# Supporting Information for
# "Bayesian compositional regression with structured priors for microbiome feature selection" by

Liangliang Zhang[1], Yushu Shi[2], Robert R. Jenq[3], Kim-Anh Do[1], and Christine B. Peterson[1]

[1] *Department of Biostatistics, University of Texas MD Anderson Cancer Center*

[2] *Department of Statistics, University of Missouri*

[3] *Department of Genomic Medicine, University of Texas MD Anderson Cancer Center*

**Contents**

## S1. DETAILS ON CONTRAST TRANSFORMATION

In this section, we provide the explicit form of the contrast transformation matrix $\boldsymbol{T}$ which corresponds to the additive log-ratio and centered log-ratio transformations. As in the main text, we let $\boldsymbol{U} = (u_{ij})$ represent the observed relative abundances, and $\boldsymbol{Z} = (\log u_{ij})$ represent their log-transformed values. Then the linear model can be expressed as

$$\boldsymbol{y} = \boldsymbol{Z}\boldsymbol{T}\boldsymbol{\theta} + \boldsymbol{\varepsilon} = \boldsymbol{X}\boldsymbol{\theta} + \boldsymbol{\varepsilon},$$

where $\boldsymbol{X} = \boldsymbol{Z}\boldsymbol{T}$ and $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_{p-1})$. In other words, the parameter space degenerates to $p-1$ dimensions after the contrast transformation $\boldsymbol{T}$ is performed on $\boldsymbol{Z}$.

For the additive log-ratio (ALR) transformation, the transformation matrix $\boldsymbol{T}$ is given as

$$\boldsymbol{T} = \begin{bmatrix} \boldsymbol{I}_{(p-1)\times(p-1)} \\ -\boldsymbol{1}_{p-1}' \end{bmatrix}_{p\times(p-1)},$$

where $\boldsymbol{I}_{(p-1)\times(p-1)}$ is a $(p-1)\times(p-1)$ identity matrix and $\boldsymbol{1}'_{p-1}$ is a $p-1$ dimensional row vector of 1s, then the transformed version of $\boldsymbol{Z}$ will be $\boldsymbol{X}_{n\times(p-1)} = \{\log(u_{ij}/u_{ip})\}$, where $i = 1, \ldots, n$. The position of the row $-\boldsymbol{1}'_{p-1}$ in matrix $\boldsymbol{T}$ determines which variable will be the reference. Here the last variable $\boldsymbol{u}_p$ is chosen as the reference, as $-\boldsymbol{1}'_{p-1}$ is the last row of $\boldsymbol{T}$.

For the centered log-ratio (CLR) transformation, the transformation matrix $\boldsymbol{T}$ is given as

$$\boldsymbol{T} = \begin{bmatrix} \boldsymbol{D}_{(p-1)\times(p-1)} \\ -\frac{1}{p} \times \boldsymbol{1}'_{p-1} \end{bmatrix}_{p\times(p-1)},$$

where $\boldsymbol{D}_{(p-1)\times(p-1)}$ is a $(p-1)\times(p-1)$ square matrix with diagonal elements $1-\frac{1}{p}$ and off-diagonal elements $-\frac{1}{p}$. Then the transformed version of $\boldsymbol{Z}$ will be $\boldsymbol{X}_{n\times(p-1)} = (x_{ij})$, where

$$x_{ij} = \log(u_{ij}) - \frac{1}{p}\sum_{k=1}^{p}\log(u_{ik}),\ \ i = 1, 2, \cdots, n;\ j = 1, 2, \cdots, p-1.$$

For each sample $i$, the transformation of each variable $j$ has the same reference $\frac{1}{p}\sum_{k=1}^{p}\log(u_{ik})$, which means that the CLR transformed results do not require the specification of a reference, and are generally more stable than ALR transformed ones.

## S2. MOTIVATION FOR AND PROPERTIES OF OF THE Z-PRIOR

$\boldsymbol{T}_\gamma$ is a $(p+1)\times p$ matrix with rank $p$. Although $\boldsymbol{T}_{\boldsymbol{\gamma}}\boldsymbol{\beta}_{\boldsymbol{\gamma}}$ consists of $p+1$ random variables, its distribution degenerates to $p$ dimensions. Therefore, we can assume $\boldsymbol{T}_{\boldsymbol{\gamma}}\boldsymbol{\beta}_{\boldsymbol{\gamma}}|\ldots \sim \mathcal{N}(0, \sigma^2\tau^2\boldsymbol{I}_{p_\gamma})$. If we multiply $\boldsymbol{T}_{\boldsymbol{\gamma}}\boldsymbol{\beta}_{\boldsymbol{\gamma}}$ by $\boldsymbol{T}'_\gamma$ on the left, then we obtain $\boldsymbol{T}'_\gamma\boldsymbol{T}_\gamma\boldsymbol{\beta}_{\boldsymbol{\gamma}}|\ldots \sim \mathcal{N}(0, \sigma^2\tau^2\boldsymbol{T}'_\gamma\boldsymbol{T}_\gamma)$. As $\boldsymbol{T}'_\gamma\boldsymbol{T}_\gamma$ is a square matrix and invertible, if we then multiply $\boldsymbol{T}'_\gamma\boldsymbol{T}_\gamma\boldsymbol{\beta}_{\boldsymbol{\gamma}}$ by $(\boldsymbol{T}'_\gamma\boldsymbol{T}_\gamma)^{-1}$ on the left, then we obtain $\boldsymbol{\beta}_{\boldsymbol{\gamma}}|\ldots \sim \mathcal{N}(0, \sigma^2\tau^2(\boldsymbol{T}'_\gamma\boldsymbol{T}_\gamma)^{-1})$.

If we multiply $\boldsymbol{\beta}_{\boldsymbol{\gamma}}$ by $\boldsymbol{T}_\gamma$ on the left, then we have $\boldsymbol{T}_{\boldsymbol{\gamma}}\boldsymbol{\beta}_{\boldsymbol{\gamma}}|\ldots \sim \mathcal{N}(0, \sigma^2\tau^2\boldsymbol{T}_\gamma(\boldsymbol{T}'_\gamma\boldsymbol{T}_\gamma)^{-1}\boldsymbol{T}'_\gamma)$. If we do the spectral decomposition of $\boldsymbol{T}_\gamma(\boldsymbol{T}'_\gamma\boldsymbol{T}_\gamma)^{-1}\boldsymbol{T}'_\gamma = \boldsymbol{U}\boldsymbol{D}\boldsymbol{U}^{-1}$, the last eigenvalue in $\boldsymbol{D}$ will be zero and all the other eigenvalues will be one (assuming the eigenvalues are ordered in decreasing order). This rank deficiency of the covariance matrix is equivalent to the singular distribution of the random variable $\boldsymbol{T}_{\boldsymbol{\gamma}}\boldsymbol{\beta}_{\boldsymbol{\gamma}}$ in $(p+1)$ dimensions, because $\boldsymbol{T}_{\boldsymbol{\gamma}}\boldsymbol{\beta}_{\boldsymbol{\gamma}}$ is distributed on $p$-dimensional space. The singular Gaussian distribution has a non-singular distribution form in a lower-dimensional space. In fact, if we keep the positive eigenvalues and the corresponding eigenvectors, then we obtain the positive definite matrix $\boldsymbol{U}_{-(p+1)}\boldsymbol{D}_{-(p+1)}\boldsymbol{U}^{-1}_{-(p+1)}$, where $\boldsymbol{D}_{-(p+1)}$ represents the exclusion of the last eigenvalue 0. $\boldsymbol{U}_{-(p+1)}$ represents the exclusion of the $(p+1)$-th row and the $(p+1)$-th column of $\boldsymbol{U}$. Then we obtain that $\boldsymbol{U}_{-(p+1)}\boldsymbol{D}_{-(p+1)}\boldsymbol{U}^{-1}_{-(p+1)} = \boldsymbol{I}_{p_\gamma}$. Therefore, we showed the consistency of our definition of the distribution of $\boldsymbol{T}_{\boldsymbol{\gamma}}\boldsymbol{\beta}_{\boldsymbol{\gamma}}$.

We now comment on the fact that the $g$-prior, which has a similar form to our proposed $z$-prior, does not satisfy the zero constraint, motivating the development of our novel prior. Specifically, a regular $g$-prior with $\tau^2\sigma^2(\boldsymbol{X}^T_\gamma\boldsymbol{X}_\gamma)^{-1}$ structure cannot guarantee the zero-constrained property. We ran a simple simulation to illustrate this point, adopting the same set-up as described in the main manuscript when $n = 50$ and $p = 30$. We then calculated the sum of all the elements in the squared matrix of log transformed design matrix $(\boldsymbol{X}^T_\gamma\boldsymbol{X}_\gamma)^{-1}$. We repeated this calculation 100 times. We obtained the mean as 0.18 and the standard deviation as 0.07, implying that $\sum_\gamma \boldsymbol{\beta}_{\boldsymbol{\gamma}}$ violates the zero-constrained property.

