Web Supplementary Material for "Bayesian Model Selection in Complex Linear Systems, as Illustrated in Genetic Association Studies"

Xiaoquan Wen

Department of Biostatistics, University of Michigan, Ann Arbor, USA

September 3, 2013

A Prior Specification in Genetic Applications

In this section, we summarize and discuss some of the existing results that can be utilized for the prior specification in the SSMR model in various genetic applications.

A.1 Prior Decomposition by Genetic Variants

Guan and Stephens (2011) argue that regression coefficients of genetic effects reflect the "causal" effects on the phenotype of interest and there is no obvious reason to suspect these causal effects among different variants are correlated spatially. (Note, it is important to distinguish the correlations among the observed genotypes and the independence of the underlying genetic

effects.) The similar type of the independent prior has also been widely used in the polygenic models. As a consequence of this reasoning, it is sensible to decompose the $\boldsymbol{W}_g$ matrix into a block diagonal structure, i.e., $\boldsymbol{W}_g = \boldsymbol{\Phi}_1 \oplus \cdots + \boldsymbol{\Phi}_p$, where each block matrix $\boldsymbol{\Phi}_i$ corresponds to a single SNP. Also, the prior distribution $\Pr\left(\xi(\boldsymbol{\beta}_g)\right)$ can be factored into the product of the prior probability of each SNP.

The simple i.i.d priors on SNPs provide a useful starting point for many applications in genetics. More recently, many authors (Veyrieras *et al.* (2008), Stingo *et al.* (2011)) have proposed to integrate SNP-level genomic annotation information into prior specifications. In the simplest case, a logit function is used to connect the genomic feature of a SNP and its marginal prior inclusion probability, and a "feature coefficient" is parametrized to quantify the impact of the genomic feature on the genetic association. The feature coefficient in this context is typically unknown and often of great interest for inference. As a consequence, the priors on different SNPs are no longer i.i.d. This approach not only is useful in integrating additional information to identify the *causal* genetic variant, but also provides an elegant parametric framework to perform feature enrichment analysis, i.e., the posterior inference results of the feature coefficients summarize all necessary statistical evidence of the enrichment of association signals in the relevant annotation categories.

A.2 Priors for Multiple Quantitative Traits Associations

The interplays between genetic variants and multiple phenotypes are complicated: not only genetic variants can directly affect multiple phenotypes, but also there are interactions between phenotypes through gene networks. As a result, genetic variants and phenotypes can be interacted in an indirect way (through some intermediate phenotypes).

Most recently, Stephens (2010) proposes a directed acyclic graph (DAG) approach to address the structured phenotype relationships. Their approach first classifies phenotypes into three groups

of directly affected, indirectly affected and unaffected with respect to a target genetic variant. Conditioning on the classification, an MVLR model is employed to model the genetic association between the genetic variant and the directly affected phenotypes. Because the true classification of the phenotypes is typically unknown, they use Bayesian model averaging technique to account for this latent structure.

Other approaches (Scott-Boyer *et al.* (2012), Stingo *et al.* (2011)) avoid directly modeling the relationship among multiple phenotypes, instead they utilize prior biological pathway and information of gene networks to prioritize the potential associations of a target variant with respect to a group of phenotypes.

A.3 Priors for Heterogeneous Genetic Effects in Subgroups

When considering the genetic effects between a genetic variant and a phenotype in various subgroups (formed either by environmental conditions, e.g. in G×E interactions, or by sampling structures, e.g. in meta-analysis), the key is to account for the heterogeneity of genetic effects. Wen and Stephens (2011) have recently proposed a flexible Bayesian prior to model heterogeneous genetic effects across multiple subgroups. For a genetic variant, this prior assumes that its genetic effects with respect to a common phenotype in s subgroup, if non-zero, are described by

$$\beta_i \sim \mathrm{N}(\bar{\beta}, \phi^2), i = 1, \ldots, s, \tag{1}$$

and

$$\bar{\beta} \sim \mathrm{N}(0, \omega^2), \tag{2}$$

where parameter ω^2 quantifies the prior magnitude of the average effect and ϕ^2 describes the prior degree of heterogeneity. Equivalently, the joint prior distribution for vector $(\beta_1, \ldots, \beta_r)$

can be represented by a multivariate normal distribution with mean 0 and variance-covariance matrix W_g, where

$$\boldsymbol{W}_g = \begin{pmatrix} \phi^2 + \omega^2 & \cdots & \omega^2 \\ \vdots & \ddots & \vdots \\ \omega^2 & \cdots & \phi^2 + \omega^2 \end{pmatrix}. \tag{3}$$

It is easy to see that $\frac{\omega^2}{\omega^2+\phi^2}$ is the prior correlation between a pair of genetic effects: when ϕ^2 is set to 0, it corresponds to the fixed effect model; whereas setting $\omega^2 = 0$ implies the effects are *a priori* independent in all subgroups.

B Bayes Factor Derivation

In this section, we show the derivation of Bayes factors based on the SSMR model.

In the SSMR model, we have defined $\boldsymbol{\mathcal{Y}}, \boldsymbol{\mathcal{X}}, \boldsymbol{\mathcal{E}}, \boldsymbol{\beta}_c$ and $\boldsymbol{\beta}_g$ in section 2 of the main text. In addition, we denote the complete collection of regression coefficients and its vectorized version by $\boldsymbol{\mathcal{B}} := \{\boldsymbol{B}_1, \ldots, \boldsymbol{B}_s\}$ and $\boldsymbol{\beta}_{\text{sys}} := \begin{pmatrix} \boldsymbol{\beta}_c \\ \boldsymbol{\beta}_g \end{pmatrix}$, respectively.

The likelihood function of the SSMR model is given by

$$p(\boldsymbol{\mathcal{Y}}|\boldsymbol{\mathcal{X}}, \boldsymbol{\mathcal{B}}, \boldsymbol{\mathcal{E}}) = (2\pi)^{-\frac{r\sum_{i=1}^{s} n_i}{2}} \cdot \prod_{i=1}^{s} |\boldsymbol{\Sigma}_i|^{-\frac{n_i}{2}} \cdot \text{etr}\left(-\frac{1}{2}\sum_{i=1}^{s} \boldsymbol{\Sigma}_i^{-1}(\boldsymbol{Y}_i - \boldsymbol{X}_i\boldsymbol{B}_i)'(\boldsymbol{Y}_i - \boldsymbol{X}_i\boldsymbol{B}_i)\right) \tag{4}$$

where function etr$(\cdot)$ denotes the exponential of the trace. Given the least squares estimate $\hat{B}_i$ for each composing MVLR, it follows that

$$(\boldsymbol{Y}_i - \boldsymbol{X}_i\boldsymbol{B}_i)'(\boldsymbol{Y}_i - \boldsymbol{X}_i\boldsymbol{B}_i) = (\boldsymbol{Y}_i - \boldsymbol{X}_i\hat{\boldsymbol{B}}_i)'(\boldsymbol{Y}_i - \boldsymbol{X}_i\hat{\boldsymbol{B}}_i) + (\boldsymbol{B}_i - \hat{\boldsymbol{B}}_i)'(\boldsymbol{X}_i'\boldsymbol{X}_i)(\boldsymbol{B}_i - \hat{\boldsymbol{B}}_i). \tag{5}$$

Note this decomposition holds even if $\boldsymbol{X}_i$ is rank-deficient (however, $\hat{\boldsymbol{B}}_i$ may not be unique, see McCullagh and Nelder (1989), page 82 for discussions). We denote $\boldsymbol{\beta}_i := \text{vec}(\boldsymbol{B}_i')$ and $\hat{\boldsymbol{\beta}}_i := \text{vec}(\hat{\boldsymbol{B}}_i')$, and use $\boldsymbol{\beta}_{\text{all}}$ and $\hat{\boldsymbol{\beta}}_{\text{all}}$ to denote the sequentially concatenated vectors of $(\boldsymbol{\beta}_1, \ldots, \boldsymbol{\beta}_s)$ and $(\hat{\boldsymbol{\beta}}_1, \ldots, \hat{\boldsymbol{\beta}}_s)$, respectively. The likelihood function (4) can be re-written as

$$\begin{aligned} p(\boldsymbol{\mathcal{Y}}|\boldsymbol{\mathcal{X}}, \boldsymbol{\mathcal{B}}, \boldsymbol{\mathcal{E}}) =& (2\pi)^{-\frac{r\sum_{i=1}^{s} n_i}{2}} \cdot \prod_{i=1}^{s} |\boldsymbol{\Sigma}_i|^{-\frac{n_i}{2}} \cdot \text{etr}\left(-\frac{1}{2}\sum_{i=1}^{s} \boldsymbol{\Sigma}_i^{-1}(\boldsymbol{Y}_i - \boldsymbol{X}_i\hat{\boldsymbol{B}}_i)'(\boldsymbol{Y}_i - \boldsymbol{X}_i\hat{\boldsymbol{B}}_i)\right) \\ & \cdot \exp\left(-\frac{1}{2}\left(\boldsymbol{\beta}_{\text{all}} - \hat{\boldsymbol{\beta}}_{\text{all}}\right)' \boldsymbol{\Phi} \left(\boldsymbol{\beta}_{\text{all}} - \hat{\boldsymbol{\beta}}_{\text{all}}\right)\right), \end{aligned} \tag{6}$$

where

$$\boldsymbol{\Phi} = \left(\boldsymbol{X}_1'\boldsymbol{X}_1 \otimes \boldsymbol{\Sigma}_s^{-1}\right) \oplus \cdots \oplus \left(\boldsymbol{X}_s'\boldsymbol{X}_s \otimes \boldsymbol{\Sigma}_s^{-1}\right).$$

Also, by the general case of Gauss-Markov theorem, we note that $\text{Var}(\hat{\boldsymbol{\beta}}_{\text{all}}) = \boldsymbol{\Phi}^{-1}$ (In case that $\boldsymbol{\Phi}$ is singular, the Moore–Penrose pseudoinverse is applied).

Although $\boldsymbol{\beta}_{\text{sys}}$ and $\boldsymbol{\beta}_{\text{all}}$ generally differ in the orders of the composing elements, they can be reconciled by a permutation operation, i.e.,

$$\boldsymbol{P}\boldsymbol{\beta}_{\text{all}} = \boldsymbol{\beta}_{\text{sys}}, \tag{7}$$

where $\boldsymbol{P}$ is a $(rps + r\sum_i^s q_i) \times (rps + r\sum_i^s q_i)$ permutation matrix. Furthermore, we denote

$$\boldsymbol{\Omega} = \boldsymbol{P}\boldsymbol{\Phi}\boldsymbol{P},$$

and it can be shown that

$$\text{Var}(\hat{\boldsymbol{\beta}}_{\text{sys}}) = \boldsymbol{\Omega}^{-1}. \tag{8}$$

As a result,

$$
\begin{aligned}
p(\boldsymbol{\mathcal{Y}}|\boldsymbol{\mathcal{X}}, \boldsymbol{\beta}_{\text{sys}}, \boldsymbol{\mathcal{E}}) =& (2\pi)^{-\frac{r\sum_i^s n_i}{2}} \cdot \prod_i^s |\boldsymbol{\Sigma}_i|^{-\frac{n_i}{2}} \cdot \text{etr}\left(-\frac{1}{2}\sum_i^s \boldsymbol{\Sigma}_i^{-1}(\boldsymbol{Y}_i - \boldsymbol{X}_i\hat{\boldsymbol{B}}_i)'(\boldsymbol{Y}_i - \boldsymbol{X}_i\hat{\boldsymbol{B}}_i)\right) \\
& \cdot \exp\left(-\frac{1}{2}\left(\boldsymbol{\beta}_{\text{sys}} - \hat{\boldsymbol{\beta}}_{\text{sys}}\right)' \boldsymbol{\Omega}\left(\boldsymbol{\beta}_{\text{sys}} - \hat{\boldsymbol{\beta}}_{\text{sys}}\right)\right).
\end{aligned} \tag{9}
$$

