our parameters of interest (Rubin, 1976), and therefore the likelihood ratio in (7) becomes

$$\frac{\prod_i \exp\left(-\frac{1}{2}(\mathbf{y_{io}} - x_i\boldsymbol{\beta}^*_{\boldsymbol{io}})^T\Sigma_{io}^{-1(j-1)}(\mathbf{y_{io}} - x_i\boldsymbol{\beta}^*_{\boldsymbol{io}})\right)}{\prod_i \exp\left(-\frac{1}{2}(\mathbf{y_{io}} - x_i\boldsymbol{\beta}^{(j)}_{\boldsymbol{io}})^T\Sigma_{io}^{-1(j-1)}(\mathbf{y_{io}} - x_i\boldsymbol{\beta}^{(j)}_{\boldsymbol{io}})\right)}$$

Another challenge is the empirical covariance matrix required to update Σ with conditioning on fixed $\boldsymbol{\beta}$, and we can apply the same process. The full posterior for Σ is proportional to

$$p(\Sigma)\prod_{i=1}^{n}\int f(\mathbf{y_i} \mid \Sigma)g(m_i \mid \mathbf{y_{io}})d\mathbf{y_{im}}$$
$$= c \cdot p(\Sigma)\prod_{i=1}^{n}\int f(\mathbf{y_i} \mid \Sigma)d\mathbf{y_{im}}$$

The likelihood $\int f(\mathbf{y_i} \mid \Sigma)d\mathbf{y_{im}}$ is no longer a function of the full matrix Σ but rather a submatrices $\Sigma_{i,o}$ for $i = 1, \cdots, n$, and it is impossible to obtain a closed form posterior. We propose to use the EM algorithm to estimate the MLE $\hat{\Sigma}$ and maintain the posterior formula (8). We find MLE $\hat{\Sigma}$ by solving the following optimization function through the EM algorithm, whose details can be found in (Little & Rubin, 2014).

$$\arg\max_{\Sigma} -\frac{1}{2}\sum_{i=1}^{n} log|\Sigma_{i,o}| - \frac{\sum_{i=1}^{n}(\mathbf{y_i} - x_i\boldsymbol{\beta}_{i,o})^T\Sigma_{i,o}^{-1}(\mathbf{y_i} - x_i\boldsymbol{\beta}_{i,o})}{2}. \tag{9}$$

The first obvious benefit of this approximation is the algorithm's simplicity. There is no intuitive proposal distribution for the covariance matrix to implement the Metropolis-Hastings algorithm, and especially when T is large, calculating the acceptance probability for a large number of covariance matrices can be computationally burdensome. Another benefit is that this EM algorithm can return a valid result even when two variables share no common subjects. For example, there is no subject that has observations both in Testis and Uterus tissues, but the MLE $\hat{\Sigma}$ can return a valid correlation value between the two

using other variables, while the posterior mean in (9) only receives information from the prior at such indices. Also, simulations show this update procedure does not interfere with the correct inference on $\boldsymbol{\gamma}$.

3. Simulation Studies

We evaluate the proposed method's performance on data sets simulated under multiple settings. The results show that the algorithm performs well even in difficult settings with low sample size, high rate of missing values, and weak signals, all of which are expected in the GTEx data. We also demonstrate that the method is robust to hyperparameter misspecification, and that posterior inclusion probability (PIP) is well calibrated in that variables with higher PIP has higher proportions of true positives. The simulation results show that the proposed method improves the statistical power compared to the univariate linear regression that assumes independence among the outcomes.

3.1 *Settings*

We construct data sets to resemble the real GTEx data. For sample size, we use $n = 71$ which is the number of African American samples and $T = 24$ which matches the number of tissues in GTEx data with available expression levels from more than 20 subjects. The missing pattern is inherited from the tissue availability of the real data. Some outcomes have more missing values than others, and on average, around 53% of the entries are missing. The covariate $\mathbf{x}$ comes from one of the gene's local ancestry data where age, sex, and global ancestry have been regressed out. We fix Σ with 1 at the diagonals and 0.5 elsewhere. Fixed error level creates a consistent environment so that we can observe how the power varies with the effect sizes.