## S3. PROPERTIES OF THE MATRIX $(\boldsymbol{T}_\gamma'\boldsymbol{T}_\gamma)^{-1}$

In this section, we discuss the properties of the matrix $(\boldsymbol{T}_\gamma'\boldsymbol{T}_\gamma)^{-1}$, which appears in the variance of the $z$-prior. We assume the $z$-prior of $\boldsymbol{\beta}_\gamma$ is given as

$$\boldsymbol{\beta}_\gamma|\mathcal{M}_\gamma,\sigma^2,\tau^2 \sim \mathcal{N}\left(\mathbf{0},\sigma^2\tau^2(\boldsymbol{T}_\gamma'\boldsymbol{T}_\gamma)^{-1}\right),$$

where the proposed generalized transformation $\boldsymbol{T}_\gamma$ is defined as

$$\boldsymbol{T}_\gamma = \begin{bmatrix} \boldsymbol{I}_{p_\gamma} \\ c * \mathbf{1}_{p_\gamma}' \end{bmatrix}_{(p_\gamma+1)\times p_\gamma}.$$

Given the generalized transformation matrix $\boldsymbol{T}_\gamma$, the inverse of the matrix $\boldsymbol{T}_\gamma'\boldsymbol{T}_\gamma$ has the explicit form $(\boldsymbol{T}_\gamma'\boldsymbol{T}_\gamma)^{-1} = \boldsymbol{I}_{p_\gamma} - \frac{c^2}{1+c^2p_\gamma}\mathbf{1}_{p_\gamma}\mathbf{1}_{p_\gamma}'$. Thus, the sum of the linear coefficients $\sum_{i\in\gamma}\beta_i$ follows a normal distribution with mean 0 and variance $\frac{p_\gamma}{1+c^2p_\gamma}\sigma^2\tau^2$. When $c$ becomes large, the variance approaches 0, which implies that more shrinkage is imposed on $\sum_{i\in\gamma}\beta_i$. We can even let $c$ be $+\infty$; then the term $(\boldsymbol{T}_\gamma'\boldsymbol{T}_\gamma)^{-1}$ converges to $\boldsymbol{I}_{p_\gamma} - \frac{1}{p_\gamma}\mathbf{1}_{p_\gamma}\mathbf{1}_{p_\gamma}'$ and $\text{var}(\sum_{i\in\gamma}\beta_i) = 0$. The matrix $\boldsymbol{I}_{p_\gamma} - \frac{1}{p_\gamma}\mathbf{1}_{p_\gamma}\mathbf{1}_{p_\gamma}'$ is an idempotent matrix, and is singular with one zero eigenvalue and all the other eigenvalues equal to one. In this singular case of the multivariate normal distribution, one of the values is constrained by the others. For more details of the singular normal distribution, please refer to [1] and [2].

We now discuss the form of $(\boldsymbol{T}_\gamma'\boldsymbol{T}_\gamma)^{-1}$ when $\boldsymbol{T}$ represents a contrast transformation. The sum of the linear coefficients does not need to satisfy the zero-constrained property. For the additive log-ratio (ALR) transformation, the transformation matrix $\boldsymbol{T}_\gamma$ is given as

$$\boldsymbol{T}_\gamma = \begin{bmatrix} \boldsymbol{I}_{(p_\gamma-1)\times(p_\gamma-1)} \\ -\mathbf{1}_{p_\gamma-1}' \end{bmatrix}_{p_\gamma\times(p_\gamma-1)}.$$

The inverse of matrix $\boldsymbol{T}_\gamma'\boldsymbol{T}_\gamma$ has the explicit form $(\boldsymbol{T}_\gamma'\boldsymbol{T}_\gamma)^{-1} = \boldsymbol{I}_{p_\gamma} - \frac{1}{p_\gamma}\mathbf{1}_{p_\gamma}\mathbf{1}_{p_\gamma}'$. The sum of the linear coefficients $\sum_{i\in\gamma}\beta_i$ follows a normal distribution with mean 0 and variance $\frac{p_\gamma-1}{p_\gamma}\sigma^2\tau^2$. For the centered log-ratio (CLR) transformation, the transformation matrix $\boldsymbol{T}_\gamma$ is given as

$$\boldsymbol{T}_\gamma = \begin{bmatrix} \boldsymbol{D}_{(p_\gamma-1)\times(p_\gamma-1)} \\ -\frac{1}{p_\gamma}\times\mathbf{1}_{p_\gamma-1}' \end{bmatrix}_{p_\gamma\times(p_\gamma-1)},$$

where $\boldsymbol{D}_{(p_\gamma-1)\times(p_\gamma-1)}$ is a $(p_\gamma-1)\times(p_\gamma-1)$ square matrix with diagonal elements $1-\frac{1}{p_\gamma}$ and off-diagonal elements $-\frac{1}{p_\gamma}$. The inverse of matrix $\boldsymbol{T}_\gamma'\boldsymbol{T}_\gamma$ is then equal to $\boldsymbol{I}_{p_\gamma} + \mathbf{1}_{p_\gamma}\mathbf{1}_{p_\gamma}'$. The sum of the linear coefficients $\sum_{i\in\gamma}\beta_i$ follows a normal distribution with mean 0 and variance $p_\gamma(p_\gamma-1)\sigma^2\tau^2$.

We now include a brief comment on how these transformations respect the geometry of the compositional data setting. Euclidean space, the standard in classical geometry, is an inner product space on the real numbers, and so allows common operations such as addition and multiplication, along with notions such as distance and orthogonality that rely on the standard inner product. However, Euclidean space is not a proper geometry for compositional data, which must satisfy a fixed sum constraint. In this setting, Aitchison geometry, which defines a space on the simplex, is appropriate [3]. The contrast transformations are defined under Aitchison geometry and satisfy linearity in that space. Analogously, in the probability space, we can define the geometry of the zero-constrained Gaussian random variables. The $z$-prior is the probability measure defined under this geometry.

## S4. DETAILS OF POSTERIOR INFERENCE

***Derivations***: Assume that the likelihood function of the observation $\boldsymbol{Y}$ given model $\mathcal{M}_\gamma$ follows a multivariate normal distribution

$$p(\boldsymbol{Y}|\mathcal{M}_{\boldsymbol{\gamma}}, \boldsymbol{\beta}_{\boldsymbol{\gamma}}, \sigma^2) = (2\pi)^{-\frac{n}{2}}(\sigma^2)^{-\frac{n}{2}} \exp^{-\frac{(\boldsymbol{Y}-\boldsymbol{X}_{\boldsymbol{\gamma}}\boldsymbol{\beta}_{\boldsymbol{\gamma}})^T(\boldsymbol{Y}-\boldsymbol{X}_{\boldsymbol{\gamma}}\boldsymbol{\beta}_{\boldsymbol{\gamma}})}{2\sigma^2}} . \tag{S.1}$$

The prior density of $\boldsymbol{\beta}_\gamma$ is given as

$$\pi(\boldsymbol{\beta}_\gamma|\mathcal{M}_\gamma, \sigma^2, \tau^2) = (2\pi)^{-\frac{p_\gamma}{2}}(\sigma^2\tau^2)^{-\frac{p_\gamma}{2}} \left|(\boldsymbol{T}'_\gamma\boldsymbol{T}_\gamma)\right|^{1/2} \exp^{-\frac{\boldsymbol{\beta}_\gamma^T(\boldsymbol{T}'_\gamma\boldsymbol{T}_\gamma)\boldsymbol{\beta}_\gamma}{2\sigma^2\tau^2}} . \tag{S.2}$$

The prior density of $\sigma^2$ is given as

$$\pi(\sigma^2|\nu, \omega) = \frac{(\frac{\nu\omega}{2})^{\nu/2}}{\Gamma(\nu/2)}(\sigma^2)^{-\nu/2-1} \exp^{-\frac{(\nu\omega)/2}{\sigma^2}} . \tag{S.3}$$

Then the joint conditional distribution of $\boldsymbol{Y}$ and $\boldsymbol{\beta}_\gamma$ is