B.1 Bayes Factor for Known $\boldsymbol{\Sigma}$

With $\boldsymbol{\mathcal{E}}$ known, the marginal likelihood $p(\boldsymbol{\mathcal{Y}}|\boldsymbol{\mathcal{X}}, \boldsymbol{\mathcal{E}})$ can be evaluated analytically, i.e.,

$$
p(\boldsymbol{\mathcal{Y}}|\boldsymbol{\mathcal{X}}, \boldsymbol{\mathcal{E}}) = \int p(\boldsymbol{\mathcal{Y}}|\boldsymbol{\mathcal{X}}, \boldsymbol{\mathcal{E}}, \boldsymbol{\beta}_{\text{sys}}) p(\boldsymbol{\beta}_{\text{sys}})\, d\boldsymbol{\beta}_{\text{sys}}. \tag{10}
$$

Recall the prior distribution defined in section 2 of the main text,

$$
\boldsymbol{\beta}_{\text{sys}} \sim \text{N}(\boldsymbol{0}, \boldsymbol{\Psi}_c \oplus \boldsymbol{W}_g).
$$

Assuming $\boldsymbol{W}_g$ is full rank, the integration yields

$$
\begin{aligned}
p(\boldsymbol{\mathcal{Y}}|\boldsymbol{\mathcal{X}}, \boldsymbol{\mathcal{E}}) =& (2\pi)^{-\frac{r\sum_i^s n_i}{2}} \cdot \prod_i^s |\boldsymbol{\Sigma}_i|^{-\frac{n_i}{2}} \cdot |\boldsymbol{W}_g|^{-\frac{1}{2}} \cdot |\boldsymbol{\Psi}_c|^{-\frac{1}{2}} \cdot |\boldsymbol{\Omega} + \boldsymbol{\Psi}_c^{-1} \oplus \boldsymbol{W}_g^{-1}|^{-\frac{1}{2}} \\
& \cdot \exp\left(-\frac{1}{2}\hat{\boldsymbol{\beta}}'_{\text{sys}}\boldsymbol{\Omega}\left(\boldsymbol{\Omega}^{-1} - (\boldsymbol{\Omega} + \boldsymbol{\Psi}_c^{-1} \oplus \boldsymbol{W}_g^{-1})^{-1}\right)\boldsymbol{\Omega}\hat{\boldsymbol{\beta}}_{\text{sys}}\right) \\
& \cdot \text{etr}\left(-\frac{1}{2}\sum_i^s \boldsymbol{\Sigma}_i^{-1}(\boldsymbol{Y}_i - \boldsymbol{X}_i\hat{\boldsymbol{B}}_i)'(\boldsymbol{Y}_i - \boldsymbol{X}_i\hat{\boldsymbol{B}}_i)\right),
\end{aligned} \tag{11}
$$

To further simplify (11), we decompose Ω into the following block matrix

$$
\boldsymbol{\Omega} = \begin{pmatrix} \boldsymbol{\Omega}_c & \boldsymbol{\Omega}_f \\ \boldsymbol{\Omega}'_f & \boldsymbol{\Omega}_g \end{pmatrix},
$$

where $\boldsymbol{\Omega}_c$ and $\boldsymbol{\Omega}_g$ match the the dimensions of the matrices $\boldsymbol{\Psi}_c$ and $\boldsymbol{W}_g$, respectively. By (8), it follows that

$$\boldsymbol{V}_g^{-1} = \boldsymbol{\Omega}_g - \boldsymbol{\Omega}_f' \boldsymbol{\Omega}_c^{-1} \boldsymbol{\Omega}_f. \tag{12}$$

Let

$$\boldsymbol{\mathcal{U}} = \boldsymbol{\Omega}_g - \boldsymbol{\Omega}_f' (\boldsymbol{\Omega}_c + \boldsymbol{\Psi}_c^{-1})^{-1} \boldsymbol{\Omega}_f + \boldsymbol{W}_g^{-1},$$

and it follows that

$$|\boldsymbol{\Omega} + \boldsymbol{\Psi}_c^{-1} \oplus \boldsymbol{W}_g^{-1}| = |\boldsymbol{\Omega}_c + \boldsymbol{\Psi}_c^{-1}| \cdot |\boldsymbol{\mathcal{U}}|. \tag{13}$$

Furthermore, the matrix product, $\boldsymbol{\Omega} \left(\boldsymbol{\Omega}^{-1} - (\boldsymbol{\Omega} + \boldsymbol{\Psi}_c^{-1} \oplus \boldsymbol{W}_g^{-1})^{-1} \right) \boldsymbol{\Omega}$, can be represented by the block matrix $\begin{pmatrix} \boldsymbol{A} & \boldsymbol{B} \\ \boldsymbol{B}' & \boldsymbol{D} \end{pmatrix}$, where

$$\begin{aligned}
\boldsymbol{A} &= \boldsymbol{\Omega}_c \left[\boldsymbol{I} - (\boldsymbol{\Omega}_c + \boldsymbol{\Psi}_c^{-1})^{-1} \boldsymbol{\Omega}_c \right] - \left[\boldsymbol{I} - \boldsymbol{\Omega}_c (\boldsymbol{\Omega}_c + \boldsymbol{\Psi}_c^{-1})^{-1} \right] \boldsymbol{\Omega}_f \boldsymbol{\mathcal{U}}^{-1} \boldsymbol{\Omega}_f' \left[\boldsymbol{I} - (\boldsymbol{\Omega}_c + \boldsymbol{\Psi}_c^{-1})^{-1} \boldsymbol{\Omega}_c \right], \\
B &= \left[\boldsymbol{I} - \boldsymbol{\Omega}_c (\boldsymbol{\Omega}_c + \boldsymbol{\Psi}_c^{-1})^{-1} \right] \boldsymbol{\Omega}_f \boldsymbol{\mathcal{U}}^{-1} \boldsymbol{W}_g^{-1} \\
D &= \boldsymbol{W}_g^{-1} - \boldsymbol{W}_g^{-1} \boldsymbol{\mathcal{U}}^{-1} \boldsymbol{W}_g^{-1} = (\boldsymbol{\mathcal{U}} - \boldsymbol{W}_g^{-1}) - (\boldsymbol{\mathcal{U}} - \boldsymbol{W}_g^{-1}) \boldsymbol{\mathcal{U}}^{-1} (\boldsymbol{\mathcal{U}} - \boldsymbol{W}_g^{-1}).
\end{aligned}$$

Although the expressions are fairly complicated, when the limit $\boldsymbol{\Psi}_c^{-1} \to 0$ is taken, $\boldsymbol{A} \to 0$ and $\boldsymbol{B} \to 0$.

The exact same calculations can be carried out with respect to the null model. In the end, we

obtain the following marginal likelihood under H_0,

$$
\begin{aligned}
P(\boldsymbol{\mathcal{Y}}|\boldsymbol{\mathcal{X}},\boldsymbol{\mathcal{E}},H_0) = (2\pi)^{-\frac{r\sum_i^s n_i}{2}} \cdot \prod_i^s |\boldsymbol{\Sigma}_i|^{-\frac{n_i}{2}} \cdot |\boldsymbol{\Psi}_c|^{-\frac{1}{2}} \cdot |\boldsymbol{\Omega}_c + \boldsymbol{\Psi}_c^{-1}|^{-\frac{1}{2}} \\
\cdot \exp\left(-\frac{1}{2}\tilde{\boldsymbol{\beta}}_c' \boldsymbol{\Omega}_c \left(\boldsymbol{\Omega}_c^{-1} - (\boldsymbol{\Omega}_c + \boldsymbol{\Psi}_c^{-1})^{-1}\right) \boldsymbol{\Omega}_c \tilde{\boldsymbol{\beta}}_c\right) \\
\cdot \operatorname{etr}\left(-\frac{1}{2}\sum_i^s \boldsymbol{\Sigma}_i^{-1}(\boldsymbol{Y}_i - \boldsymbol{X}_{c,i}\tilde{\boldsymbol{B}}_i)'(\boldsymbol{Y}_i - \boldsymbol{X}_{c,i}\tilde{\boldsymbol{B}}_i)\right),
\end{aligned}
\tag{14}
$$

where $\tilde{\boldsymbol{\beta}}_c$ and $\tilde{\boldsymbol{B}}_i$ are the MLEs of regression coefficients obtained under the null model (i.e. restricting $\boldsymbol{\beta}_g \equiv 0$). Note the relationship of the least squares estimates between the target and the null models:

$$
\tilde{\boldsymbol{B}}_i = \hat{\boldsymbol{B}}_{c,i} + (\boldsymbol{X}_{c,i}'\boldsymbol{X}_{c,i})^{-1}\boldsymbol{X}_{c,i}'\boldsymbol{X}_{g,i}\hat{\boldsymbol{B}}_{g,i}, \tag{15}
$$

and

$$
\begin{aligned}
&(\boldsymbol{Y}_i - \boldsymbol{X}_{c,i}\tilde{\boldsymbol{B}}_i)'(\boldsymbol{Y}_i - \boldsymbol{X}_{c,i}\tilde{\boldsymbol{B}}_i) - (\boldsymbol{Y}_i - \boldsymbol{X}_i\hat{\boldsymbol{B}}_i)'(\boldsymbol{Y}_i - \boldsymbol{X}_i\hat{\boldsymbol{B}}_i) \\
&= \hat{\boldsymbol{B}}_{g,i}' \left(\boldsymbol{X}_{g,i}'\boldsymbol{X}_{g,i} - \boldsymbol{X}_{g,i}'\boldsymbol{X}_{c,i}(\boldsymbol{X}_{c,i}'\boldsymbol{X}_{c,i})^{-1}\boldsymbol{X}_{c,i}'\boldsymbol{X}_{g,i}\right)\hat{\boldsymbol{B}}_{g,i}
\end{aligned}
\tag{16}
$$

It follows that

$$
\begin{aligned}
&\operatorname{etr}\left(\frac{1}{2}\sum_i^s \boldsymbol{\Sigma}_i^{-1}\left[(\boldsymbol{Y}_i - \boldsymbol{X}_{c,i}\tilde{\boldsymbol{B}}_i)'(\boldsymbol{Y}_i - \boldsymbol{X}_{c,i}\tilde{\boldsymbol{B}}_i) - (\boldsymbol{Y}_i - \boldsymbol{X}_i\hat{\boldsymbol{B}}_i)'(\boldsymbol{Y}_i - \boldsymbol{X}_i\hat{\boldsymbol{B}}_i)\right]\right) \\
&= \exp\left(\frac{1}{2}\hat{\boldsymbol{\beta}}_g'\boldsymbol{V}_g^{-1}\hat{\boldsymbol{\beta}}_g\right).
\end{aligned}
\tag{17}
$$

This also gives the explicit expression for $\boldsymbol{V}_g^{-1}$, i.e.,

$$
\boldsymbol{V}_g^{-1} = \oplus_{i=1}^s \boldsymbol{V}_{g,i}^{-1} = \oplus_{i=1}^s \left[\left(\boldsymbol{X}_{g,i}'\boldsymbol{X}_{g,i} - \boldsymbol{X}_{g,i}'\boldsymbol{X}_{c,i}(\boldsymbol{X}_{c,i}'\boldsymbol{X}_{c,i})^{-1}\boldsymbol{X}_{c,i}'\boldsymbol{X}_{g,i}\right) \otimes \boldsymbol{\Sigma}_i^{-1}\right]. \tag{18}
$$

Because of the block-diagonal nature of the $\boldsymbol{V}_g^{-1}$ matrix, the following expression also holds true

$$\hat{\boldsymbol{\beta}}_g' \boldsymbol{V}_g^{-1} \hat{\boldsymbol{\beta}}_g = \sum_{i=1}^{s} \hat{\boldsymbol{\beta}}_{g,i}' \boldsymbol{V}_{g,i}^{-1} \hat{\boldsymbol{\beta}}_{g,i}, \tag{19}$$

which provides convenience for Laplace approximation later on.