We generate the data sets as follows. For each simulation, we first draw π from the specified

Beta distribution, and then draw $\gamma_t|\pi \sim \text{Ber}(\pi)$ for $t = 1, \cdots, T$. Then according to $\boldsymbol{\gamma}$ we draw $\beta_t|(\gamma_t = 1) \sim N(0, \sigma_\beta^2)$, or $\beta_t|(\gamma_t = 0) = 0$. Lastly, we draw error $\epsilon \sim N_T(\mathbf{0}, \Sigma)$ and construct $Y = \mathbf{x}\boldsymbol{\beta}^T + \epsilon$ with fixed $\mathbf{x}$. We use two effect sizes, $\sigma_\beta^2 = 5$ and $\sigma_\beta^2 = 1$. We suspect that the scenario with small effect sizes resemble the real data application. We also use two distributions $\pi \sim \text{Be}(0.05, 0.5)$ and $\pi \sim \text{Be}(1, 1)$ to see how parameter misspecification affects the performance of the algorithm. The simulation scenarios are summarized in Table 1, and we generate 500 different Y for each scenario to run the algorithm with constant hyperparameter specifications. Note that for the real data application, we use a more conservative prior $a = 0.01, b = 0.5$. However, drawing $\pi \sim \text{Be}(0.01, 0.5)$ creates too few true signals, making it difficult to examine the algorithm's performance. So we deliberately create more true signals for our simulations by using a more liberal prior for π.

[Table 1 about here.]

3.2 *Results*

We first define some of the terms that are used to analyze the result of the algorithm. We observe at each iteration $\hat{\gamma}_{jst}$ and $\hat{\beta}_{jst}$ where $j = 1, \cdots, J$ is the iteration index after removing burn-in, $s = 1, \cdots, S$ is simulation index, and $t = 1, \cdots, T$ is the outcome index. We define posterior inclusion probability (PIP) as

$$p(\hat{\gamma}_{st} = 1) = \frac{\sum_{j=1}^{J} \hat{\gamma}_{sjt}}{J}.$$

We use the following definitions to analyze Type I and Type II errors given the PIP threshold c.

- ·True Positive : $p(\hat{\gamma}_{st} = 1) \geqslant c$ and $\gamma_{st} = 1$
- ·False Positive : $p(\hat{\gamma}_{st} = 1) \geqslant c$ and $\gamma_{st} = 0$
- ·True Negative : $p(\hat{\gamma}_{st} = 0) \geqslant c$ and $\gamma_{st} = 0$
- ·True Negative : $p(\hat{\gamma}_{st} = 0) \geqslant c$ and $\gamma_{st} = 1$

We also define FDR and power given c. Although these concepts are fundamentally frequentist, they are useful when we compare the result with the univariate linear regression. We reject the null $\gamma_{st} = 0$ if PIP $\geqslant c$, and not reject the null if PIP $< c$.

$$\text{FDR}_c = \frac{\sum_{s,t} \mathbf{1}\{p(\hat{\gamma}_{st} = 1) \geqslant c \text{ and } \gamma_{st} = 0\}}{\sum_{s,t} \mathbf{1}\{p(\hat{\gamma}_{st} = 1) \geqslant c\}}$$

$$\text{Power}_c = \frac{\sum_{s,t} \mathbf{1}\{p(\hat{\gamma}_{st} = 1) \geqslant c \text{ and } \gamma_{st} = 1\}}{\sum_{s,t} \mathbf{1}\{\gamma_{st} = 1\}}$$

We also define posterior mean of β,

$$\bar{\beta}_{st} = \sum_{j:\hat{\gamma}_{sjt}=1} \frac{\hat{\beta}_{sjt}}{\sum_j \hat{\gamma}_{sjt}}$$

to check the algorithm's performance on the estimation of the effect size.