$$\begin{aligned}
&p(\boldsymbol{Y}, \boldsymbol{\beta}_{\boldsymbol{\gamma}}|\mathcal{M}_{\boldsymbol{\gamma}}, \sigma^2, \tau^2)\\
=&p(\boldsymbol{Y}|\boldsymbol{\beta}_{\boldsymbol{\gamma}}, \sigma^2)\pi(\boldsymbol{\beta}_{\boldsymbol{\gamma}}|\sigma^2, \tau^2)\\
=&(2\pi)^{-n/2}(\sigma^2)^{-n/2}(2\pi)^{-p_\gamma/2}(\sigma^2)^{-p_\gamma/2}(\tau^2)^{-p_\gamma/2}\left|(\boldsymbol{T}'_\gamma\boldsymbol{T}_\gamma)\right|^{1/2}\\
&\times \exp\left\{-\frac{1}{2\sigma^2}\left[\boldsymbol{\beta}_\gamma^T\left(\boldsymbol{X}_\gamma^T\boldsymbol{X}_\gamma + \frac{1}{\tau^2}(\boldsymbol{T}'_\gamma\boldsymbol{T}_\gamma)\right)\boldsymbol{\beta}_\gamma - \boldsymbol{\beta}_\gamma^T\boldsymbol{X}_\gamma^T\boldsymbol{Y} - \boldsymbol{Y}^T\boldsymbol{X}_\gamma\boldsymbol{\beta}_\gamma + \boldsymbol{Y}^T\boldsymbol{Y}\right]\right\}\\
=&(2\pi)^{-n/2}(\sigma^2)^{-n/2}(2\pi)^{-p_\gamma/2}(\sigma^2)^{-p_\gamma/2}(\tau^2)^{-p_\gamma/2}\left|(\boldsymbol{T}'_\gamma\boldsymbol{T}_\gamma)\right|^{1/2}\\
&\times \exp\left\{-\frac{1}{2\sigma^2}\left[\left(\boldsymbol{\beta}_\gamma - \boldsymbol{A}_\gamma^{-1}\boldsymbol{X}_\gamma^T\boldsymbol{Y}\right)^T\boldsymbol{A}_\gamma\left(\boldsymbol{\beta}_\gamma - \boldsymbol{A}_\gamma^{-1}\boldsymbol{X}_\gamma^T\boldsymbol{Y}\right) + \boldsymbol{Y}^T\boldsymbol{Y} - \boldsymbol{Y}^T\boldsymbol{X}_\gamma\boldsymbol{A}_\gamma^{-1}\boldsymbol{X}_\gamma^T\boldsymbol{Y}\right]\right\}\\
=&(2\pi)^{-n/2}(\sigma^2)^{-n/2}(\tau^2)^{-p_\gamma/2}\left|\boldsymbol{A}_\gamma\right|^{-1/2}\left|(\boldsymbol{T}'_\gamma\boldsymbol{T}_\gamma)\right|^{1/2}\exp\left\{-\frac{1}{2\sigma^2}\left[\boldsymbol{Y}^T\boldsymbol{Y} - \boldsymbol{Y}^T\boldsymbol{X}_\gamma\boldsymbol{A}_\gamma^{-1}\boldsymbol{X}_\gamma^T\boldsymbol{Y}\right]\right\}\\
&\times (2\pi)^{-p_\gamma/2}(\sigma^2)^{-p_\gamma/2}\left|\boldsymbol{A}_\gamma\right|^{1/2}\exp\left\{-\frac{1}{2\sigma^2}\left(\boldsymbol{\beta}_\gamma - \hat{\boldsymbol{\beta}}_\gamma\right)^T\boldsymbol{A}_\gamma\left(\boldsymbol{\beta}_\gamma - \hat{\boldsymbol{\beta}}_\gamma\right)\right\},
\end{aligned}$$

where $\hat{\boldsymbol{\beta}}_\gamma = \boldsymbol{A}_\gamma^{-1}\boldsymbol{X}_\gamma^T\boldsymbol{Y}$ and $\boldsymbol{A}_\gamma = \boldsymbol{X}_\gamma^T\boldsymbol{X}_\gamma + \frac{1}{\tau^2}(\boldsymbol{T}'_\gamma\boldsymbol{T}_\gamma)$. As we can observe, $\boldsymbol{\beta}_\gamma$ is inside a normal density function, so we can integrate it out. We have

$$\begin{aligned}
&p(\boldsymbol{Y}|\mathcal{M}_\gamma, \sigma^2, \tau^2)\\
=&(2\pi)^{-n/2}(\sigma^2)^{-n/2}(\tau^2)^{-p_\gamma/2}\left|\boldsymbol{A}_\gamma\right|^{-1/2}\left|(\boldsymbol{T}'_\gamma\boldsymbol{T}_\gamma)\right|^{1/2}\exp\left\{-\frac{1}{2\sigma^2}\left[\boldsymbol{Y}^T\boldsymbol{Y} - \boldsymbol{Y}^T\boldsymbol{X}_\gamma\boldsymbol{A}_\gamma^{-1}\boldsymbol{X}_\gamma^T\boldsymbol{Y}\right]\right\}.
\end{aligned}$$

Then the joint conditional distribution of $\boldsymbol{Y}$ and $\sigma^2$ is

$$\begin{aligned}
&p(\boldsymbol{Y}, \sigma^2|\mathcal{M}_\gamma, \tau^2, \nu, \omega)\\
=&p(\boldsymbol{Y}|\mathcal{M}_\gamma, \sigma^2, \tau^2)\pi(\sigma^2|\nu, \omega)\\
=&(2\pi)^{-n/2}(\tau^2)^{-p_\gamma/2}\left|\boldsymbol{A}_\gamma\right|^{-1/2}\left|(\boldsymbol{T}'_\gamma\boldsymbol{T}_\gamma)\right|^{1/2}\\
&\times \frac{(\frac{\nu\omega}{2})^{\nu/2}}{\Gamma(\nu/2)}(\sigma^2)^{-\frac{n+\nu}{2}-1}\exp\left\{-\frac{1}{\sigma^2}\frac{1}{2}\left[\boldsymbol{Y}^T\boldsymbol{Y} - \boldsymbol{Y}^T\boldsymbol{X}_\gamma\boldsymbol{A}_\gamma^{-1}\boldsymbol{X}_\gamma^T\boldsymbol{Y} + \nu\omega\right]\right\}.
\end{aligned}$$

As we can observe, $\sigma^2$ is inside an Inverse-Gamma function, so we can integrate it out.

*Marginal likelihood.* The marginal posterior of $\boldsymbol{Y}$ is given as

$$\begin{aligned}
&p(\boldsymbol{Y}|\mathcal{M}_\gamma, \tau^2, \nu, \omega)\\
=&(2\pi)^{-n/2}(\tau^2)^{-p_\gamma/2}\left|\boldsymbol{A}_\gamma\right|^{-1/2}\left|(\boldsymbol{T}'_\gamma\boldsymbol{T}_\gamma)\right|^{1/2}(\frac{\nu\omega}{2})^{\nu/2}\frac{\Gamma(\frac{n+\nu}{2})}{\Gamma(\frac{\nu}{2})}\left\{\frac{1}{2}\left[\boldsymbol{Y}^T\boldsymbol{Y} - \boldsymbol{Y}^T\boldsymbol{X}_\gamma\boldsymbol{A}_\gamma^{-1}\boldsymbol{X}_\gamma^T\boldsymbol{Y} + \nu\omega\right]\right\}^{-\frac{n+\nu}{2}}.
\end{aligned}$$

When $\nu = \omega = 0$ in the inverse-gamma prior of equation (7), the inverse-gamma reduces to a non-informative prior. Given this choice of hyperparameters, we obtain the marginal likelihood of $\boldsymbol{Y}$ given model $\mathcal{M}_{\boldsymbol{\gamma}}$ and the fixed hyperparameter $\tau^2$ as

$$p(\boldsymbol{Y}|\mathcal{M}_{\boldsymbol{\gamma}}, \tau^2) = (\pi)^{-n/2}\Gamma(\frac{n}{2})(\tau^2)^{-p_{\boldsymbol{\gamma}}/2} \left|\boldsymbol{A}_{\boldsymbol{\gamma}}\right|^{-1/2} \left|(\boldsymbol{T}_{\boldsymbol{\gamma}}'\boldsymbol{T}_{\boldsymbol{\gamma}})\right|^{1/2} \left[\boldsymbol{Y}^T\boldsymbol{Y} - \boldsymbol{Y}^T\boldsymbol{X}_{\boldsymbol{\gamma}}\boldsymbol{A}_{\boldsymbol{\gamma}}^{-1}\boldsymbol{X}_{\boldsymbol{\gamma}}^T\boldsymbol{Y}\right]^{-\frac{n}{2}}.$$

From the marginal likelihood of model of $\boldsymbol{Y}$ given model $\mathcal{M}_{\boldsymbol{\gamma}}$, the Bayes factor is

$$F(\boldsymbol{\gamma}'|\boldsymbol{\gamma}) = (\tau^2)^{-(\frac{p_{\boldsymbol{\gamma}'}-p_{\boldsymbol{\gamma}}}{2})} \frac{|\boldsymbol{A}_{\boldsymbol{\gamma}}|^{1/2}\left|(\boldsymbol{T}_{\boldsymbol{\gamma}'}'\boldsymbol{T}_{\boldsymbol{\gamma}'})\right|^{1/2}}{|\boldsymbol{A}_{\boldsymbol{\gamma}'}|^{1/2}\left|(\boldsymbol{T}_{\boldsymbol{\gamma}}'\boldsymbol{T}_{\boldsymbol{\gamma}})\right|^{1/2}} \left(\frac{\boldsymbol{Y}^T\boldsymbol{Y} - \boldsymbol{Y}^T\boldsymbol{X}_{\boldsymbol{\gamma}}\boldsymbol{A}_{\boldsymbol{\gamma}}^{-1}\boldsymbol{X}_{\boldsymbol{\gamma}}^T\boldsymbol{Y} + \nu\omega}{\boldsymbol{Y}^T\boldsymbol{Y} - \boldsymbol{Y}^T\boldsymbol{X}_{\boldsymbol{\gamma}'}\boldsymbol{A}_{\boldsymbol{\gamma}'}^{-1}\boldsymbol{X}_{\boldsymbol{\gamma}'}^T\boldsymbol{Y} + \nu\omega}\right)^{\frac{n+\nu}{2}}, \tag{S.4}$$

where $\boldsymbol{A}_{\boldsymbol{\gamma}'} = \boldsymbol{X}_{\boldsymbol{\gamma}'}^T\boldsymbol{X}_{\boldsymbol{\gamma}'} + \frac{1}{\tau^2}(\boldsymbol{T}_{\boldsymbol{\gamma}'}'\boldsymbol{T}_{\boldsymbol{\gamma}'})$. $\boldsymbol{\gamma}'$ and $\boldsymbol{\gamma}$ only differ by one index position, therefore $p_{\boldsymbol{\gamma}'} - p_{\boldsymbol{\gamma}}$ will be 1 or $-1$.