Finally, by taking the limit $\boldsymbol{\Psi}_c^{-1} \to 0$ and noting

$$\lim_{\psi_c^{-1} \to 0} \boldsymbol{\mathcal{U}} = \boldsymbol{V}_g^{-1} + \boldsymbol{W}_g^{-1}, \tag{20}$$

we obtain

$$\mathrm{BF}(\boldsymbol{W}_g) = |\boldsymbol{I} + \boldsymbol{V}_g^{-1}\boldsymbol{W}_g|^{-\frac{1}{2}} \cdot \exp\left(\frac{1}{2} \hat{\boldsymbol{\beta}}_g' \boldsymbol{V}_g^{-1} \left[\boldsymbol{W}_g (\boldsymbol{I} + \boldsymbol{V}_g^{-1}\boldsymbol{W}_g)^{-1} \right] \boldsymbol{V}_g^{-1} \hat{\boldsymbol{\beta}}_g \right), \tag{21}$$

which proves LEMMA 1.

B.2 Approximate Bayes Factors for Unknown $\boldsymbol{\Sigma}$

When $\boldsymbol{\mathcal{E}}$ is unknown, we assign independent inverse Wishart priors, $\mathrm{IW}_r(\nu_i \boldsymbol{H}_i, m_i)$, to each $\boldsymbol{\Sigma}_i$ and additional integrals are required for computing the marginal likelihood. More specifically, the goal is to evaluate

$$p(\boldsymbol{\mathcal{Y}}|\boldsymbol{\mathcal{X}}) = \int p(\boldsymbol{\mathcal{Y}}|\boldsymbol{\mathcal{X}}, \boldsymbol{\mathcal{E}}) \prod_i p(\boldsymbol{\Sigma}_i^{-1}) \, d\boldsymbol{\Sigma}_1^{-1} \ldots d\boldsymbol{\Sigma}_s^{-1}, \tag{22}$$

where

$$p(\boldsymbol{\Sigma}_i^{-1}) \propto |\boldsymbol{\Sigma}_i^{-1}|^{\frac{m_i - r - 1}{2}} \mathrm{etr}\left(-\frac{1}{2} \nu_i \boldsymbol{H}_i \boldsymbol{\Sigma}_i^{-1} \right). \tag{23}$$

The desired Bayes factor is therefore computed as

$$\mathrm{BF}(\boldsymbol{W}_g) = \lim_{\boldsymbol{\Psi}_c^{-1}\to 0} \frac{\int p(\boldsymbol{\mathcal{Y}}|\boldsymbol{\mathcal{X}},\boldsymbol{\mathcal{E}}) \prod_i p(\boldsymbol{\Sigma}_i^{-1})\; d\,\boldsymbol{\Sigma}_1^{-1}\cdots d\,\boldsymbol{\Sigma}_s^{-1}}{\int p(\boldsymbol{\mathcal{Y}}|\boldsymbol{\mathcal{X}},\boldsymbol{\mathcal{E}},H_0) \prod_i p(\boldsymbol{\Sigma}_i^{-1})\; d\,\boldsymbol{\Sigma}_1^{-1}\cdots d\,\boldsymbol{\Sigma}_s^{-1}}. \tag{24}$$

By plugging in (11) and (14) and noting the cancellation of $|\boldsymbol{\Psi}_c|^{-\frac{1}{2}}$ terms along with the fact that $\boldsymbol{\Omega}^{-1} - (\boldsymbol{\Omega} + \boldsymbol{\Psi}_c^{-1} \oplus \boldsymbol{W}_g^{-1})^{-1}$ is positive definite, it is easy to see that the remaining integrands, both are functions of $\boldsymbol{\Psi}_c^{-1}$, are bounded. It is then justified by bounded convergence theorem (BCT) to switch the limit and integration operations. As a result, we obtain

$$\mathrm{BF}(\boldsymbol{W}_g) = \frac{\int K_{H_a}\, d\,\boldsymbol{\Sigma}_1^{-1}\cdots d\,\boldsymbol{\Sigma}_s^{-1}}{\int K_{H_0}\, d\,\boldsymbol{\Sigma}_1^{-1}\cdots d\,\boldsymbol{\Sigma}_s^{-1}}, \tag{25}$$

where

$$\begin{aligned} K_{H_a} = {} & |\boldsymbol{I} + \boldsymbol{V}_g^{-1}\boldsymbol{W}_g|^{-\frac{1}{2}} \cdot \exp\left(\frac{1}{2}\hat{\boldsymbol{\beta}}_g' \left[\boldsymbol{V}_g^{-1}\boldsymbol{W}_g(\boldsymbol{I} + \boldsymbol{V}_g^{-1}\boldsymbol{W}_g)^{-1}\boldsymbol{V}_g^{-1}\right]\hat{\boldsymbol{\beta}}_g\right) \\ & \cdot \prod_{i=1}^{s} |\boldsymbol{\Sigma}_i^{-1}|^{\frac{n_i+m_i-q_i-r-1}{2}} \cdot \mathrm{etr}\left(-\frac{1}{2}\sum_{i=1}^{s}\boldsymbol{\Sigma}_i^{-1}\left(\nu_i\boldsymbol{H}_i + (\boldsymbol{Y}_i - \boldsymbol{X}_{c,i}\tilde{\boldsymbol{B}}_i)'(\boldsymbol{Y}_i - \boldsymbol{X}_{c,i}\tilde{\boldsymbol{B}}_i)\right)\right), \end{aligned} \tag{26}$$

$$K_{H_0} = \prod_{i=1}^{s} |\boldsymbol{\Sigma}_i^{-1}|^{\frac{n_i+m_i-q_i-r-1}{2}} \cdot \mathrm{etr}\left(-\frac{1}{2}\sum_{i=1}^{s}\boldsymbol{\Sigma}_i^{-1}\left(\nu_i\boldsymbol{H}_i + (\boldsymbol{Y}_i - \boldsymbol{X}_{c,i}\tilde{\boldsymbol{B}}_i)'(\boldsymbol{Y}_i - \boldsymbol{X}_{c,i}\tilde{\boldsymbol{B}}_i)\right)\right), \tag{27}$$

Because $\boldsymbol{V}_g^{-1}$ and (potentially) $\boldsymbol{W}_g$ are both functions of $\boldsymbol{\mathcal{E}}$, the analytic integration of K_{H_a} is generally implausible. Here we approximate the integrals of both K_{H_a} and K_{H_0} by Laplace's method. Note, although the analytic integration of K_{H_0} is straightforward, it is been shown (Wen and Stephens (2011)) that simultaneously applying Laplace's methods to both K_{H_a} and K_{H_0} achieves better numerical accuracy for desired Bayes factor.

Laplace's method approximates an integral with respect to a $d \times d$ symmetric matrix $\boldsymbol{Z}$ (or equivalently the corresponding half-vectorized $(d+1)d/2$ dimensional vector $\mathrm{vech}(\boldsymbol{Z})$) in the

following way,

$$\int_D h(\boldsymbol{Z}) \exp\left(g(\boldsymbol{Z}) \right) \, d\,\boldsymbol{Z} \approx (2\pi)^{d(d+1)/4} |\boldsymbol{H}_{\hat{\boldsymbol{Z}}}|^{-1/2} h(\hat{\boldsymbol{Z}}) \exp\left(g(\hat{\boldsymbol{Z}}) \right), \tag{28}$$

where

$$\hat{\boldsymbol{Z}} = \arg\max_{\boldsymbol{Z}} g(\boldsymbol{Z}),$$

and $|\boldsymbol{H}_{\hat{\boldsymbol{Z}}}|$ is the absolute value of the determinant of the Hessian matrix of the function g evaluated at $\hat{\boldsymbol{Z}}$. The technical requirements on the factorization are that $h(\cdot)$ is smooth and positively valued and $g(\cdot)$ is smooth and obtains its unique maximum in the interior of D. Although different factorization schemes generally achieve different approximation accuracies for finite sample sizes, the asymptotic error bounds are typically the same. For a detailed discussion, see Butler (2007) chapter 2.

To evaluate the desired Bayes factor, we sequentially apply the Laplace's method with respect to each $\boldsymbol{\Sigma}_i^{-1}$ for both K_{H_a} and K_{H_0}.

B.2.1 General Derivation

By (17) and (19), we note the exponential term

$$\operatorname{tr}\left[\boldsymbol{\Sigma}_j^{-1} \left(\nu_j \boldsymbol{H}_j + (\boldsymbol{Y}_j - \boldsymbol{X}_{c,j} \tilde{\boldsymbol{B}}_j)' (\boldsymbol{Y}_j - \boldsymbol{X}_{c,j} \tilde{\boldsymbol{B}}_j) \right) \right], \tag{29}$$

is presented in both the alternative and the null models for each multivariate linear regression model j, and it can be generally decomposed into

$$\begin{aligned}&\operatorname{tr}\left[\boldsymbol{\Sigma}_j^{-1}\left(\nu_j\boldsymbol{H}_j+(1-\alpha_j)(\boldsymbol{Y}_j-\boldsymbol{X}_{c,j}\tilde{\boldsymbol{B}}_j)'(\boldsymbol{Y}_j-\boldsymbol{X}_{c,j}\tilde{\boldsymbol{B}}_j)+\alpha_j\left(\boldsymbol{Y}_j-\boldsymbol{X}_{c,j}\hat{\boldsymbol{B}}_j\right)'(\boldsymbol{Y}_j-\boldsymbol{X}_{c,j}\hat{\boldsymbol{B}}_j)\right)\right]\\&+\alpha_j\,\hat{\boldsymbol{\beta}}_{g,i}'\boldsymbol{V}_{g,i}^{-1}\hat{\boldsymbol{\beta}}_{g,i},\end{aligned}\tag{30}$$

where $\alpha_j\in[0,1]$. Thus, when applying Laplace's method, we start by factoring K_{H_a} into

$$K_{H_a}=h_a(\boldsymbol{\Sigma}_1^{-1},\ldots,\boldsymbol{\Sigma}_s^{-1})\exp\left(\,g_a(\boldsymbol{\Sigma}_1^{-1},\ldots,\boldsymbol{\Sigma}_s^{-1})\,\right),\tag{31}$$

where

$$\begin{aligned}h_a(\boldsymbol{\Sigma}_1^{-1},\ldots,\boldsymbol{\Sigma}_s^{-1})=|\boldsymbol{I}+\boldsymbol{V}_g^{-1}\boldsymbol{W}_g|^{-\frac{1}{2}}\cdot\exp\left(\frac{1}{2}\hat{\boldsymbol{\beta}}_g'\left[\boldsymbol{V}_g^{-1}\boldsymbol{W}_g(\boldsymbol{I}+\boldsymbol{V}_g^{-1}\boldsymbol{W}_g)^{-1}\boldsymbol{V}_g^{-1}\right]\hat{\boldsymbol{\beta}}_g\right)\\\cdot\exp\left(-\frac{1}{2}\sum_{i=1}^{s}\alpha_i\hat{\boldsymbol{\beta}}_{g,i}'\boldsymbol{V}_{g,i}^{-1}\hat{\boldsymbol{\beta}}_{g,i}\right)\end{aligned}\tag{32}$$

and

$$\begin{aligned}&g_a(\boldsymbol{\Sigma}_1^{-1},\ldots,\boldsymbol{\Sigma}_s^{-1})=\sum_{i=1}^{s}\frac{n_i+\nu_i}{2}\log|\boldsymbol{\Sigma}_i^{-1}|\\&-\frac{1}{2}\sum_{i=1}^{s}\operatorname{tr}\left[\boldsymbol{\Sigma}_i^{-1}\left(\nu_i\boldsymbol{H}_i+\alpha_i(\boldsymbol{Y}_i-\boldsymbol{X}_i\hat{\boldsymbol{B}}_i)'(\boldsymbol{Y}_i-\boldsymbol{X}_i\hat{\boldsymbol{B}}_i)+(1-\alpha_i)(\boldsymbol{Y}_i-\boldsymbol{X}_i\tilde{\boldsymbol{B}}_i)'(\boldsymbol{Y}_i-\boldsymbol{X}_i\tilde{\boldsymbol{B}}_i)\right)\right].\end{aligned}\tag{33}$$