We first examine the number of false and true positives for varying PIP threshold c. Figure 2 (a) shows that all scenarios show consistent behaviors. Scenarios 1 and 2 has more false discovery rate at low 0 because π is drawn from $\text{Be}(0.05, 0.5)$ and there are not that many true signals in the data. When the effect sizes are small ($\sigma_\beta^2 = 1$, scenarios 2 and 4), power decreases more quickly as c increases. Even when hyperparameters are mis-specified ($\pi \sim \text{Be}(1, 1)$, secnarios 3 and 4), the algorithm performs well.

To evaluate the inference on β, for each simulation s and variable t, we compute the posterior mean and plot it against the true value in Figure 2 (b). For each scenario, we compute the FDR and power as defined in the previous section, and use PIP threshold c where FDR_c reaches 0.05. The red points are the ones not selected by the algorithm($p(\hat{\gamma}_{st} = 1) < c$), and the blue points selected ($p(\hat{\gamma}_{st} = 1) \geqslant c$). We also divide the true β values into bins and investigate the change in power in Table 2. PIP increases as the effect size increases, proving the calibration of PIP for variable selection.

Figure 2 (c) shows the power improvement compared to the univariate analysis. For the marginal linear regression, we record the $-log_{10}(p)$ values for the $24 \times 500 = 12000$ variables to test the null hypothesis $\beta = 0$. We discretize the log-transformed p-value threshold and compute FDR and power just as we do with posterior inclusion probability thresholds. Then we match the FDR level with the multivariate result to create Figure 2 (d) plot. The power of the proposed method is consistently higher than that of the marginal result when FDR level is fixed.

Figure 2 (d) shows the calibration of PIP as a selection criterion. We divide PIP into 5 bins and compute the mean of PIP and the proportion of true positives for each bin for each scenario. The higher the PIP, the higher the proportion of true positives. This means that we can decide on a PIP threshold to effectively control for type I error. Scenarios 1 and 2 show more inconsistent pattern compared to scenarios 3 and 4, and this is simply due to the size of true positives in the simulations. When π is drawn from $\text{Be}(0.05, 0.5)$, only around 20% of the variables are non-zero and they're divided into 10 bins. Especially in scenario 2, since the effect sizes are small ($\sigma_\beta^2 = 1$), a very small number of variables are placed into bins with PIP greater than 0.5.

We also run a null simulation where $\pi = 0$ and the rest of the data generating process is the same. This is designed to check the algorithm's susceptibility to false positives. The result returned no variables with PIP higher than 0.95, and only one variable returned PIP higher than 0.9 out of $500 \times 24 = 12,000$ variables. This shows that the the algorithm is quite robust to false positives, and we believe 0.95 is a conservative enough threshold that can effectively control the error.

[Figure 2 about here.]

[Table 2 about here.]

4. Application to Local Ancestry and Gene Expression Level Data

As mentioned, the proposed method is motivated by the multi-tissue gene expression level data. We aim to discover the genes whose expression levels are affected by the local or global ancestry.

4.1 *Data*

We study the data from GTEx V6p release (Consortium et al., 2015). In order to select African Americans from the available samples, we first infer the local ancestry of the samples who identified themselves as European Americans or African Americans and verify that their genetic ancestry is consistent. For local ancestry inference, we use the software LAMP that reaches as high as 98% accuracy level for distinguishing YRI and CEU ancestry (Paşaniuc et al., 2009). The software requires the user inputs of the genotypes of the admixed population, chromosomal positions of the SNPs, and the reference minor allele frequency of pure population CEU (Utah residents with Northern and Western European ancestry) and YRI (Yoruba in Ibadan, Nigeria). The genotype data of GTEx was obtained using Illumina OMNI 5M SNP array, and the rest of the SNPs have been imputed using 1000 Genomes Project Phase I, version 3. After imputation, genotypes were filtered using call-rate threshold 95% and info score threshold 0.4. Only the SNPs with minor allele frequency $> 5\%$ are included in the analysis. The reference minor allele frequency was obtained from the 1000 Genome Project (1000GP). We use 7 for the number of generations of admixture, 0.2 and 0.8 for the initial proportion of CEU and YRI population, and 10^{-8} for recombination rate, but the results are robust to these initial parameters.