Next, we calculate the posterior of $\boldsymbol{\beta}_{\boldsymbol{\gamma}}$. The joint conditional distribution of $\boldsymbol{Y}$, $\boldsymbol{\beta}_\gamma$, $\sigma^2$ is

$$\begin{aligned}
&p(\boldsymbol{Y}, \boldsymbol{\beta}_{\boldsymbol{\gamma}}, \sigma^2|\mathcal{M}_{\boldsymbol{\gamma}}, \tau^2, \nu, \omega)\\
=&p(\boldsymbol{Y}, \boldsymbol{\beta}_{\boldsymbol{\gamma}}|\mathcal{M}_{\boldsymbol{\gamma}}, \sigma^2, \tau^2)\pi(\sigma^2|\nu, \omega)\\
=&(2\pi)^{-n/2}(2\pi)^{-p_{\boldsymbol{\gamma}}/2}(\tau^2)^{-p_{\boldsymbol{\gamma}}/2}\frac{(\frac{\nu\omega}{2})^{\nu/2}}{\Gamma(\nu/2)}(\sigma^2)^{-\frac{n+p_{\boldsymbol{\gamma}}+\nu}{2}-1}\left|(\boldsymbol{T}_{\boldsymbol{\gamma}}'\boldsymbol{T}_{\boldsymbol{\gamma}})\right|^{1/2}\\
&\exp\left\{-\frac{1}{2\sigma^2}\left[\left(\boldsymbol{\beta}_{\boldsymbol{\gamma}} - \hat{\boldsymbol{\beta}}_{\boldsymbol{\gamma}}\right)^T\boldsymbol{A}_{\boldsymbol{\gamma}}\left(\boldsymbol{\beta}_{\boldsymbol{\gamma}} - \hat{\boldsymbol{\beta}}_{\boldsymbol{\gamma}}\right) + \boldsymbol{Y}^T\boldsymbol{Y} - \boldsymbol{Y}^T\boldsymbol{X}_{\boldsymbol{\gamma}}\boldsymbol{A}_{\boldsymbol{\gamma}}^{-1}\boldsymbol{X}_{\boldsymbol{\gamma}}^T\boldsymbol{Y} + \nu\omega\right]\right\}.
\end{aligned}$$

As we can observe, $\sigma^2$ is inside an Inverse-Gamma density function, so we can integrate it out. We have

$$\begin{aligned}
&p(\boldsymbol{Y}, \boldsymbol{\beta}_{\boldsymbol{\gamma}}|\mathcal{M}_{\boldsymbol{\gamma}}, \tau^2, \nu, \omega)\\
=&(2\pi)^{-n/2}(2\pi)^{-p_{\boldsymbol{\gamma}}/2}(\tau^2)^{-p_{\boldsymbol{\gamma}}/2}\frac{(\frac{\nu\omega}{2})^{\nu/2}}{\Gamma(\nu/2)}\Gamma(\frac{n+p_{\boldsymbol{\gamma}}+\nu}{2})\left|(\boldsymbol{T}_{\boldsymbol{\gamma}}'\boldsymbol{T}_{\boldsymbol{\gamma}})\right|^{1/2}\\
&\left\{\frac{1}{2}\left[\left(\boldsymbol{\beta}_{\boldsymbol{\gamma}} - \hat{\boldsymbol{\beta}}_{\boldsymbol{\gamma}}\right)^T\boldsymbol{A}_{\boldsymbol{\gamma}}\left(\boldsymbol{\beta}_{\boldsymbol{\gamma}} - \hat{\boldsymbol{\beta}}_{\boldsymbol{\gamma}}\right) + \boldsymbol{Y}^T\boldsymbol{Y} - \boldsymbol{Y}^T\boldsymbol{X}_{\boldsymbol{\gamma}}\boldsymbol{A}_{\boldsymbol{\gamma}}^{-1}\boldsymbol{X}_{\boldsymbol{\gamma}}^T\boldsymbol{Y} + \nu\omega\right]\right\}^{-\frac{n+p_{\boldsymbol{\gamma}}+\nu}{2}}\\
=&\pi^{-n/2}(\tau^2)^{-p_{\boldsymbol{\gamma}}/2}(\nu\omega)^{\nu/2}\frac{\Gamma(\frac{n+\nu}{2})}{\Gamma(\nu/2)}|\boldsymbol{A}_{\boldsymbol{\gamma}}|^{-1/2}\left|(\boldsymbol{T}_{\boldsymbol{\gamma}}'\boldsymbol{T}_{\boldsymbol{\gamma}})\right|^{1/2}(C_{\boldsymbol{\gamma}} + \nu\omega)^{-\frac{n+\nu}{2}}\\
&\frac{\Gamma(\frac{n+\nu+p_{\boldsymbol{\gamma}}}{2})|\frac{C_{\boldsymbol{\gamma}}+\nu\omega}{n+\nu}\boldsymbol{A}_{\boldsymbol{\gamma}}^{-1}|^{-1/2}}{\Gamma(\frac{n+\nu}{2})(n+\nu)^{p_{\boldsymbol{\gamma}}/2}\pi^{p_{\boldsymbol{\gamma}}/2}}\left[1 + \frac{1}{n+\nu}\left(\boldsymbol{\beta}_{\boldsymbol{\gamma}} - \hat{\boldsymbol{\beta}}_{\boldsymbol{\gamma}}\right)^T\left[(n+\nu)(C_{\boldsymbol{\gamma}} + \nu\omega)^{-1}\boldsymbol{A}_{\boldsymbol{\gamma}}\right]\left(\boldsymbol{\beta}_{\boldsymbol{\gamma}} - \hat{\boldsymbol{\beta}}_{\boldsymbol{\gamma}}\right)\right]^{-\frac{n+\nu+p_{\boldsymbol{\gamma}}}{2}},
\end{aligned}$$

where $C_{\boldsymbol{\gamma}} = \boldsymbol{Y}^T\boldsymbol{Y} - \boldsymbol{Y}^T\boldsymbol{X}_{\boldsymbol{\gamma}}\boldsymbol{A}_{\boldsymbol{\gamma}}^{-1}\boldsymbol{X}_{\boldsymbol{\gamma}}^T\boldsymbol{Y}$. Obviously, $\boldsymbol{\beta}_{\boldsymbol{\gamma}}$ follows a multivariate $t$-distribution, given as

$$\begin{aligned}
&p(\boldsymbol{\beta}_{\boldsymbol{\gamma}}|\boldsymbol{Y}, \mathcal{M}_{\boldsymbol{\gamma}}, \tau^2, \nu, \omega)\\
=&p(\boldsymbol{Y}, \boldsymbol{\beta}_{\boldsymbol{\gamma}}|\mathcal{M}_{\boldsymbol{\gamma}}, \tau^2, \nu, \omega)/p(\boldsymbol{Y}|\mathcal{M}_{\boldsymbol{\gamma}}, \tau^2, \nu, \omega)\\
=&\frac{\Gamma(\frac{n+\nu+p_{\boldsymbol{\gamma}}}{2})|\frac{C_{\boldsymbol{\gamma}}+\nu\omega}{n+\nu}\boldsymbol{A}_{\boldsymbol{\gamma}}^{-1}|^{-1/2}}{\Gamma(\frac{n+\nu}{2})(n+\nu)^{p_{\boldsymbol{\gamma}}/2}\pi^{p_{\boldsymbol{\gamma}}/2}}\left[1 + \frac{1}{n+\nu}\left(\boldsymbol{\beta}_{\boldsymbol{\gamma}} - \hat{\boldsymbol{\beta}}_{\boldsymbol{\gamma}}\right)^T\left[(n+\nu)(C_{\boldsymbol{\gamma}} + \nu\omega)^{-1}\boldsymbol{A}_{\boldsymbol{\gamma}}\right]\left(\boldsymbol{\beta}_{\boldsymbol{\gamma}} - \hat{\boldsymbol{\beta}}_{\boldsymbol{\gamma}}\right)\right]^{-\frac{n+\nu+p_{\boldsymbol{\gamma}}}{2}},
\end{aligned}$$

with mean $\hat{\boldsymbol{\beta}}_{\boldsymbol{\gamma}} = \boldsymbol{A}_{\boldsymbol{\gamma}}^{-1}\boldsymbol{X}_{\boldsymbol{\gamma}}^T\boldsymbol{Y}$ and covariance $\frac{1}{n+\nu-2}(C_{\boldsymbol{\gamma}} + \nu\omega)\boldsymbol{A}_{\boldsymbol{\gamma}}^{-1}$. The mean squared error of prediction is defined as $\text{MSE} = \frac{1}{n-p_{\boldsymbol{\gamma}}}(\boldsymbol{Y} - \boldsymbol{X}_{\boldsymbol{\gamma}}\boldsymbol{A}_{\boldsymbol{\gamma}}^{-1}\boldsymbol{X}_{\boldsymbol{\gamma}}^T\boldsymbol{Y})^T(\boldsymbol{Y} - \boldsymbol{X}_{\boldsymbol{\gamma}}\boldsymbol{A}_{\boldsymbol{\gamma}}^{-1}\boldsymbol{X}_{\boldsymbol{\gamma}}^T\boldsymbol{Y})$, when $p_{\boldsymbol{\gamma}} < n$.