It is straightforward to show that the unique maximum of $g(\boldsymbol{\Sigma}^{-1},\ldots,\boldsymbol{\Sigma}_s^{-1})$ can be obtained by performing sequential analytic maximization with respect to each individual $\boldsymbol{\Sigma}_i$. More specifi-

cally, the maximum is attained at

$$\check{\boldsymbol{\Sigma}}_i = \frac{\nu_i}{n_i + \nu_i}\boldsymbol{H}_i + \frac{n_i}{n_i + \nu_i}\left[\alpha_i\hat{\boldsymbol{\Sigma}}_i + (1-\alpha_i)\tilde{\boldsymbol{\Sigma}}_i\right],\ \forall i, \tag{34}$$

where

$$\hat{\boldsymbol{\Sigma}}_i = \frac{1}{n_i}(\boldsymbol{Y}_i - \boldsymbol{X}_i\hat{\boldsymbol{B}}_i)'(\boldsymbol{Y}_i - \boldsymbol{X}_i\hat{\boldsymbol{B}}_i), \tag{35}$$

and

$$\tilde{\boldsymbol{\Sigma}}_i = \frac{1}{n_i}(\boldsymbol{Y}_i - \boldsymbol{X}_{c,i}\tilde{\boldsymbol{B}}_i)'(\boldsymbol{Y}_i - \boldsymbol{X}_{c,i}\tilde{\boldsymbol{B}}_i), \tag{36}$$

are commonly used MLEs of $\boldsymbol{\Sigma}_i$ evaluated under the target and the null models respectively.

Following Minka (2000), it can be shown that the Hessian matrix $H_{g_a}(\boldsymbol{\Sigma}_i^{-1})$ for each $\boldsymbol{\Sigma}_i^{-1}$ is given by

$$\begin{aligned} H_{g_a}(\boldsymbol{\Sigma}_i^{-1}) =& \frac{\mathrm{d}^2 g_a}{\mathrm{dvech}(\boldsymbol{\Sigma}_i^{-1})\mathrm{dvech}(\boldsymbol{\Sigma}_i^{-1})'} \\ =& -\frac{n_i}{2}\boldsymbol{D}_s'\left(\boldsymbol{\Sigma}_i \otimes \boldsymbol{\Sigma}_i\right)\boldsymbol{D}_s, \end{aligned} \tag{37}$$

where $\boldsymbol{D}_s$ denotes the duplication matrix for $s \times s$ symmetric matrices. As it is evaluated at $\check{\boldsymbol{\Sigma}}_i^{-1}$, its absolute determinant results in the following simple form,

$$|H_{g_a}(\check{\boldsymbol{\Sigma}}_i^{-1})| = 2^{-r} n_i^{r(r+1)/2}|\check{\boldsymbol{\Sigma}}_i|^{r+1}. \tag{38}$$

Similarly, we factor K_{H_0} in the same way, i.e.,

$$K_{H_0} = h_0(\boldsymbol{\Sigma}_1^{-1}, \ldots, \boldsymbol{\Sigma}_s^{-1})\exp(\, g_0(\boldsymbol{\Sigma}_1^{-1}), \ldots, \boldsymbol{\Sigma}_s^{-1}\,), \tag{39}$$

where

$$h_0(\boldsymbol{\Sigma}_1^{-1},\ldots,\boldsymbol{\Sigma}_s^{-1}) = \exp\left(-\frac{1}{2}\sum_{i=1}^{s}\alpha_i\hat{\boldsymbol{\beta}}_{g,i}'\boldsymbol{V}_{g,i}^{-1}\hat{\boldsymbol{\beta}}_{g,i}\right) \tag{40}$$

and

$$\begin{aligned} g_0(\boldsymbol{\Sigma}_1^{-1},\ldots,\boldsymbol{\Sigma}_s^{-1}) &= \sum_{i=1}^{s}\frac{n_i+\nu_i}{2}\log|\boldsymbol{\Sigma}_i^{-1}| \\ -\frac{1}{2}\sum_{i=1}^{s}\mathrm{tr}\left[\boldsymbol{\Sigma}_i^{-1}\left(\nu_i\boldsymbol{H}_i + \alpha_i(\boldsymbol{Y}_i-\boldsymbol{X}_i\hat{\boldsymbol{B}}_i)'(\boldsymbol{Y}_i-\boldsymbol{X}_i\hat{\boldsymbol{B}}_i) + (1-\alpha_i)(\boldsymbol{Y}_i-\boldsymbol{X}_i\tilde{\boldsymbol{B}}_i)'(\boldsymbol{Y}_i-\boldsymbol{X}_i\tilde{\boldsymbol{B}}_i)\right)\right]. \end{aligned} \tag{41}$$

Note that $g_0(\boldsymbol{\Sigma}_1^{-1},\ldots,\boldsymbol{\Sigma}_s^{-1})$ and $g_a(\boldsymbol{\Sigma}_1^{-1}\ldots,\boldsymbol{\Sigma}_s^{-1})$ are identical, $(\check{\boldsymbol{\Sigma}}_1,\ldots,\check{\boldsymbol{\Sigma}}_s)$ also uniquely maximizes g_0 function.

Following (28), the desired Bayes factor is computed as

$$\mathrm{BF}(\boldsymbol{W}_g) = |\boldsymbol{I}+\check{\boldsymbol{V}}_g^{-1}\check{\boldsymbol{W}}_g|^{-\frac{1}{2}}\cdot\exp\left(\frac{1}{2}\hat{\boldsymbol{\beta}}_g'\check{\boldsymbol{V}}_g^{-1}\left[\check{\boldsymbol{W}}_g(\boldsymbol{I}+\check{\boldsymbol{V}}_g^{-1}\check{\boldsymbol{W}}_g)^{-1}\right]\check{\boldsymbol{V}}_g^{-1}\hat{\boldsymbol{\beta}}_g\right)\cdot\prod_{i=1}^{s}\left(1+O(\frac{1}{n_i})\right) \tag{42}$$

where $\check{\boldsymbol{V}}_g^{-1}$ and $\check{\boldsymbol{W}}_g$ are the corresponding $\boldsymbol{V}_g^{-1}$ and $\boldsymbol{W}_g$ evaluated at $(\check{\boldsymbol{\Sigma}}_1,\ldots,\check{\boldsymbol{\Sigma}}_s)$. In particular,

$$\check{\boldsymbol{V}}_g^{-1} = \oplus_{i=1}^{s}\left[\left(\boldsymbol{X}_{g,i}'\boldsymbol{X}_{g,i} - \boldsymbol{X}_{g,i}'\boldsymbol{X}_{c,i}(\boldsymbol{X}_{c,i}'\boldsymbol{X}_{c,i})^{-1}\boldsymbol{X}_{c,i}'\boldsymbol{X}_{g,i}\right)\otimes\check{\boldsymbol{\Sigma}}_i^{-1}\right]. \tag{43}$$

This leads to the final expression of ABF

$$\mathrm{ABF}(\boldsymbol{W}_g,\boldsymbol{\alpha}) = |\boldsymbol{I}+\check{\boldsymbol{V}}_g^{-1}\check{\boldsymbol{W}}_g|^{-\frac{1}{2}}\cdot\exp\left(\frac{1}{2}\hat{\boldsymbol{\beta}}_g'\check{\boldsymbol{V}}_g^{-1}\left[\check{\boldsymbol{W}}_g(\boldsymbol{I}+\check{\boldsymbol{V}}_g^{-1}\check{\boldsymbol{W}}_g)^{-1}\right]\check{\boldsymbol{V}}_g^{-1}\hat{\boldsymbol{\beta}}_g\right), \tag{44}$$

which also completes the proof for PROPOSITION 1.

C Computational Stability of Bayes Factor

In this section, we demonstrate the computational stability of the derived Bayes factors. In particular, we show that the derived Bayes factor and its approximations can be stably evaluated even if some design matrix $\boldsymbol{X}_i \in \boldsymbol{\mathcal{X}}$ is rank deficient.

First, assuming $\boldsymbol{X}'_{c,i}\boldsymbol{X}_{c,i}$ can be inverted in the general sense $\forall i = 1, \ldots, s$, we define

$$\boldsymbol{G}_i = \left(\boldsymbol{I} - \boldsymbol{X}_{c,i}\left(\boldsymbol{X}'_{c,i}\boldsymbol{X}_{c,i}\right)^{-1}\boldsymbol{X}'_{c,i}\right)\boldsymbol{X}_{g,i}, \tag{45}$$

and denote its $p \times n_i$ Moore-Penrose pseudo inverse matrix by $\boldsymbol{G}_i^+$. By the general least squares theory, it can be shown (regardless if $\boldsymbol{G}_i$ is full-rank) that

$$\hat{\boldsymbol{B}}_{g,i} = \boldsymbol{G}_i^+\boldsymbol{Y}_i, \tag{46}$$

$$\hat{\boldsymbol{\beta}}_{g,i} = \mathrm{vec}(\hat{\boldsymbol{B}}'_{g,i}) = (\boldsymbol{G}_i^+ \otimes \boldsymbol{I})\mathrm{vec}(\boldsymbol{Y}'_i) \tag{47}$$

$$\tag{48}$$

and

$$\boldsymbol{V}_{g,i}^{-1} = (\boldsymbol{G}'_i\boldsymbol{G}_i) \otimes \boldsymbol{\Sigma}_i^{-1}. \tag{49}$$

It is then follows from the general property of Moore-Penrose pseudo inverse, such that

$$\begin{aligned}\boldsymbol{V}_{g,i}^{-1}\hat{\boldsymbol{\beta}}_{g,i} &= \left[\left(\boldsymbol{G}'_i\boldsymbol{G}_i\boldsymbol{G}_i^+\right) \otimes \boldsymbol{\Sigma}_i^{-1}\right]\mathrm{vec}(\boldsymbol{Y}'_i) \\ &= (\boldsymbol{G}'_i \otimes \boldsymbol{\Sigma}_i^{-1})\mathrm{vec}(\boldsymbol{Y}'_i) \\ &= \mathrm{vec}(\boldsymbol{\Sigma}^{-1}\boldsymbol{Y}'_i\boldsymbol{G}_i).\end{aligned} \tag{50}$$

Finally, $\boldsymbol{V}_g^{-1}\hat{\boldsymbol{\beta}}_g$ is computed by sequentially concatenating $\boldsymbol{V}_{g,i}^{-1}\hat{\boldsymbol{\beta}}_{g,i}$ for $i = 1, ..., s$. Note, in this

computational procedure

1. there is no matrix inversion operation on $\boldsymbol{X}'_{g,i}\boldsymbol{X}_{g,i}$ (which we allow to be rank deficient).
2. there is no matrix inversion operation on $\boldsymbol{W}_g$.
3. matrix $(\boldsymbol{I} + \boldsymbol{V}_g^{-1}\boldsymbol{W}_g)$ is guaranteed positive definite.

In case that $\boldsymbol{\mathcal{E}}$ is unknown and some $\boldsymbol{X}_{g,i}$ is rank deficient, it becomes inevitable to perform Moore-Penrose pseudo-inverse of $\boldsymbol{X}'_i\boldsymbol{X}_i$ for evaluation of $\hat{\boldsymbol{\Sigma}}_i$. This would cost the computational efficiency but unlikely affect the computational stability of the ABF.