LAMP returns local ancestry at each SNP as the count of African chromosomes (0, 1, or 2) at each locus, and we use the SNP closest to the center of the gene to represent the local ancestry of the entire gene. Around 92% of the genes showed no recombination event in all of the subjects, and less than 3% of the genes have more than one individual with ancestry switch within the gene, so we believe this is a valid approximation.

We define global ancestry as a value between 0 and 1 that quantifies the proportion of African chromosome in each subject. We first estimate it by averaging the inferred local ancestry, and this estimate is cross-checked with principal component analysis which can effectively cluster the subjects into subpopulations (Pritchard et al., 2000). We also include pure YRI and CEU population for PCA, and most African Americans lie strictly between the YRI and CEU population showing a two-way admixture between pure Europeans and pure Africans. We observed some outliers and so we remoevd them. We also observed some self-identified Europeans whose genetic ancestry is more than 10% African, and we include them in our analysis.

The expression levels provided by GTEx were measured using RNA-seq for 38,498 genes in the autosomal chromosomes. For each tissue, only genes with RPKM higher than 0.1 were included. Then the expression levels are normalized, log-transformed, and corrected for technical artifacts by GTEx. In addition, for each tissue, we regressed out the first two principal components of the expression level matrix to remove the effects of the technical noise confounded with the measurements of each individual. Different subjects were sequenced for different sets of tissues, introducing a number of missing values in the expression level matrices. Lastly, we didn't analyze genes that were expressed in only one tissue, because

linear regression suffices to analyze them.

In summary, we study the expression levels of 32,006 genes from 71 African American individuals with reliable measurements of local and global ancestry.

4.2 *Multivariate Analysis*

We use the proposed method to test the effects of both local and global ancestry for 32,006 genes. These genes were expressed in at least 2 tissues. We use the hyperparameters (Φ, ν, a, b) as specified in Section 2.3. Based on the simulations, we use the PIP threshold 0.95. The results are summarized in Table 3 and Table 4.

We also compare the result with simple linear regression that assumes inter-tissue independence of the expression levels and analyze the data separately for each tissue t. We used the same demographical covariates including the two principal components of the expression level. We do not find any signal that stood out when we use FDR threshold 0.05 with Benjamini Hochberg procedure (Benjamini & Hochberg, 1995).

[Table 3 about here.]

[Table 4 about here.]

5. Discussion

We have developed a Bayesian variable selection method that can explain the relationship between a covariate and correlated multiple phenotypes. The non-informative prior for the proportion of variance explained (PVE) aids the interpretation, and the algorithm can also analyze highly incomplete data. This method allows us to analyze multi-tissue expression level data against a covariate, for example local ancestry, and it has a wide range of other

possible applications.

The simulation section shows that the proposed method works better than the linear regression that assumes independence across the tissues, especially in scenario 2 where the signals are scarce ($\pi \sim \text{Be}(0.05, 0.5)$) and small ($\sigma_\beta^2 = 1$). In recent challenges in biology, single variable rarely explains a significant portion of trait variability, and it is common to search for weak and sparse signals, in which our proposed method shows an advantage.

The algorithm could take a few possible other directions. First, as briefly mentioned, we could use a prior $\boldsymbol{\gamma}$ that allows correlation. However, this can induce false confidence in the selection of $\boldsymbol{\gamma}$ compared to relying on the data to infer correlation. It can also impose more computational burden to the algorithm. Second, it is possible to expand this algorithm to consider multiple covariates simultaneously. However, it is common to focus on one explanatory variable, and it is straightforward to regress out other covariates beforehand.