Lastly, the posterior of $\sigma^2$ follows an Inverse-Gamma distribution given as

$$\begin{aligned}
p(\sigma^2|\boldsymbol{Y}, \mathcal{M}_{\boldsymbol{\gamma}}, \tau^2, \nu, \omega) &= p(\boldsymbol{Y}, \sigma^2|\mathcal{M}_{\boldsymbol{\gamma}}, \tau^2, \nu, \omega)/p(\boldsymbol{Y}|\mathcal{M}_{\boldsymbol{\gamma}}, \tau^2, \nu, \omega)\\
&= \frac{\left[\frac{1}{2}(C_{\boldsymbol{\gamma}} + \nu\omega)\right]^{\frac{n+\nu}{2}}}{\Gamma(\frac{n+\nu}{2})}(\sigma^2)^{-\frac{n+\nu}{2}-1}\exp^{\left(\frac{(C_{\boldsymbol{\gamma}}+\nu\omega)/2}{\sigma^2}\right)}.
\end{aligned}$$

with the shape parameter $\frac{n+\nu}{2}$ and the scale parameter $\frac{1}{2}(C_{\boldsymbol{\gamma}} + \nu\omega)$. The mean is given by $\frac{(C_{\boldsymbol{\gamma}}+\nu\omega)}{n+\nu}$.

***MCMC algorithm***: We now outline the construction of the Gibbs sampler on $\boldsymbol{\gamma}$, which searches over the space of models $\{0,1\}^p$. Let $\gamma_{(-i)} = \{\gamma_j : j \neq i\}$, and $I_{(-i)}$ be $\{\gamma_j = 1 : j \neq i\}$, the set of indices for the selected variables other than $i$. $\tau$ is fixed at 1. The posterior distribution of $\boldsymbol{\gamma}$ given the data can be decomposed by Bayes formula as

$$P(\gamma_i = 1|\gamma_{(-i)}, \boldsymbol{y}) = \frac{P(\gamma_i = 1|\gamma_{(-i)})}{P(\gamma_i = 1|\gamma_{(-i)}) + F(\boldsymbol{\gamma}'|\boldsymbol{\gamma})^{-1} \times P(\gamma_i = 0|\gamma_{(-i)})}\ , \tag{S.5}$$

where $F(\boldsymbol{\gamma}'|\boldsymbol{\gamma})$ is the Bayes factor for the indicator vectors $\boldsymbol{\gamma}'$ and $\boldsymbol{\gamma}$, and is defined as

$$F(\boldsymbol{\gamma}'|\boldsymbol{\gamma}) = \frac{|\boldsymbol{A}_{\boldsymbol{\gamma}}|^{1/2}\left|(\boldsymbol{T}'_{\boldsymbol{\gamma}'}\boldsymbol{T}_{\boldsymbol{\gamma}'})\right|^{1/2}}{|\boldsymbol{A}_{\boldsymbol{\gamma}'}|^{1/2}\left|(\boldsymbol{T}'_{\boldsymbol{\gamma}}\boldsymbol{T}_{\boldsymbol{\gamma}})\right|^{1/2}} \left(\frac{\boldsymbol{y}^T\boldsymbol{y} - \boldsymbol{y}^T\boldsymbol{X}_{\boldsymbol{\gamma}}\boldsymbol{A}_{\boldsymbol{\gamma}}^{-1}\boldsymbol{X}_{\boldsymbol{\gamma}}^T\boldsymbol{y}}{\boldsymbol{y}^T\boldsymbol{y} - \boldsymbol{y}^T\boldsymbol{X}_{\boldsymbol{\gamma}'}\boldsymbol{A}_{\boldsymbol{\gamma}'}^{-1}\boldsymbol{X}_{\boldsymbol{\gamma}'}^T\boldsymbol{y}}\right)^{\frac{n}{2}}. \tag{S.6}$$

From equation (8), the conditional distribution of $\gamma_i$ under the prior is given by

$$P(\gamma_i|\gamma_{(-i)}) = \frac{e^{\gamma_i a_i + \sum_{j\in I_{(-i)}} q_{ij}\gamma_i\gamma_j}}{1 + e^{a_i + \sum_{j\in I_{(-i)}} q_{ij}\gamma_j}}.$$

In each iteration, we select an index $i$ at random, and then sample a Bernoulli random variable with probability $P(\gamma_i = 1|\gamma_{(-i)}, \boldsymbol{y})$ following equation (S.5). Since we update only one index at a time, $p_{\boldsymbol{\gamma}'} - p_{\boldsymbol{\gamma}}$ will be 1 or $-1$, and $\boldsymbol{\gamma}'$ and $\boldsymbol{\gamma}$ differ only in the $i$th position. If the proposed value equals the current $\gamma_i$, the model is unchanged; otherwise, we update $\boldsymbol{\gamma}$ accordingly. To accelerate the computationally intensive step of evaluating $F(i|\gamma_{(-i)})$, we adopt the same procedure to calculate the matrix inverse and determinant as in [4].

The computational speed of the proposed method is quite fast, especially when the true model space is sparse. For all of the simulations provided in Section 4, each MCMC run has 20,000 iterations with the first 15,000 as burn-in. On average for data with the dependent covariate structure, it takes 80 seconds to run 20,000 iterations with an average posterior model size of 24 on an Intel Core(TM) i5-6500 with 3.2GHz CPU.

## S5. JUSTIFICATION OF INVERSE-GAMMA PRIOR ON $\sigma^2$

In this section, we justify why posterior inference for our model is robust to small values of $\nu$ and $\omega$ using both theory and sensitivity analysis. For our theoretical discussion, let us recall the the hierarchical model in Gelman (2006) [5].

**Model 1:**

$$\begin{aligned} y_{ij} &\sim \mathcal{N}(\mu + \xi\eta_j, \sigma_y^2) && \text{Hierarchy I} \\ \eta_j &\sim \mathcal{N}(0, \sigma_\eta^2) && \text{Hierarchy II} \\ \sigma_\eta^2 &\sim \text{InverseGamma}(\alpha, \beta) && \text{Hierarchy III} \end{aligned}$$

Then we know the likelihood for $\xi$ has the form of a normal distribution $(y_{ij} - \mu)/\eta_j \sim \mathcal{N}(\xi, \sigma^2/\eta_j)$. Next we examine the impact of small values of $\alpha$ and $\beta$ on the likelihood of $\xi$. If $\alpha$ and $\beta$ are close to zero (Hierarchy III), then $\sigma_\eta^2$ will have a high concentration around zero. Then $\eta_j$ will be close to zero (Hierarchy II). The likelihood for $\xi$ will be unstable, as $\eta_j$ is in the denominator. This is intuitively the main reason why small values of Inverse Gamma parameters are not recommended in this hierarchical model.

Although our model is hierarchical as well, the prior on $\sigma^2$ plays a different role in governing the other layers compared with the prior in Gelman's paper. We write our proposed model as follows:

**Model 2:**

$$\begin{aligned} p(\boldsymbol{Y}|\mathcal{M}_{\boldsymbol{\gamma}}, \boldsymbol{\beta}_{\boldsymbol{\gamma}}, \sigma^2) &= (2\pi)^{-\frac{n}{2}}(\sigma^2)^{-\frac{n}{2}} \exp^{-\frac{(\boldsymbol{Y}-\boldsymbol{X}_{\boldsymbol{\gamma}}\boldsymbol{\beta}_{\boldsymbol{\gamma}})^T(\boldsymbol{Y}-\boldsymbol{X}_{\boldsymbol{\gamma}}\boldsymbol{\beta}_{\boldsymbol{\gamma}})}{2\sigma^2}} \\ \pi(\boldsymbol{\beta}_\gamma|\mathcal{M}_{\boldsymbol{\gamma}}, \sigma^2, \tau^2) &= (2\pi)^{-\frac{p_\gamma}{2}}(\sigma^2\tau^2)^{-\frac{p_\gamma}{2}} \left|(\boldsymbol{T}'_{\boldsymbol{\gamma}}\boldsymbol{T}_{\boldsymbol{\gamma}})\right|^{1/2} \exp^{-\frac{\boldsymbol{\beta}_{\boldsymbol{\gamma}}^T(\boldsymbol{T}'_{\boldsymbol{\gamma}}\boldsymbol{T}_{\boldsymbol{\gamma}})\boldsymbol{\beta}_{\boldsymbol{\gamma}}}{2\sigma^2\tau^2}} \\ \pi(\sigma^2|\nu,\omega) &= \frac{(\frac{\nu\omega}{2})^{\nu/2}}{\Gamma(\nu/2)}(\sigma^2)^{-\nu/2-1} \exp^{-\frac{(\nu\omega)/2}{\sigma^2}} \end{aligned}$$