D Computing Bayes Factors with Singular $\boldsymbol{W}_g$

We first give the proof for PROPOSITION 2 in below.

Proof. In case that $\boldsymbol{\mathcal{E}}$ is known, the proof is trivial by noting that there is no matrix inversion of $\boldsymbol{W}_g$ in the Bayes factor formula of LEMMA 1.

If $\boldsymbol{\mathcal{E}}$ is unknown, the desired Bayes factor is computed by

$$\mathrm{BF}(\boldsymbol{W}_g) = \frac{\lim_{\lambda\to 0}\int K_{H_a}(\boldsymbol{W}_g^{\dagger}(\lambda))\, d\boldsymbol{\Sigma}_1^{-1} \ldots d\boldsymbol{\Sigma}_s^{-1}}{\int K_{H_0}\, d\boldsymbol{\Sigma}^{-1} \ldots d\boldsymbol{\Sigma}_s^{-1}}, \tag{51}$$

where the integrands K_{H_a} and K_{H_0} are defined in (26) and (27) respectively. It should be clear that

$$K_{H_a}(\boldsymbol{W}_g^{\dagger}(\lambda)) \leq \prod_{i=1}^{s}\left[|\boldsymbol{\Sigma}_i^{-1}|^{\frac{n_i+m_i-q_i-r-1}{2}} \cdot \mathrm{etr}\left(-\frac{1}{2}\boldsymbol{\Sigma}_i^{-1}\left(\boldsymbol{H}_i + (\boldsymbol{Y}_i - \boldsymbol{X}_i\hat{\boldsymbol{B}}_i)'(\boldsymbol{Y}_i - \boldsymbol{X}_i\hat{\boldsymbol{B}}_i)\right)\right)\right]. \tag{52}$$

Because the RHS is clearly integrable with respect to $\boldsymbol{\Sigma}_1^{-1}, \ldots, \boldsymbol{\Sigma}_s^{-1}$, by bounded convergence theorem, it follows that

$$\mathrm{BF}(\boldsymbol{W}_g) = \frac{\int \lim_{\lambda \to 0} K_{H_a}(\boldsymbol{W}_g^{\dagger}(\lambda))\, d\boldsymbol{\Sigma}_1^{-1} \ldots d\boldsymbol{\Sigma}_s^{-1}}{\int K_{H_0}\, d\boldsymbol{\Sigma}_1^{-1} \ldots d\boldsymbol{\Sigma}_s^{-1}}. \tag{53}$$

Because the computation of K_{H_a} does not require inversion of $\boldsymbol{W}_g$ and the matrix sum $(\boldsymbol{I} + \boldsymbol{V}_g^{-1}\boldsymbol{W}_g)$ is guaranteed to be full rank, we conclude that

$$\lim_{\lambda \to 0} K_{H_a}(\boldsymbol{W}_g^{\dagger}(\lambda)) = K_{H_a}(\boldsymbol{W}_g), \tag{54}$$

and

$$\lim_{\lambda \to 0} \mathrm{BF}(\boldsymbol{W}_g^{\dagger}(\lambda)) = \mathrm{BF}(\boldsymbol{W}_g) = \frac{\int K_{H_a}(\boldsymbol{W}_g)\, d\boldsymbol{\Sigma}_1^{-1} \ldots d\boldsymbol{\Sigma}_s^{-1}}{\int K_{H_0}\, d\boldsymbol{\Sigma}_1^{-1} \ldots d\boldsymbol{\Sigma}_s^{-1}}, \tag{55}$$

provided that $\boldsymbol{W}_g$ is positive semidefinite. □

In case $\boldsymbol{W}_g$ is singular, to evaluate the approximate Bayes factor using Laplace's method, we modify the factorization in (31) to account for the imposed linear restrictions. More specifically, we factor K_{H_a} into

$$\begin{aligned} h_a(\boldsymbol{\Sigma}_1^{-1}, \ldots, \boldsymbol{\Sigma}_s^{-1}) = |\boldsymbol{I} + \boldsymbol{V}_g^{-1}\boldsymbol{W}_g|^{-\frac{1}{2}} \cdot \exp\left(\frac{1}{2}\hat{\boldsymbol{\beta}}_g' \left[\boldsymbol{V}_g^{-1}\boldsymbol{W}_g(\boldsymbol{I} + \boldsymbol{V}_g^{-1}\boldsymbol{W}_g)^{-1}\boldsymbol{V}_g^{-1}\right] \hat{\boldsymbol{\beta}}_g\right) \\ \cdot \exp\left(-\frac{1}{2}\sum_{i=1}^{s} \alpha_i \hat{\boldsymbol{\beta}}_{g,i}^{r'} \boldsymbol{V}_{g,i}^{-1} \hat{\boldsymbol{\beta}}_{g,i}^{r}\right), \end{aligned} \tag{56}$$

and

$$\begin{aligned} g_a(\boldsymbol{\Sigma}_1^{-1}, \ldots, \boldsymbol{\Sigma}_s^{-1}) = \sum_{i=1}^{s} \frac{n_i + \nu_i}{2} \log|\boldsymbol{\Sigma}_i^{-1}| \\ - \frac{1}{2}\sum_{i=1}^{s} \mathrm{tr}\left[\boldsymbol{\Sigma}_i^{-1}\left(\nu_i \boldsymbol{H}_i + \alpha_i(\boldsymbol{Y}_i - \boldsymbol{X}_i\hat{\boldsymbol{B}}_i^{r})'(\boldsymbol{Y}_i - \boldsymbol{X}_i\hat{\boldsymbol{B}}_i^{r}) + (1-\alpha_i)(\boldsymbol{Y}_i - \boldsymbol{X}_i\tilde{\boldsymbol{B}}_i)'(\boldsymbol{Y}_i - \boldsymbol{X}_i\tilde{\boldsymbol{B}}_i)\right)\right], \end{aligned}$$

(57)

where $\hat{\boldsymbol{B}}_i^r$ is the least squares estimate of $\boldsymbol{B}_i$ subject to the linear constraints imposed by $\boldsymbol{W}_g$ and $\hat{\boldsymbol{\beta}}_{g,i}^r$ is the corresponding vectorized estimates. The remaining steps for Laplace's method are the same as we have shown in appendix B.2.1, however $\hat{\boldsymbol{\Sigma}}_i$ is now taking the following form:

$$\hat{\boldsymbol{\Sigma}}_i = \frac{1}{n_i}\left(\boldsymbol{H}_i + (\boldsymbol{Y}_i - \boldsymbol{X}_i\hat{\boldsymbol{B}}_i^r)'(\boldsymbol{Y}_i - \boldsymbol{X}_i\hat{\boldsymbol{B}}_i^r)\right). \tag{58}$$

E Numerical Evaluation of Approximate Bayes Factors

We perform numerical experiments to assess the finite-sample accuracies of the derived approximate Bayes factors.

We simulate data under the SSLR model (mainly because its Bayes factors can be numerically evaluated using the adaptive Gaussian quadrature method as the number of groups (s) is small). Except for the very last case, our simulated data sets always have sample size $n = 75$ and subgroup number $s = 3$. We also vary the number of covariates for $p = 2, 4, 8$ and 16 in different simulations.

For each (n, p) combination, we simulate 500 data sets using the SSLR model. We intentionally choose small to modest effect sizes, for which accuracies of the Bayes factors matter most. For every simulated data set, we evaluate its "true value" using the adaptive Gaussian quadrature procedure implemented in the GNU Scientific Library (GSL) and compare it with the ABFs computed under $\boldsymbol{\alpha} = 0, 0.5$ and 1. These results are summarized in Figure 1 and Table 1. As values of $\boldsymbol{\alpha}$ are set to 0.5 for all subgroups, the resulting ABFs yield most accurate approximations in all cases with small sample sizes. In comparison, setting $\boldsymbol{\alpha} = 1$ tends to yield anti-conservative approximations whereas setting $\boldsymbol{\alpha} = 0$ leads to conservative approximations.

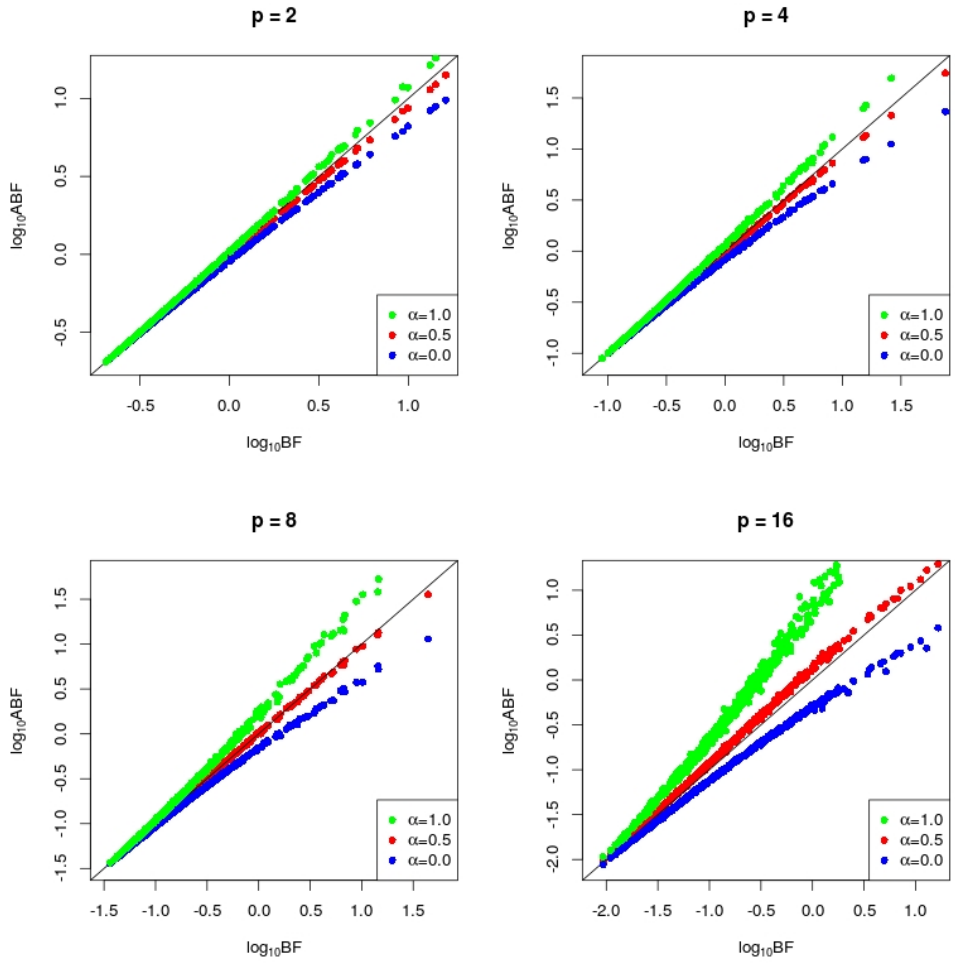

Figure 1: Accuracy of the approximate Bayes factors with small sample sizes. Each data point on the plots represents a single comparison of the ABF of certain $\boldsymbol{\alpha}$ value with the true value using a data set simulated from the SSLR model ($n = 75$ and $s = 3$). The four different panels represent the different numbers of covariates (p) allowed in the model.

Finally, to demonstrate a situation that is close to the preferred asymptotic settings, we simulate data for $n = 1000$ and $p = 16$. The result is shown in Figure 2. It suggests as the sample size increases, all approximations become quite accurate.