One thing this algorithm lacks is an effective error control which is especially difficult when the number of tests is as large as the number of genes. The simulation shows that the empirical FDR reaches 0.05 at approximately PIP=0.5 threshold. However, FDR is a frequentist concept, and it is difficult to apply it to the Bayesian variable selection problem. The null simulation of 500 data sets with 24 variables each returns 1 case with PIP higher than 0.95, and this multiplies fast as the number of tests increases to more than 30,000 genes. It is possible to use a much more conservative prior, but it would violate our assumption that some genes have many cross-tissue signals. It is also possible to pre-select some of the genes to reduce the number of tests, but then we will have to consider the selection bias. Still, our method allows us to observe the top genes with the strongest signals, and it gives

valuable biological insights to better understand African American genome and the effects of genetic ancestry.

References

Benjamini, Y. & Hochberg, Y. (1995): Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the royal statistical society. Series B (Methodological)* pages 289–300.

Brown, P. J.; Vannucci, M. & Fearn, T. (1998): Multivariate Bayesian variable selection and prediction. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **60**(3):627–641.

Consortium, G. et al. (2015): The Genotype-Tissue Expression (GTEx) pilot analysis: multitissue gene regulation in humans. *Science* **348**(6235):648–660.

Fagny, M.; Paulson, J. N.; Kuijjer, M. L.; Sonawane, A. R.; Chen, C.-Y.; Lopes-Ramos, C. M.; Glass, K.; Quackenbush, J. & Platig, J. (2017): Exploring regulation in tissues with eQTL networks. *Proceedings of the National Academy of Sciences* **114**(37):E7841–E7850.

George, E. I. & McCulloch, R. E. (1993): Variable selection via Gibbs sampling. *Journal of the American Statistical Association* **88**(423):881–889.

George, E. I. & McCulloch, R. E. (1997): Approaches for Bayesian variable selection. *Statistica sinica* pages 339–373.

Guan, Y. & Stephens, M. (2011): Bayesian variable selection regression for genome-wide association studies and other large-scale problems. *The Annals of Applied Statistics* pages 1780–1815.

Hastings, W. K. (1970): Monte Carlo sampling methods using Markov chains and their applications .

Hernández-Lobato, D.; Hernández-Lobato, J. M. & Dupont, P. (2013): Generalized spike-

and-slab priors for Bayesian group feature selection using expectation propagation. *The Journal of Machine Learning Research* **14**(1):1891–1945.

Lee, K. H.; Tadesse, M. G.; Baccarelli, A. A.; Schwartz, J. & Coull, B. A. (2017): Multivariate Bayesian variable selection exploiting dependence structure among outcomes: Application to air pollution effects on DNA methylation. *Biometrics* **73**(1):232–241.

Li, G.; Shabalin, A. A.; Rusyn, I.; Wright, F. A. & Nobel, A. B. (2017): An empirical Bayes approach for multiple tissue eQTL analysis. *Biostatistics* **19**(3):391–406.

Liang, F.; Paulo, R.; Molina, G.; Clyde, M. A. & Berger, J. O. (2008): Mixtures of g priors for Bayesian variable selection. *Journal of the American Statistical Association* **103**(481):410–423.

Little, R. J. & Rubin, D. B. (2014): *Statistical analysis with missing data*, volume 333. John Wiley & Sons.

Mitchell, T. J. & Beauchamp, J. J. (1988): Bayesian variable selection in linear regression. *Journal of the American Statistical Association* **83**(404):1023–1032.

Narisetty, N. N.; He, X. et al. (2014): Bayesian variable selection with shrinking and diffusing priors. *The Annals of Statistics* **42**(2):789–817.

Oba, S.; Sato, M.-a.; Takemasa, I.; Monden, M.; Matsubara, K.-i. & Ishii, S. (2003): A Bayesian missing value estimation method for gene expression profile data. *Bioinformatics* **19**(16):2088–2096.