Here, the observed data can directly contribute information regarding inference on $\sigma^2$, because $\sigma^2$ is included in the likelihood function, while in Gelman's paper the likelihood function does not contain $\sigma^2_\eta$. Moreover, the hyperparameters $\alpha$ and $\beta$ are not analytically tractable in Model 1, but we can express the explicit form of the posterior of $\sigma^2$ in our model (see Section 3 above). Just like a regular Gaussian regression model with an Inverse-Gamma prior imposed on the variance term, the posterior of $\sigma^2$ in our model follows an Inverse-Gamma distribution:

$$\begin{aligned} p(\sigma^2|\boldsymbol{Y}, \mathcal{M}_{\boldsymbol{\gamma}}, \tau^2, \nu, \omega) &= p(\boldsymbol{Y}, \sigma^2|\mathcal{M}_{\boldsymbol{\gamma}}, \tau^2, \nu, \omega)/p(\boldsymbol{Y}|\mathcal{M}_{\boldsymbol{\gamma}}, \tau^2, \nu, \omega) \\ &= \frac{\left[\frac{1}{2}(C_{\boldsymbol{\gamma}} + \nu\omega)\right]^{\frac{n+\nu}{2}}}{\Gamma(\frac{n+\nu}{2})}(\sigma^2)^{-\frac{n+\nu}{2}-1} \exp^{\left(\frac{(C_{\boldsymbol{\gamma}}+\nu\omega)/2}{\sigma^2}\right)}, \end{aligned}$$

with shape parameter $\frac{n+\nu}{2}$ and scale parameter $\frac{1}{2}(C_{\boldsymbol{\gamma}} + \nu\omega)$. The mean is given by $\frac{(C_{\boldsymbol{\gamma}}+\nu\omega)}{n+\nu}$. Note that $C_{\boldsymbol{\gamma}} = \boldsymbol{Y}^T\boldsymbol{Y} - \boldsymbol{Y}^T\boldsymbol{X}_{\boldsymbol{\gamma}}\boldsymbol{A}_{\boldsymbol{\gamma}}^{-1}\boldsymbol{X}_{\boldsymbol{\gamma}}^T\boldsymbol{Y}$ and $\boldsymbol{A}_{\boldsymbol{\gamma}} = \boldsymbol{X}_{\boldsymbol{\gamma}}^T\boldsymbol{X}_{\boldsymbol{\gamma}} + \frac{1}{\tau^2}(\boldsymbol{T}'_{\boldsymbol{\gamma}}\boldsymbol{T}_{\boldsymbol{\gamma}})$. Small values of $\nu$ and $\omega$ do not have a strong impact on the posterior, as the shape parameter is dominated by $n/2$ and the scale parameter is dominated by $C_{\boldsymbol{\gamma}}/2$. Additionally, the posterior is a proper density even when $\nu = \omega = 0$.

$C_{\boldsymbol{\gamma}}$ is actually the sum of squared estimate of errors, which measures the fitness of the selected variables. $C_{\boldsymbol{\gamma}}$ is decided by the posterior of the selection index $\boldsymbol{\gamma}$. The conditional posterior distribution of $\boldsymbol{\gamma}$ given the data can be decomposed by Bayes formula as

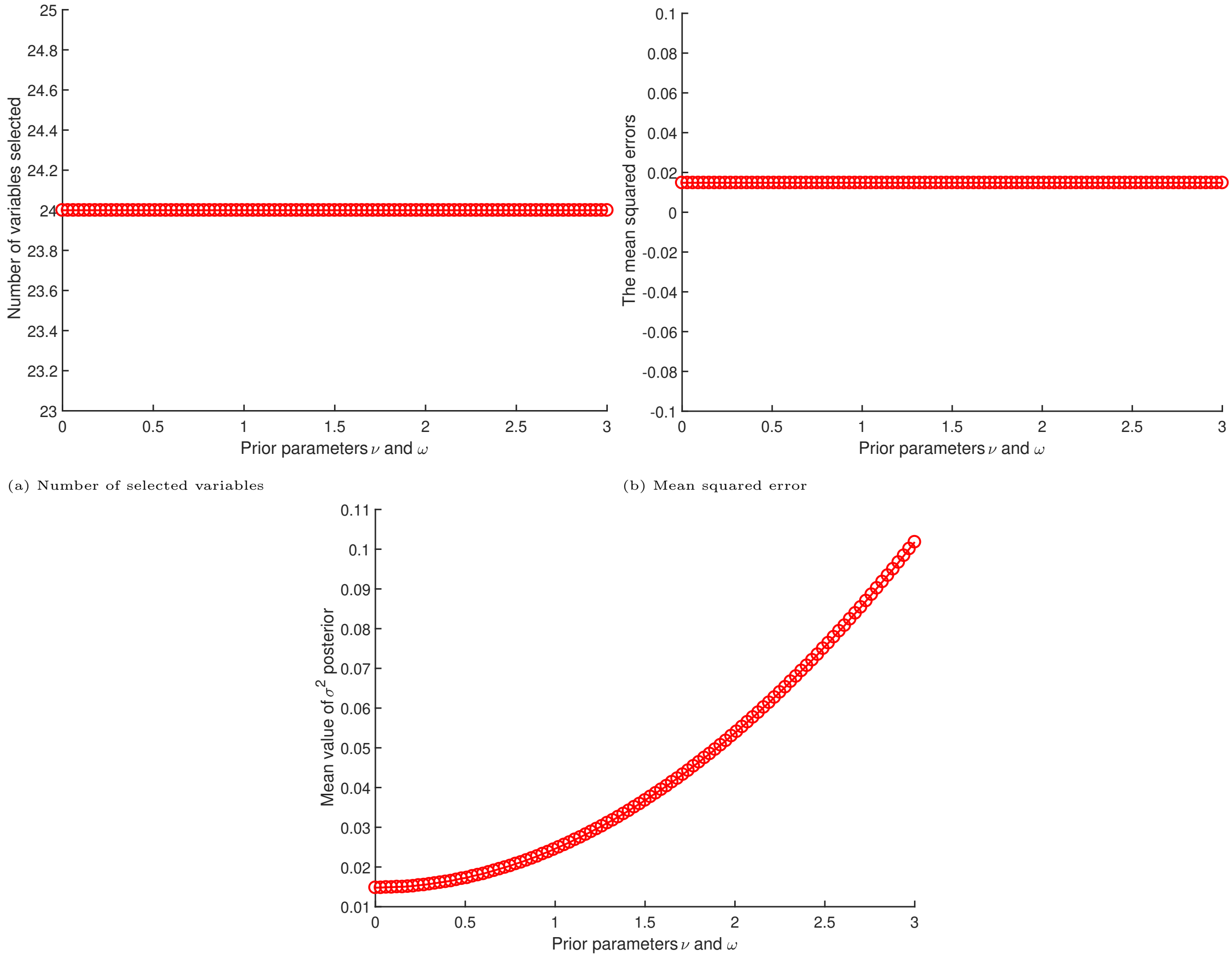
$$P(\gamma_i = 1|\gamma_{(-i)}, \boldsymbol{y}) = \frac{P(\gamma_i = 1|\gamma_{(-i)})}{P(\gamma_i = 1|\gamma_{(-i)}) + F(\boldsymbol{\gamma}'|\boldsymbol{\gamma})^{-1} \times P(\gamma_i = 0|\gamma_{(-i)})},$$

and the Bayes factor is

$$F(\boldsymbol{\gamma}'|\boldsymbol{\gamma}) = (\tau^2)^{-(\frac{p_{\boldsymbol{\gamma}'}-p_{\boldsymbol{\gamma}}}{2})} \frac{|\boldsymbol{A}_{\boldsymbol{\gamma}}|^{1/2}\left|(\boldsymbol{T}'_{\boldsymbol{\gamma}'}\boldsymbol{T}_{\boldsymbol{\gamma}'})\right|^{1/2}}{|\boldsymbol{A}_{\boldsymbol{\gamma}'}|^{1/2}\left|(\boldsymbol{T}'_{\boldsymbol{\gamma}}\boldsymbol{T}_{\boldsymbol{\gamma}})\right|^{1/2}} \left(\frac{\boldsymbol{Y}^T\boldsymbol{Y} - \boldsymbol{Y}^T\boldsymbol{X}_{\boldsymbol{\gamma}}\boldsymbol{A}_{\boldsymbol{\gamma}}^{-1}\boldsymbol{X}_{\boldsymbol{\gamma}}^T\boldsymbol{Y} + \nu\omega}{\boldsymbol{Y}^T\boldsymbol{Y} - \boldsymbol{Y}^T\boldsymbol{X}_{\boldsymbol{\gamma}'}\boldsymbol{A}_{\boldsymbol{\gamma}'}^{-1}\boldsymbol{X}_{\boldsymbol{\gamma}'}^T\boldsymbol{Y} + \nu\omega}\right)^{\frac{n+\nu}{2}}.$$

The Bayes factor depends on $\nu$ and $\omega$, but, in the same way, small values of $\nu$ and $\omega$ do not have a strong impact on the Bayes factor. Therefore, in summary, small values of $\nu$ and $\omega$ in the prior do not impact the robustness of the posterior of $\sigma^2$.

To complement the analytical discussion above, we performed a simulation-based sensitivity analysis to confirm the behavior of our model for changing parameter values. We use the same simulated data with dependent covariates as described in section 4.2 in the main manuscript, where 24 of the 1000 variables are set to be non-zero, and the true $\sigma^2$ is set as 0.0138. We let $\nu = \omega$, and vary their values from 0 to 3 with a step size of 0.03. Then we run the model and plot the results of variable selection and posterior estimates of $\sigma^2$. As shown in Figure S1, the model selects 24 variables and the MSE $\big($i.e., $\frac{C_{\boldsymbol{\gamma}}}{n}\big)$ is 0.0148, which means the selection results are correct and robust to the change of prior parameters. The posterior mean of $\sigma^2$ increases because $\nu$ and $\omega$ are getting bigger. But $\nu = \omega = 0$ give the best estimate of the true $\sigma^2$ value. Therefore, the posterior of $\sigma^2$ is robust to small values of $\nu$ and $\omega$.