	RMSE of $\log_{10}$(ABF)		
	$\boldsymbol{\alpha} = 0$	$\boldsymbol{\alpha} = 0.5$	$\boldsymbol{\alpha} = 1.0$
$n = 75, p = 2$	0.032	0.009	0.016
$n = 75, p = 4$	0.052	0.011	0.041
$n = 75, p = 8$	0.074	0.008	0.096
$n = 75, p = 16$	0.102	0.035	0.268
$n = 1000, p = 16$	0.044	0.006	0.032

Table 1: Root Mean Square Errors (RMSE) of $\log_{10}$(ABF) for different $\boldsymbol{\alpha}$ values under different model settings. The approximate Bayes factors are computed based on the SSLR model with three subgroups ($s = 3$) and different (n, p) settings. Under each setting, we compute $\log_{10}$(ABF) for $\boldsymbol{\alpha} = 0, 0.5, 1.0$ and report the RMSE by comparing the approximations with the true values.

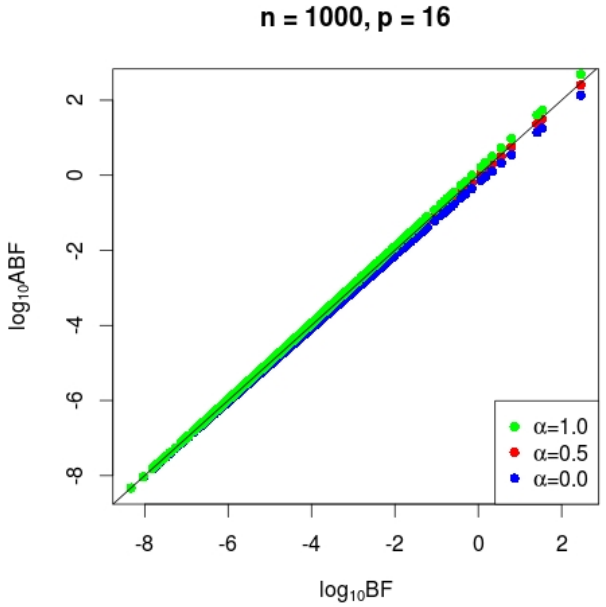

Figure 2: Accuracy of the approximate Bayes factors when the sample size is relatively large. In this plot, the data simulated from the SSLR model with $n = 1000, s = 3$ and $p = 16$. Approximate Bayes factors computed using different $\boldsymbol{\alpha}$ values all show good agreement with the true values.

F Bayes Factor, Multivariate Test Statistics and the BIC

In this section, we show that the derived Bayes factor and its approximations are connected to various frequentist multivariate test statistics and the BIC.

F.1 Connection to Multivariate Test Statistics

Under the following prior specification for $\boldsymbol{W}_g$:

1. $\boldsymbol{W}_g = c\boldsymbol{V}_g$, where c is a positive scalar constant.
2. $\boldsymbol{V}_g$ is full-rank
3. $\boldsymbol{\Sigma}_i \sim \text{IW}(\nu_i \boldsymbol{H}_i, m_i)$ under the limiting conditions, $\nu_i \to 0, \forall i = 1, \ldots, s$

It can be shown that

$$\text{ABF}(\boldsymbol{W}_g, \boldsymbol{\alpha} = 1) = \left(\sqrt{\frac{1}{1+c}}\right)^{rps} \cdot \exp\left(\frac{1}{2} \cdot \frac{c}{1+c} \cdot T_{\text{wald}}\right), \tag{59}$$

and

$$\text{ABF}(\boldsymbol{W}_g, \boldsymbol{\alpha} = 0) = \left(\sqrt{\frac{1}{1+c}}\right)^{rps} \cdot \exp\left(\frac{1}{2} \cdot \frac{c}{1+c} \cdot T_{\text{score}}\right), \tag{60}$$

where T_{wald} and T_{score} represent the multivariate Wald statistic and the Rao's score statistic, respectively. Both statistics can be used for testing $H_0 : \boldsymbol{\beta}_g = 0$ based on the SSMR model. Obtaining (59) is straightforward. To establish (60), we compute the score statistic following Chen (1983). This yields

$$\begin{aligned} T_{\text{score}} &= \sum_{i=1}^{s} \text{vec}[(\boldsymbol{Y}_i - \boldsymbol{X}_{c,i}\tilde{\boldsymbol{B}}_i)']' \left(\boldsymbol{X}'_{g,i}\boldsymbol{X}_{g,i} \otimes \tilde{\boldsymbol{\Sigma}}_i^{-1}\right) \text{vec}[(\boldsymbol{Y}_i - \boldsymbol{X}_{c,i}\tilde{\boldsymbol{B}}_i)'] \\ &= \tilde{\boldsymbol{\beta}}_c' \left[\oplus_{i=1}^{s} \left(\boldsymbol{X}'_{g,i}\boldsymbol{X}_{g,i} \otimes \tilde{\boldsymbol{\Sigma}}_i^{-1}\right)\right] \tilde{\boldsymbol{\beta}}_c, \end{aligned} \tag{61}$$

where $\tilde{\boldsymbol{B}}_i$ and $\tilde{\boldsymbol{\beta}}_c$ are MLEs of $\boldsymbol{B}_i$ and $\boldsymbol{\beta}_c$ estimated from the null model, respectively. Under the specified conditions,

$$\begin{aligned}
&\hat{\boldsymbol{\beta}}_g' \tilde{\boldsymbol{V}}_g^{-1} \left[\tilde{\boldsymbol{W}}_g (\boldsymbol{I} + \tilde{\boldsymbol{V}}_g^{-1} \tilde{\boldsymbol{W}}_g)^{-1} \right] \tilde{\boldsymbol{V}}_g^{-1} \hat{\boldsymbol{\beta}}_g \\
=&\frac{c}{1+c} \cdot \hat{\boldsymbol{\beta}}_g' \tilde{\boldsymbol{V}}_g^{-1} \hat{\boldsymbol{\beta}}_g \\
=&\frac{c}{1+c} \cdot \tilde{\boldsymbol{\beta}}_c' \left[\oplus_{i=1}^{s} \left(\boldsymbol{X}_{g,i}' \boldsymbol{X}_{g,i} \otimes \tilde{\boldsymbol{\Sigma}}_i^{-1} \right) \right] \tilde{\boldsymbol{\beta}}_c.
\end{aligned}$$

As a consequence, the approximate Bayes factors and the corresponding frequentist test statistics yield the same ranking for a set of candidate models.

Albeit the connections, we do not advocate the use of these test statistics as model comparison devices in practice. Especially, caution should be taken when interpreting this prior in specific contexts: for example, Wakefield (2009) and Wen and Stephens (2011) have shown some counter-intuitive implications of this prior in genetic applications (e.g., $|\boldsymbol{W}_g|$ is inversely proportional to sample sizes).

F.2 Connections to the BIC

Under the conditions that

1. $\boldsymbol{V}_g$ and $\boldsymbol{W}_g$ are full-rank.
2. $\lim_{n_i \to 0} \frac{\log |\boldsymbol{W}_g|}{n_i} = 0, \forall i.$
3. $n_i \gg p, r, s, \forall i,.$

We show that the BIC can be derived as a rough approximation to the Bayes factor and its approximations under the SSMR model.

First, we assume that

$$\lim_{n_i \to \infty} \frac{1}{n_i} \left(\boldsymbol{X}'_{g,i} \boldsymbol{X}_{g,i} - \boldsymbol{X}'_{g,i} \boldsymbol{X}_{c,i} (\boldsymbol{X}'_{c,i} \boldsymbol{X}_{c,i})^{-1} \boldsymbol{X}'_{c,i} \boldsymbol{X}_{g,i} \right) = \boldsymbol{Q}_i, \tag{62}$$

and Q_i is also full-rank. Hence,

$$\lim_{n_i \to \infty} \boldsymbol{V}_g = \oplus_{i=1}^{s} \left[\frac{1}{n_i} \left(\boldsymbol{Q}_i^{-1} \otimes \boldsymbol{\Sigma}_i \right) \right]. \tag{63}$$

When $\boldsymbol{\mathcal{E}}$ is known, as $n_i \to \infty$ for each i, based on (63)

$$\lim_{n_i \to \infty, \forall i} \left(\boldsymbol{I} + \boldsymbol{V}_g^{-1} \boldsymbol{W}_g \right) = \boldsymbol{V}_g^{-1} \boldsymbol{W}_g, \tag{64}$$

and

$$\lim_{n_i \to \infty, \forall i} \mathrm{BF}(\boldsymbol{W}_g) = |\boldsymbol{V}_g|^{1/2} \cdot |\boldsymbol{W}_g|^{-1/2} \cdot \exp\left(\frac{1}{2} \hat{\boldsymbol{\beta}}'_g \boldsymbol{V}_g^{-1} \hat{\boldsymbol{\beta}}_g \right). \tag{65}$$

Note that

$$\lim_{n_i \to \infty} |\boldsymbol{V}_g| = \prod_{i=1}^{s} \left(n_i^{-pr} \cdot |\boldsymbol{Q}_i|^{-r} \cdot |\boldsymbol{\Sigma}_i|^{p} \right), \tag{66}$$

and the likelihood ratio

$$L_1 / L_0 = \frac{p(\boldsymbol{\mathcal{Y}} | \boldsymbol{\mathcal{X}}, \hat{\boldsymbol{\mathcal{B}}}, \boldsymbol{\mathcal{E}})}{p(\boldsymbol{\mathcal{Y}} | \boldsymbol{\mathcal{X}}, \tilde{\boldsymbol{\mathcal{B}}}, \boldsymbol{\mathcal{E}}, H_0)} = \exp\left(\frac{1}{2} \hat{\boldsymbol{\beta}}'_g \boldsymbol{V}_g^{-1} \hat{\boldsymbol{\beta}}_g \right) \tag{67}$$

It follows that

$$\begin{aligned}\log \mathrm{BF}(\boldsymbol{W}_g) &\approx (\log L_1 - \log L_0) - \frac{pr}{2}\sum_{i=1}^{s}\log n_i + \left(\frac{r}{2}\sum_{i=1}^{s}\log|\boldsymbol{Q}_i| - \frac{p}{2}\sum_{i=1}^{s}\log|\boldsymbol{\Sigma}_i| - \frac{1}{2}\log|\boldsymbol{W}_g|\right) \\ &= (\log L_1 - \log L_0) - \frac{pr}{2}\sum_{i=1}^{s}\log n_i + O(1), \\ &= \mathrm{BIC} + O(1).\end{aligned} \tag{68}$$

The BIC is asymptotically consistent, meaning that as sample size increases to infinity and under other suitable conditions, the BIC selects the fixed true model among a finite set of candidates with probability 1 (Haughton (1988), Schwarz (1978)). Consequently, our Bayes factor and its approximations also enjoy this asymptotic consistency property.

It is worth pointing out that the BIC is not a universal approximation of Bayes factors. In our case, BIC fails to approximate desired Bayes factors with the advocated error bound if the pre-specified conditions are violated. In particular,

1. $\boldsymbol{W}_g$ or $\boldsymbol{V}_g$ is singular. Intuitively, in this case, linear constraints on parameter space would change the way that "free" parameters are counted. Nonetheless, it is usually possible to resolve the linear constraints by transformation and re-parametrization.

2. $\boldsymbol{W}_g$ is some function of sample sizes, e.g., this may lead that $\lim_{n_i\to 0}\frac{\log|\boldsymbol{W}_g|}{n_i} \neq 0$, for some i. An example of this sort is the prior specification, $\boldsymbol{W}_g = c\boldsymbol{V}_g$. It is easy to see that BIC fails to approximate the resulting Bayes factor with the advocated error bound.

3. Parameters p, r and s are *not* small comparing with sample sizes. In particular, under the high-dimensional settings, the BIC becomes a very poor approximation of the desired Bayes factor.