Paşaniuc, B.; Sankararaman, S.; Kimmel, G. & Halperin, E. (2009): Inference of locus-specific ancestry in closely related populations (WINPOP). *Bioinformatics* **25**(12):i213–i221.

Pritchard, J. K.; Stephens, M. & Donnelly, P. (2000): Inference of population structure using multilocus genotype data. *Genetics* **155**(2):945–959.

Robert, C. & Casella, G. (2013): *Monte Carlo statistical methods*. Springer Science & Business Media.

Rubin, D. B. (1976): Inference and missing data. *Biometrika* **63**(3):581–592.

Wang, J.; Gamazon, E. R.; Pierce, B. L.; Stranger, B. E.; Im, H. K.; Gibbons, R. D.; Cox, N. J.; Nicolae, D. L. & Chen, L. S. (2016): Imputing gene expression in uncollected tissues within and beyond GTEx. *The American Journal of Human Genetics* **98**(4):697–708.

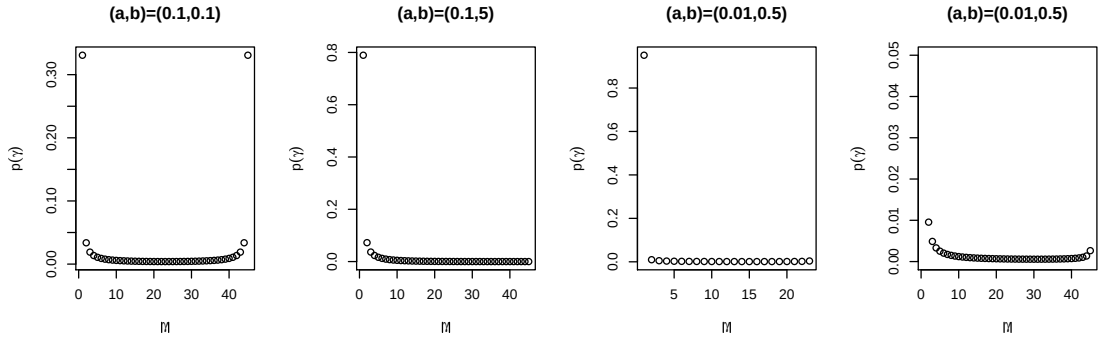

Figure 1. Marginal distribution of $\boldsymbol{\gamma}$ for different hyperparameter settings. The y-axis shows the probability mass of $\boldsymbol{\gamma}$ with given $|\boldsymbol{\gamma}|$. When $a = 0.01$ and $b = 0.5$, $p(\boldsymbol{\gamma})$ puts around 90% of the weight on the null ($\boldsymbol{\gamma} = \mathbf{0}$) and distributes the rest of the weight on the rest with a rough U-shape.