(a) Number of selected variables

(b) Mean squared error

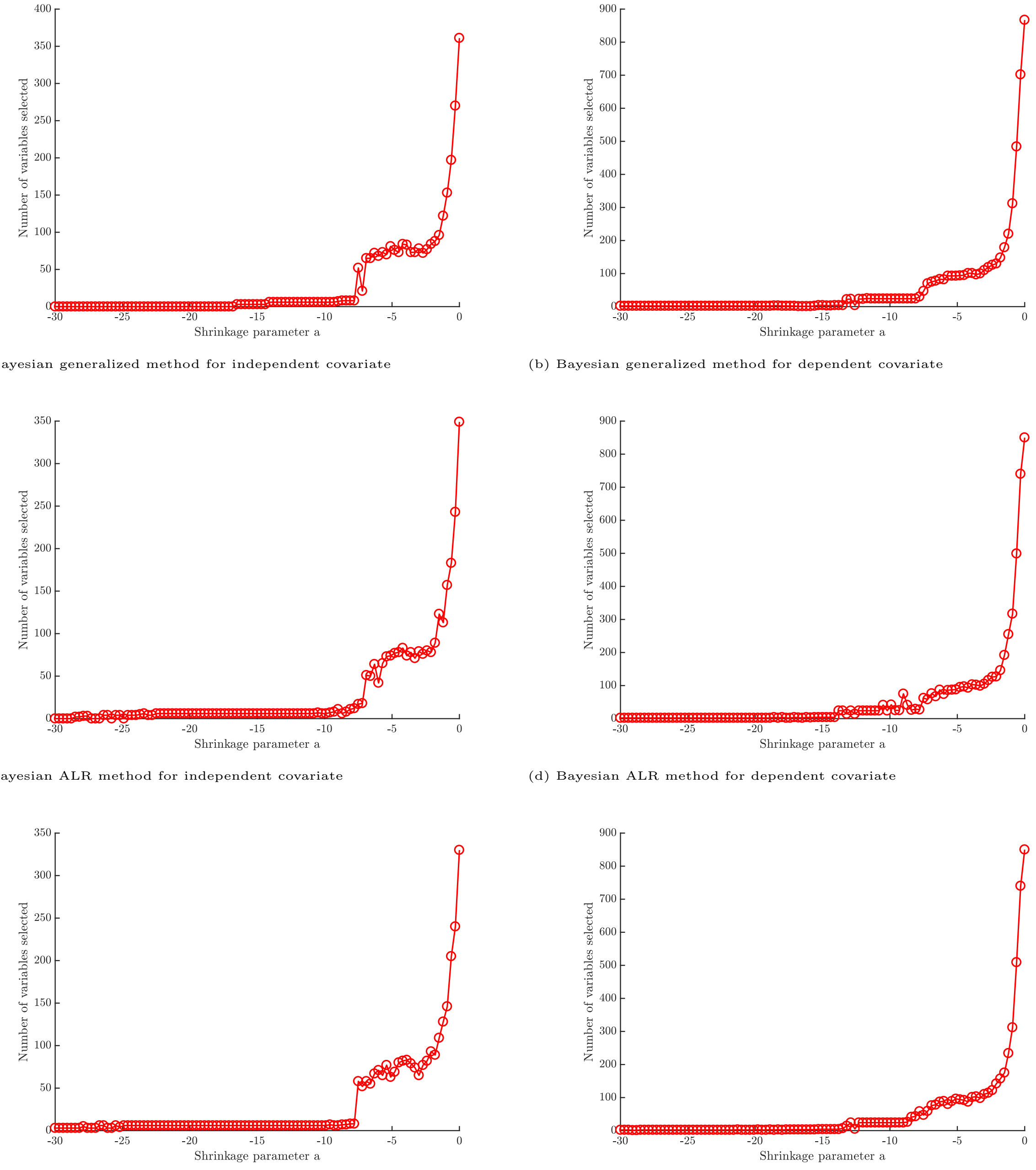
(c) Posterior mean of $\sigma^2$

FIG. S1: Sensitivity analysis of the Inverse-Gamma prior on $\sigma^2$. We use the same simulated data with dependent covariates as described in section 4.2 in the main manuscript, where 24 of the 1000 variables are set to be non-zero, and the true $\sigma^2$ is set as 0.0148.

## S6. SENSITIVITY TO SHRINKAGE PARAMETER

As microbial abundances vary across different OTUs and different experiments, we standardize the design matrix $\boldsymbol{X}$ in the regression model to ensure a consistent scale. We also scale each column of the transformation $\boldsymbol{T}_\gamma$ by the standard deviation of each column of $\boldsymbol{X}$. We use these standardized data to perform sensitivity analysis, posterior inference, and variable selection. Although this standardization is not required to apply the model, it makes the parameter choice easier to calibrate and compare across settings.

As the shrinkage parameter $a$ affects the model sparsity, we provide an illustration of the number of selected variables as a function of $a$ in Figure S2. For the independent covariate structure, when $a$ is between $-14$ and $-8$, the number of selected variables is nonzero and stable. Similarly, for the dependent covariate structure, when $a$ is between $-13$ and $-9$, the number of selected variables is nonzero and stable. Therefore, we set $a$ around -12 (in the middle of the stable range where around 24 variables are selected) to run the proposed method in all the simulation scenarios. This parameter choice reflects a preference for sparsity, which, in the absence of further information, allows for the selection of an interpretable and parsimonious model.



(a) Bayesian generalized method for independent covariate

(b) Bayesian generalized method for dependent covariate

(c) Bayesian ALR method for independent covariate

(d) Bayesian ALR method for dependent covariate

(e) Bayesian CLR method for independent covariate

(f) Bayesian CLR method for dependent covariate

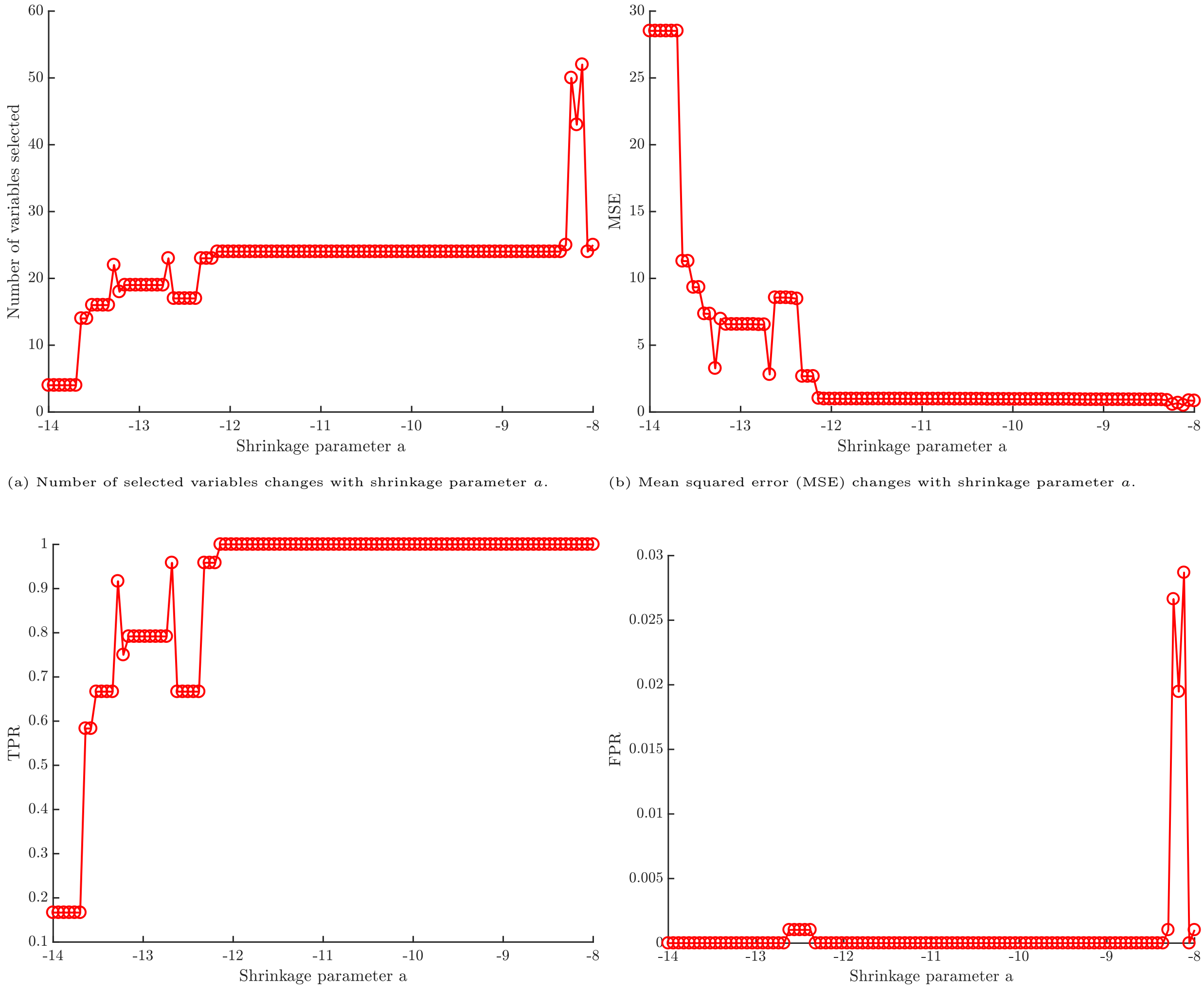
FIG. S2: Trace plot of sensitivity analysis of shrinkage parameter $a$ for simulations when $n = 100, p = 1000, \mathrm{SNR} = 1$.