When the $\boldsymbol{\mathcal{E}}$ is unknown, it can be shown that

$$\log \mathrm{ABF}(\boldsymbol{W}_g, \boldsymbol{\alpha}) \approx \frac{1}{2}\hat{\boldsymbol{\beta}}_g' \check{\boldsymbol{V}}_g^{-1} \hat{\boldsymbol{\beta}}_g - \frac{pr}{2}\sum_{i=1}^{s} \log n_i + \left(\frac{r}{2}\sum_{i=1}^{s} \log |\boldsymbol{Q}_i| - \frac{p}{2}\sum_{i=1}^{s} \log |\check{\boldsymbol{\Sigma}}_i| - \frac{1}{2}\log |\check{\boldsymbol{W}}_g| \right). \tag{69}$$

In particular,

$$\log \mathrm{ABF}(\boldsymbol{W}_g, \boldsymbol{\alpha} = 1) \approx \frac{1}{2}\hat{\boldsymbol{\beta}}_g' \hat{\boldsymbol{V}}_g^{-1} \hat{\boldsymbol{\beta}}_g - \frac{pr}{2}\sum_{i=1}^{s} \log n_i + \left(\frac{r}{2}\sum_{i=1}^{s} \log |\boldsymbol{Q}_i| - \frac{p}{2}\sum_{i=1}^{s} \log |\hat{\boldsymbol{\Sigma}}_i| - \frac{1}{2}\log |\hat{\boldsymbol{W}}_g| \right), \tag{70}$$

Asymptotically, under the conditions stated

$$\lim_{n_i \to \infty, \forall i} \hat{\boldsymbol{\beta}}_g' \hat{\boldsymbol{V}}_g^{-1} \hat{\boldsymbol{\beta}}_g \to \hat{\boldsymbol{\beta}}_g \boldsymbol{V}_g^{-1} \hat{\boldsymbol{\beta}}_g. \tag{71}$$

Furthermore, it can be shown that

$$\lim_{n_i \to \infty} \tilde{\boldsymbol{\Sigma}}_i = \hat{\boldsymbol{\Sigma}}_i + \hat{\boldsymbol{B}}_{g,i}' \boldsymbol{Q}_i \hat{\boldsymbol{B}}_{g,i}. \tag{72}$$

In general, this ensures that

$$\hat{\boldsymbol{\beta}}_g' \check{\boldsymbol{V}}_g^{-1} \hat{\boldsymbol{\beta}}_g = \hat{\boldsymbol{\beta}}_g' \boldsymbol{V}_g^{-1} \hat{\boldsymbol{\beta}}_g + O(1). \tag{73}$$

This yields our final results: under the conditions stated

$$\log \mathrm{ABF}(\boldsymbol{W}_g, \boldsymbol{\alpha}) = (\log L_1 - \log L_0) - \frac{pr}{2}\sum_{i=1}^{s} \log n_i + O(1). \tag{74}$$

G Extension to Non-normal Data

Without loss of generality, we consider a system of generalized linear models which resembles the SSLR. The MLE of the system can be numerically computed for the vectorized regression coefficients $\boldsymbol{\beta}_{\text{sys}}$. Following the standard asymptotic maximum likelihood theory, the likelihood of the system can be approximated by a quadratic expansion around its maximum likelihood estimate. This can be equivalently expressed by the following asymptotic approximation,

$$\hat{\boldsymbol{\beta}}_{\text{sys}} \,|\, \boldsymbol{\beta}_{\text{sys}} \sim \text{N}\left(\boldsymbol{\beta}_{\text{sys}}\,,\, \text{Var}(\hat{\boldsymbol{\beta}}_{\text{sys}})\right), \tag{75}$$

where $\text{Var}(\hat{\boldsymbol{\beta}}_{\text{sys}})$ is typically approximated using observed Fisher information. Combining with the prior distribution

$$\boldsymbol{\beta}_{\text{sys}} \sim \text{N}(\mathbf{0}\,,\, \boldsymbol{\Psi}_c \oplus \boldsymbol{W}_g), \tag{76}$$

it is then straightforward to show that the resulting Bayes factor under this setting maintains the same functional form as in LEMMA 1.

H MCMC Algorithm for Model Selection in MVLR

We implement an Markov Chain Monte Carlo (MCMC) algorithm to generate samples for posterior analysis of $\xi(\boldsymbol{\beta}_g)$. Here we detail the algorithm for the MVLR model, and point out that generalizing this algorithm for the general SSMR model is trivial.

H.1 Description of Algorithm

In the SSMR model, the posterior distribution of $\xi(\boldsymbol{\beta}_g)$ is given by

$$\begin{aligned} \Pr(\xi(\boldsymbol{\beta}_g) \mid \boldsymbol{Y}, \boldsymbol{X}) &\propto \Pr(\xi(\boldsymbol{\beta}_g)) \cdot p(\boldsymbol{Y} \mid \xi(\boldsymbol{\beta}_g), \boldsymbol{X}) \\ &\propto \Pr(\xi(\boldsymbol{\beta}_g)) \cdot \mathrm{BF}(\xi(\boldsymbol{\beta}_g)). \end{aligned} \tag{77}$$

In the main text, we have discussed the computation of $\mathrm{BF}(\xi(\boldsymbol{\beta}_g))$. Assuming the prior distribution $\Pr(\xi(\boldsymbol{\beta}_g))$ is provided and easy to compute, it is straightforward to apply the Metropolis-Hastings algorithm. The practical difficulty in applying this algorithm in high-dimensional settings is to find an efficient proposal distribution to ensure the fast mixing of the Markov chain.

In solving Bayesian variable selection problem in the multiple linear regression context, Guan and Stephens (2011) proposed a novel proposal distribution that prioritizes updates on variables showing strong marginal associations, an idea related to the *sure-independence screening* (Fan and Lv (2008)). We generalize their idea in the context of the SSMR model. In our implementation, we utilize two types of simple "local" proposal updates:

1. changing the configuration of a candidate covariate.
2. swapping the configurations of two different covariates.

More specifically, each covariate i is proposed according to a weight w_i computed by

$$w_i = \sum_{j=1}^{n-1} p_j \mathrm{BF}_i^{[j]} + p_n. \tag{78}$$

The quantity $\mathrm{BF}_i^{[j]}$ represents the single-variate Bayes factor of covariate i obtained by averaging (equally) over its all non-zero configuration Bayes factors and controlling for previously identified $(j-1)$ top association signals. We construct the weights by starting with an empty set of controlling covariates and compute the single covariate Bayes factors; we then select the covariate

with the highest marginal Bayes factor into the set of covariates to be controlled for in the next round; we repeat this procedure $(n-1)$ times and in the n-th round, we simply assign each covariate uniform weight. Finally, we combine these weights into w_i by a sequence of non-increasing probabilities $p_1 > p_2 > ... > p_n$. The general idea of this proposal distribution is largely due to Matthew Stephens (personal communication). In the simulation and data application examples of this paper, we set $n = 4$ and $p_1 = 0.624,\ p_2 = 0.250,\ p_3 = 0.125,\ p_4 = 0.010$. In practice, once a SNP is proposed, we randomly assign 85% of the proposals to move type 1 and the 15% of the proposals to move type 2.

In addition, when processing the posterior samples to compute posterior inclusion probabilities of covariates, we utilize Rao-Blackwellization techniques to reduce Monte Carlo variance of the estimates.

H.2 Convergence Diagnostics

We describe two convergence diagnostics of the proposed MCMC algorithm in this section. The first method is a direct adaption of Brooks *et al.* (2003), which is a formal convergence testing procedure and requires running multiple chains. The other informal diagnostic we found useful is to utilize (77), which essentially is the posterior model probability up to a unknown normalizing constant. For each MCMC run, we compute the rank correlation between the posterior sampling frequencies and corresponding posterior scores for the sampled models. When the MCMC algorithm reaches convergence, we expect this correlation is high for the top ranked posterior models. Our observation is that the rank correlation is indeed high, the formal testing of convergence usually becomes redundant and can be avoided. As a result, it reduces the computational burden to run multiple Markov chains.

H.3 Computational Benchmark

We benchmark the computational performance of the MCMC algorithm (implemented in C++) analyzing the imputed SNP data set of Gene C21orf57. The data set contains 4797 SNPs, 75 individuals and expression levels from three cell types. The program is running on a computer with 8-core Intel Xeon 2.13GHz processors and uses 25 Megabtypes of memory space. For 25000 burning steps and 50000 MCMC repeats, the full computation takes 14 minutes 22 seconds real time.

I Additional Simulation Results

We perform additional simulation studies to fully investigate the difference in performance of BMS and LASSO. In the end, we identify two primary factors that may explain the observed performance patterns:

1. the correlation structure of the random errors in the MVLR model.
2. the prior correlation information of non-zero regression coefficients.

Notably, vanilla version of the LASSO algorithm takes account of neither. To evaluate their individual effects on model selection, we simulate additional data for $n = 100, p = 250$ and $r = 3$ under the MVLR model, for which all candidate covariates are independently generated.

I.1 Impact of Error Variance Matrix

We first investigate the impact of the error variance on model selection. To do so, we simulate independent regression coefficients across subgroups for each selected covariate, but alter the

error variance matrix Σ for the MVLR model. In particular, we use the following three different settings for the Σ matrix:

1. $\boldsymbol{\Sigma} = \sigma^2 \boldsymbol{I}$
2. $\boldsymbol{\Sigma}$ is diagonal, but the diagonal elements are unequal (i.e., unequal error variances in different subgroups).
3. $\boldsymbol{\Sigma}$ has non-zero correlations between subgroups and unequal diagonal elements.

In all three settings, we run both LASSO and BMS on the simulated data sets. For BMS, we assume the prior effect sizes are independent within each covariate in all cases; and in specifying the Wishart prior for $\boldsymbol{\Sigma}$, we set $\boldsymbol{H} \to 0$ and $\nu \to 0$ (i.e. $\boldsymbol{\Sigma}$ are directly estimated from the data with essentially no prior influence).

We plot the trade-off between the true positives and false positives from both methods in Figure 3. Our result indicates that when $\boldsymbol{\Sigma} = \sigma^2 \boldsymbol{I}$, the two methods perform very similarly. However, as the true Σ departs further away from the diagonal and equal variance structure, the performance of LASSO becomes worse. In comparison, the performance of BMS is stable in all three settings.

Rothman *et al.* (2010) also discovered the structure of $\boldsymbol{\Sigma}$ matrix has significant impacts on the performance of regularized model selection method. As a remedy, they propose to regularize $\boldsymbol{\Sigma}$ matrix jointly with $\boldsymbol{\beta}$ in the L_1 penalty term. However in our context, $\boldsymbol{\Sigma}$ is considered to be low dimensional ($r = 3$) and the motivation to regularize $\boldsymbol{\Sigma}$ is unclear to us.

I.2 Importance of Prior Information

We also examine the importance of utilizing prior correlation information of non-zero coefficients on the performance of model selection. Again, we limited our comparisons to BMS and LASSO

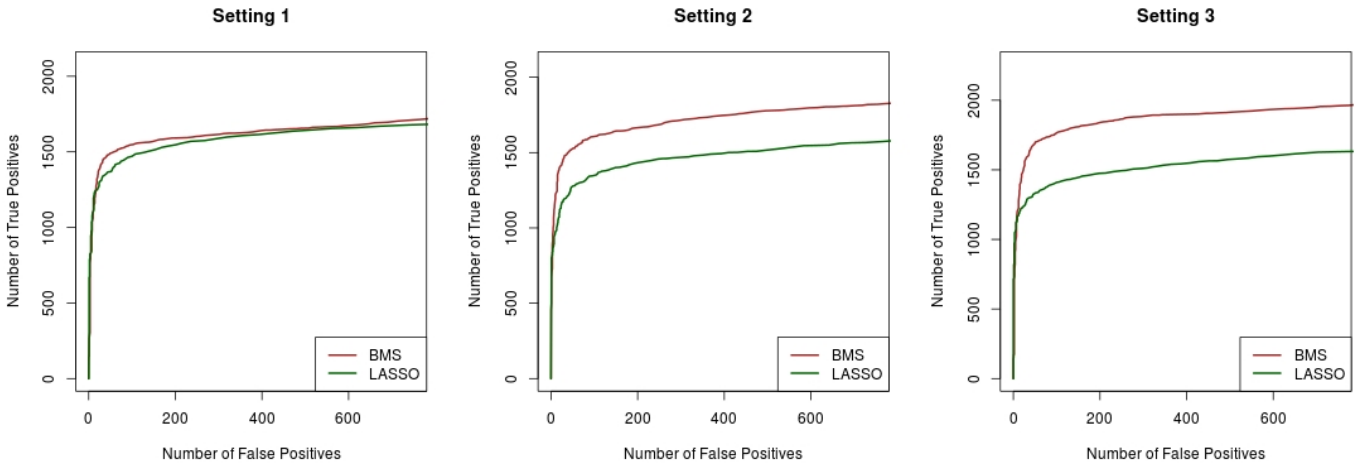

Figure 3: Evaluation of impact of error variance on model selection methods. In setting 1, the true $\mathbf{\Sigma} = \sigma^2 \boldsymbol{I}$; in setting 2, the true $\mathbf{\Sigma}$ is diagonal but with unequal diagonal elements; in setting 3, the true $\mathbf{\Sigma}$ has the most general form, with non-zero positive correlations and unequal diagonal elements. BMS has similar performance across the three settings, LASSO seems performing worse when the true $\mathbf{\Sigma}$ departs further away from $\sigma^2 \boldsymbol{I}$.

using only simulated independent covariate data. Furthermore, we use $\mathbf{\Sigma} = \sigma^2 \boldsymbol{I}$ to generate random errors for the MVLR model in this part of the simulation study.