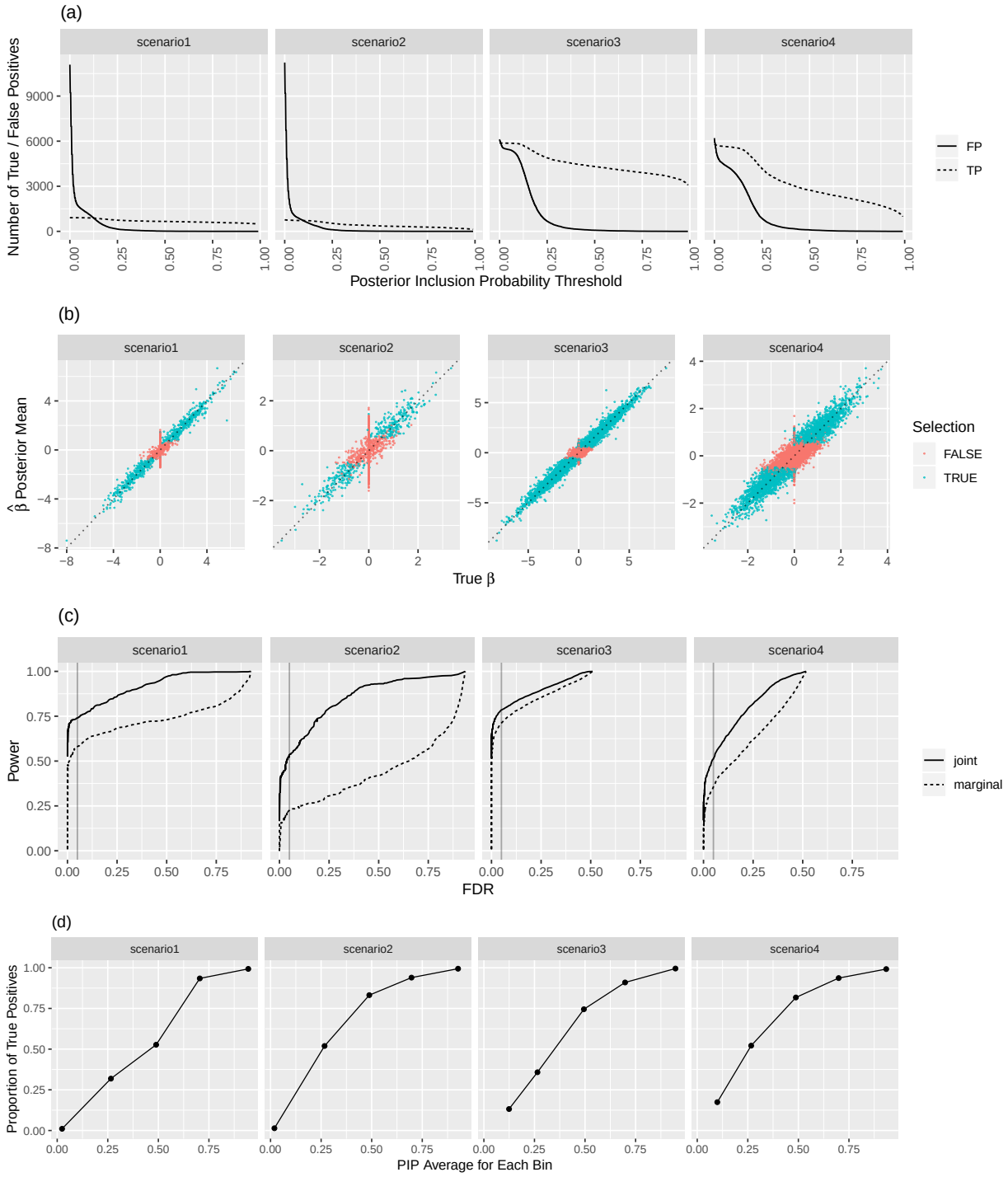

Figure 2. Results from simulation studies. (a) Number of false positives and true positives on varying PIP thresholds. (b) Average of $\hat{\beta}_j$ for iterations j with $\gamma_j = 1$. Red if $\hat{\gamma} = 0$, and blue if $\hat{\gamma} = 1$. (c) Power comparison with univariate analyses at a given FDR. (d) Calibration of PIP. We place each variable into one of 5 bins. Each point on the graph represents a single bin. x coordinate is the mean of the PIPs and y coordinate being the proportion of true positives within the bin.

Table 1
Simulation settings. Scenarios 1 to 4 compare the algorithm's behavior for different effect sizes (σ^2_β) and hyperparameters (a, b). Scenario 5 runs a null simulation with $\boldsymbol{\beta} = 0$, and we observe the algorithm's resistance toward Type I error.

	Scenario 1	Scenario 2	Scenario 3	Scenario 4	Scenario 5 (Null)
n	71				
T	44				
missing proportion	67%				
Σ	$\Sigma_{ii} = 1$, $\Sigma_{ij} = 0.5$				
π	Be(0.05, 0.5)	Be(0.05, 0.5)	Be(1, 1)	Be(1,1)	0
σ^2_β	5	1	5	1	NA
(a, b)	(0.05, 0.5)				
Φ	$\Phi_{ii} = 1$, $\Phi_{ij} = 0$				
ν	n				

Table 2
Average of posterior inclusion probability (PIP) of each scenario given the effect size in simulated data along with the coverage probability of posterior distribution of β.