We create two different schemes in generating regression coefficients. The first scheme is the same as we described in the main text (i.e., conditioning on a non-zero configuration, $\boldsymbol{\gamma} = (111)$ is with probability 0.50 and others are equally likely). In the second scheme, we assign the activity configuration $\boldsymbol{\gamma} = (111)$ with probability 1 to the selected covariate. For the ith selected covariate, the effect sizes in the three subgroups are subsequently simulated from $\mathrm{N}(\bar{\beta}_i, \frac{\bar{\beta}_i^2}{100})$, where $\bar{\beta}_i$ is drawn from a $\mathrm{N}(0, 1)$ distribution. The resulting correlation structure of regression coefficients is most similar to what have been observed in a meta-analysis.

We run both BMS and LASSO on 200 data sets simulated in each scheme. To specify the distribution of non-zero activity configurations for BMS, we use both the default "objective" prior (which assigns equal probability mass to each non-zero activity configuration) and the "perfect" prior (which is the true generative distribution of the simulation data sets).

We show the simulation results in Figure 4 by plotting the trade-off between the true positives and the false positives for each method. The results show that BMS with perfect prior information always achieves the best performance. (We again emphasize that in many genomic applications, it is possible to accurately estimate this "perfect" prior from data, see examples from Flutre *et al.* (2013)). Although the "objective" prior is clearly not optimal, because it captures the correlations between non-zero effects within a covariate, it still outperforms LASSO in both cases. Finally, we expect a prior assuming independence of effects of regression coefficients (i.e. a diagonal $\mathbf{\Gamma}_g$ matrix) will behave similarly to LASSO, based on our observation in setting 1 of Figure 3. Therefore, we conclude that the performance of model selection methods are likely to have significant improvement if the *a priori* information in data can be accurately utilized.

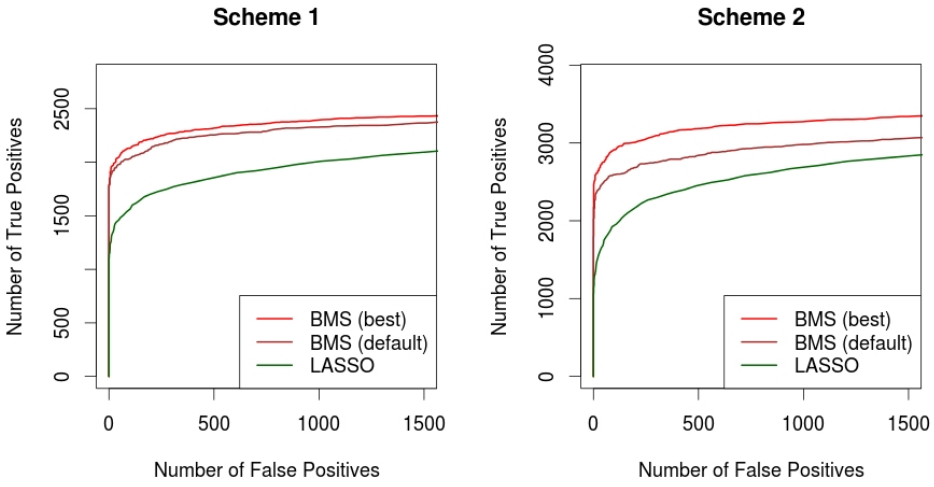

Figure 4: Evaluation of impact of prior information on model selection methods. Scheme 1 and 2 correspond to two distinct generating distributions used for simulating data. BMS(best) is our Bayesian model selection method using the true generative distribution as the prior, whereas BMS(default) uses an "objective" prior. In scheme 1, the objective prior is "closer" to the truth than in scheme 2. LASSO does not utilize the prior correlation information and essentially assumes that the regression coefficients are *a priori* independent.

J Single SNP Analysis Result for Gene C21orf57

In this section, we show the single SNP analysis results of the eQTL mapping for gene C21orf57 using Dimas data. More specifically, the aim is to examine the results of the tissue specificity inference from our BMS approach.

As a visual diagnostic, we first fit a simple linear regression model for each SNP in each cell type, we then examine the resulting regression coefficients across all three cell types for each SNP using a forest plot. We show the results for the three distinct signals identified by the BMS approach in Figure 5. By this simple diagnostic, the tissue specificity inference seems intuitively sensible.

References

Brooks, S. P., Giudici, P., and Philippe, A. (2003). Nonparametric Convergence Assessment for MCMC Model Selection. *Journal of Computational and Graphical Statistics*, **12**(1), 1–22.

Butler, R. (2007). *Saddlepoint Approximations with Applications*. Cambridge University Press, 1st edition.

Chen, C.-F. (1983). Score Tests for Regression Models. *Journal of the American Statistical Association*, **78**(381), 158–161.

Fan, J. and Lv, J. (2008). Sure independence screening for ultrahigh dimensional feature space (with discussion). *Journal of the Royal Statistical Society - Series B: Statistical Methodology*, **70**(5), 849–911.

Flutre, T., Wen, X., Pritchard, J. K., and Stephens, M. (2013). A statistical framework for joint eQTL analysis in multiple tissues. *PLoS Genetics*, **9**(5), e1003486.

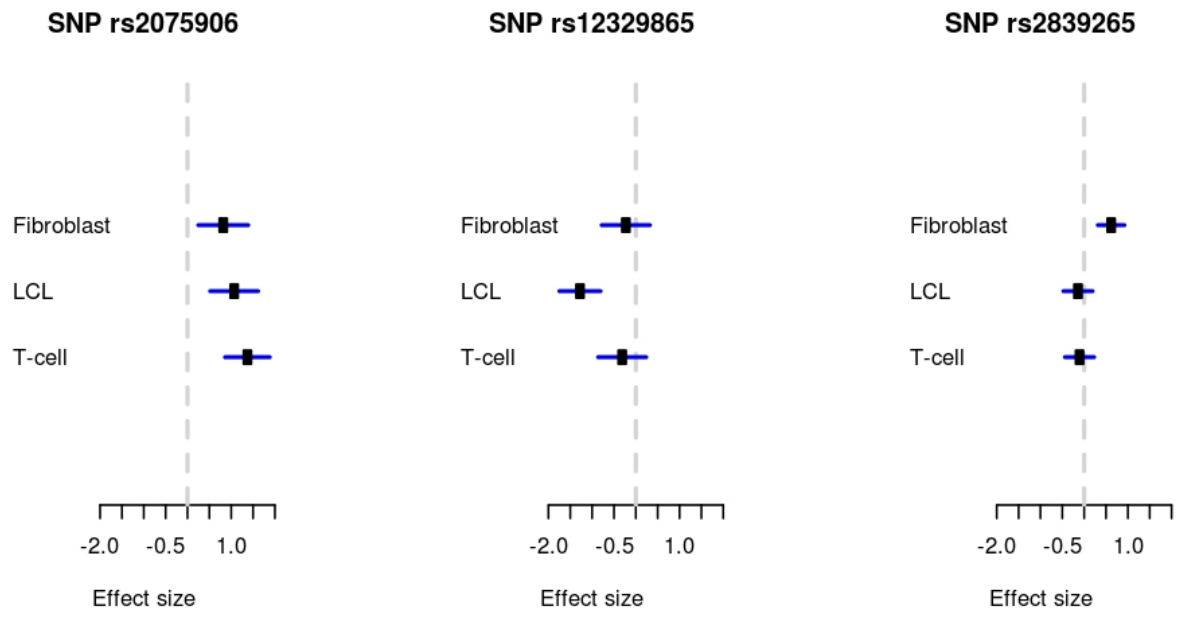

Figure 5: Examining the single-SNP effects of each identified eQTL in each separate cell type for Gene C21orf57. For each SNP in each cell type, we obtain the estimates of the effect size and its standard error by fitting a simple linear regression model. The estimated effect sizes and their corresponding 95% confidence intervals are plotted for different cell types in a forest plot for each SNP. SNP rs2839265 has shorter intervals because its minor allele frequency (0.28) is greater than the frequencies of rs2075906 (0.09) and rs12329865 (0.11). Recall our method infers that rs2075906 is tissue consistent; SNPs rs12329865 and rs2839265 are LCL-specific and Fibroblast-specific eQTLs, respectively.

Guan, Y. and Stephens, M. (2011). Bayesian variable selection regression for genome-wide association studies, and other large-scale problems. *Annals of Applied Statistics*, **5**(3), 1780–1815.

Haughton, D. M. A. (1988). On the Choice of a Model to Fit Data from an Exponential Family. *The Annals of Statistics*, **16**(1), 342–355.

McCullagh, P. and Nelder, J. (1989). *Generalized Linear Models*. Chapman and Hall, 2nd edition.

Minka, T. P. (2000). Old and new matrix algebra useful for statistics.

Rothman, A. J., Levina, E., and Zhu, J. (2010). Sparse Multivariate Regression With Covariance Estimation. *Journal of Computational and Graphical Statistics*, **19(4)**, 947–962.

Schwarz, G. E. (1978). Estimating the dimension of a model. *Annals of Statistics*, **6(2)**, 461–464.

Scott-Boyer, M. P., Imholte, G. C., Tayeb, A., Labbe, A., Deschepper, C. F., and Gottardo, R. (2012). An integrated hierarchical Bayesian model for multivariate eQTL mapping. *Statistical Applications in Genetics and Molecular Biology*, **11**(4).

Stephens, M. (2010). A unified framework for testing multiple phenotypes for association with genetic variants. *Presented at the 60th Annual Meeting of The American Society of Human Genetics, Washington D.C.*

Stingo, F. C., Chen, Y. A., Tadesse, M. G., and Vannucci, M. (2011). Incorporating biological information into linear models: A Bayesian approach to the selection of pathways and genes. *Annals of Applied Statistics*, **5**(3), 1978–2002.

Veyrieras, J., Kudaravalli, S., Kim, S. Y., Dermitzakis, E. T., Gilad, Y., Stephens, M., and Pritchard, J. K. (2008). High-resolution mapping of expression-qtls yields insight into human gene regulation. *PLoS Genetics*, **4**(10).

Wakefield, J. (2009). Bayes factors for genome-wide association studies: comparison with p-values. *Genetic Epidemiology*, **33**(1), 79–86.

Wen, X. and Stephens, M. (2011). Bayesian methods for genetic association analysis with heterogeneous subgroups: from meta-analyses to gene-environment interactions. *arXiv pre-print: 1111.1210*.