	Scenario1	Scenario2	Scenario3	Scenario4
	$E(\hat{\gamma})$ (SD)	$E(\hat{\gamma})$ (SD)	$E(\hat{\gamma})$ (SD)	$E(\hat{\gamma})$ (SD)
$\|\beta\| = 0$	0.061 (0.001)	0.058 (0.001)	0.084 (0.001)	0.084 (0.001)
$\|\beta\| \in (0, 0.2]$	0.227 (0.009)	0.226 (0.01)	0.22 (0.009)	0.229 (0.008)
$\|\beta\| \in (0.2, 0.4]$	0.282 (0.011)	0.257 (0.011)	0.27 (0.01)	0.282 (0.009)
$\|\beta\| \in (0.4, 0.6]$	0.348 (0.014)	0.373 (0.014)	0.348 (0.012)	0.357 (0.011)
$\|\beta\| \in (0.6, 0.8]$	0.455 (0.015)	0.505 (0.014)	0.491 (0.014)	0.497 (0.013)
$\|\beta\| \in (0.8, 1]$	0.6 (0.015)	0.643 (0.014)	0.645 (0.013)	0.633 (0.013)
$\|\beta\| \in (1, 1.2]$	0.79 (0.011)	0.809 (0.01)	0.776 (0.011)	0.792 (0.009)
$\|\beta\| \in (1.2, 1.4]$	0.881 (0.008)	0.875 (0.009)	0.873 (0.007)	0.861 (0.008)
$\|\beta\| \in (1.4, 1.6]$	0.921 (0.006)	0.941 (0.005)	0.95 (0.003)	0.934 (0.005)
$\|\beta\| \in (1.6, 1.8]$	0.976 (0.002)	0.962 (0.003)	0.966 (0.002)	0.966 (0.002)
$\|\beta\| \in (1.8, 2]$	0.99 (0.001)	0.989 (0.001)	0.987 (0.001)	0.988(0.001)
$\|\beta\| > 2$	1 (0)	0.999 (0)	0.999 (0)	0.999 (0)

Table 3
PIP greater than 0.95

Tissue	Local	Global
Adipose Subcutaneous	Z98048.1, **FO393419.3**	
Adipose Visceral Omentum	TRAV21, AL354989.1	
Adrenal Gland	CYP3A5, AC139495.1, AC026369.2	**AP000255.1**, AL356966.1
Artery Aorta	MFGE8, RPS15AP36	AL096803.2 , AC011444.1 , AL589765.6
Artery Coronary	**CHP2**, AC090044.1	VN1R81P, BX255923.1 , **IGBP1P1**
Artery Tibial	IGLV1-51	SGK494, AL445435.1 , **IGBP1P1** , AL450263.1, LINC00930
Breast Mammary Tissue	MIR635	AC135507.1
Colon Transverse	**APCDD1L**	
Esophagus Mucosa	ASCL2	AL121655.1 , **AP000255.1**
Esophagus Muscularis		B4GALT6
Heart Atrial Appendace		MYOT
Lung	**CHP2**	
Nerve Tibial	KLB, ADPGK-AS1	
Skin Not Sun Exposed Suprapubic		PLN
Skin Sun Exposed Lower Leg	ZNF788, ADGRG5, **APCDD1L**, CBR3-AS1	CTSV, SEC14L6, AC015914.1
Stomach	**FO393419.3**	
Testis	AC011444.3, AC010327.3	
Thyroid	MYPN, LINC01301	SORCS1, HEMGN
Whole Blood	AC131056.3	

Table 4
Functional Categorization

Pathway	Genes
Immune Response	TRAV21, IGLV1-51
Metabolism	CYP3A5, MFGE8