We now take a closer look at the effect of varying $a$ within the stable range of $-14$ to $-8$, considering additional performance measures available in the context of simulated data including mean squared error, true positive rates, and false positive rates. The results from this analysis are provided in Figure S3 and Figure S4. Although we did not select $a$ based on TPR or FPR, to avoid optimistic bias in our results, this analysis shows that the posterior estimates are robust to variation in $a$ within a reasonable range. For example, in Figure S3, when $a = -10$ , 24 covariates are selected, the mean squared error is 0.9782, TPR is 1, and FPR is 0. In Figure S4, when $a = -10.5$, 24 covariates are selected, the mean squared error is 0.9782, TPR is 1, and FPR is 0.

In simulating the data with dependence (Section 4 of the main manuscript), we constructed a scheme with five sets among the $p = 1000$ variables designed to have varying levels of signal and dependence structures. To recap, the true variables are $\boldsymbol{j} = \{160 + 20l\}_{l=1}^{12}$ and $\{560 + 20l\}_{l=1}^{12}$, the false variables are $\boldsymbol{j} = \{44 + l\}_{l=1}^{16}$, $\{444 + l\}_{l=1}^{16}$ and $\{944+l\}_{l=1}^{16}$, and all other covariates are uncorrelated noise variables. For the results provided in the main manuscript, the structural prior parameter $\boldsymbol{Q}$ is set to have nonzero entries for relations among the true variables and, to avoid giving an advantage to the Bayesian methods, also for relations among the false variables. Here, we also consider a more challenging scenario, where we additionally specify the regions $\boldsymbol{j} = \{340 + 10l\}_{l=1}^{11}$ to be nonzero in $\boldsymbol{Q}$. This setup allows, for instance, variables 390 and 400 to be linked.

As shown in Figure S5, in the setting of $\boldsymbol{Q}$ used in Section 4 of the main manuscript, when we set $a = -10$, the model is able to select all of the 24 true covariates correctly, and the MCMC converges well. In the more challenging setting of $\boldsymbol{Q}$, when we set $a = -10.5$, the model is able to select all of the 24 true covariates correctly, and the MCMC converges well. Therefore, even though dependence between true variables and their false neighbours has been added to the specification of $\boldsymbol{Q}$, the results are reasonable for values of $a$ within the stable range identified above.

As shown in Figure S6, when $a$ is greater than $-10$, the number of selected variables starts to be greater than 0. MSE is at a reasonable amount when $a$ ia around $-9$. After integrative considerations, we set $a$ around -9 to run for all Bayesian ALR, CLR, and generalized methods in real data analysis.



(a) Number of selected variables changes with shrinkage parameter $a$.

(b) Mean squared error (MSE) changes with shrinkage parameter $a$.

(c) True positive rates change with shrinkage parameter $a$.

(d) False positive rates change with shrinkage parameter $a$.

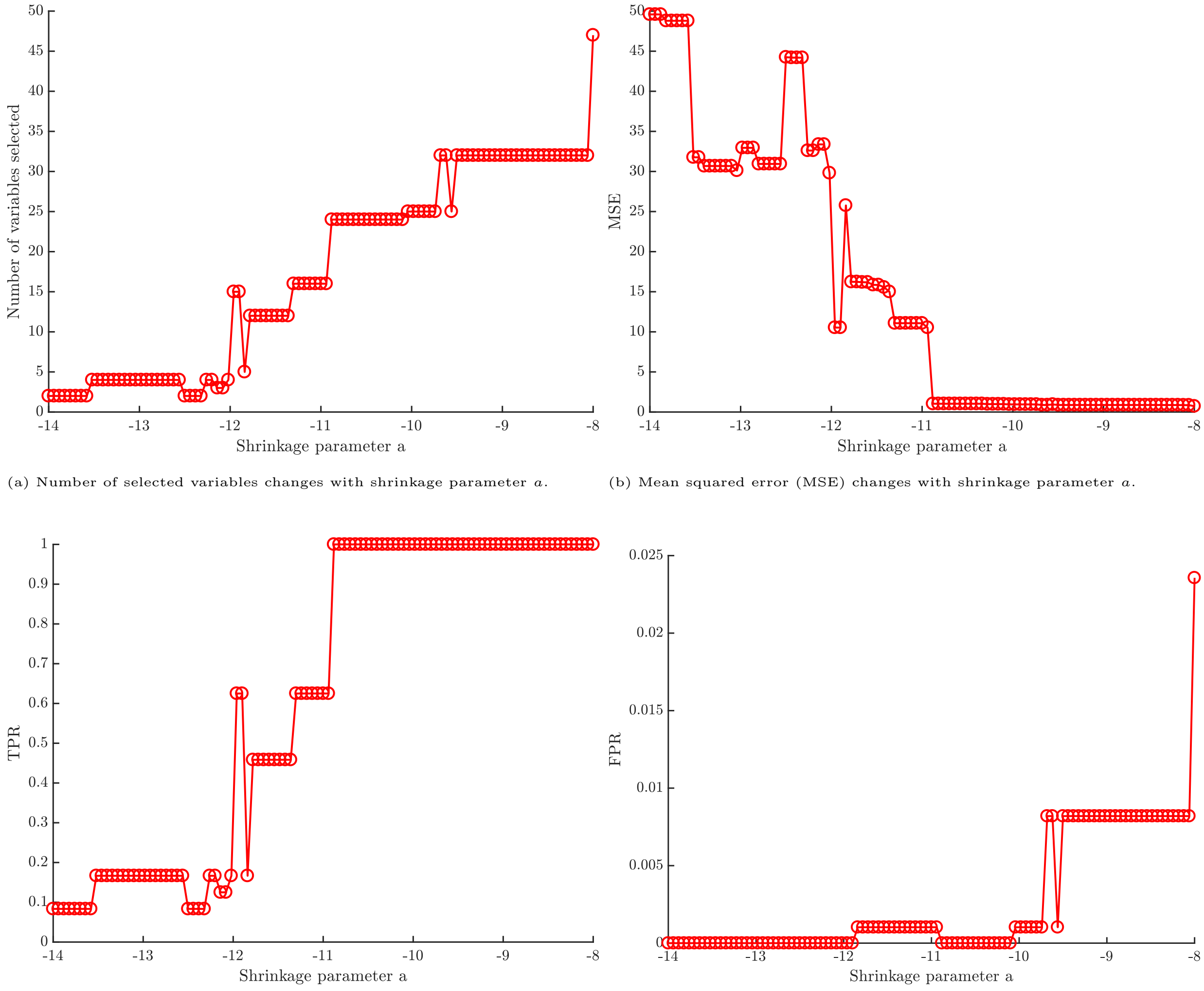
FIG. S3: The plot of sensitivity analysis of shrinkage parameter $a$ for simulations when $n = 100, p = 1000, \text{SNR} = 1$. 24 of the 1000 variables are set to be non-zero.

## S7. ADDITIONAL SIMULATION RESULTS

We run simulations on independent and dependent covariate structures. In Figure S7, we provide the selection and convergence results of three scenarios, from easy to difficult. When $n = 50$, $p = 30$, and $\text{SNR} = 1$, the MSE sequences started to converge at around 100 iterations, and all 6 true non-zero variables were correctly selected. When $n = 100$, $p = 1000$, and $\text{SNR} = 1$, the MSE sequences started to converge at around 4000 iterations, and all 24 true non-zero variables were correctly selected. When $n = 100$, $p = 1000$, and $\text{SNR} = 0.1$, the MSE sequences did not converge very well, and only 14 out of 24 truly non-zero variables were correctly selected.

Table S1 and S2 show the simulation results for independent data structure with different sample size $n$ and number of covariates $p$. The conclusion is the same with the independent data structure when $n = 100, p = 1000$ in the main manuscript.

As shown in Figure S8, we did histograms of phylogeny-induced correlations conducted by two types of correlation structures. The left panel shows the Euclidean correlation. The right panel shows the Exponential correlation when



(a) Number of selected variables changes with shrinkage parameter $a$.

(b) Mean squared error (MSE) changes with shrinkage parameter $a$.

(c) True positive rates change with shrinkage parameter $\boldsymbol{a}$.

(d) False positive rates change with shrinkage parameter $a$.

FIG. S4: The plot of sensitivity analysis of shrinkage parameter $a$ for simulations when $n = 100, p = 1000, \mathrm{SNR} = 1$. 24 of the 1000 variables are set to be non-zero. $Q$ is set to be one for false variables $\boldsymbol{j} = \{340 + 10l\}_{l=1}^{11}$.

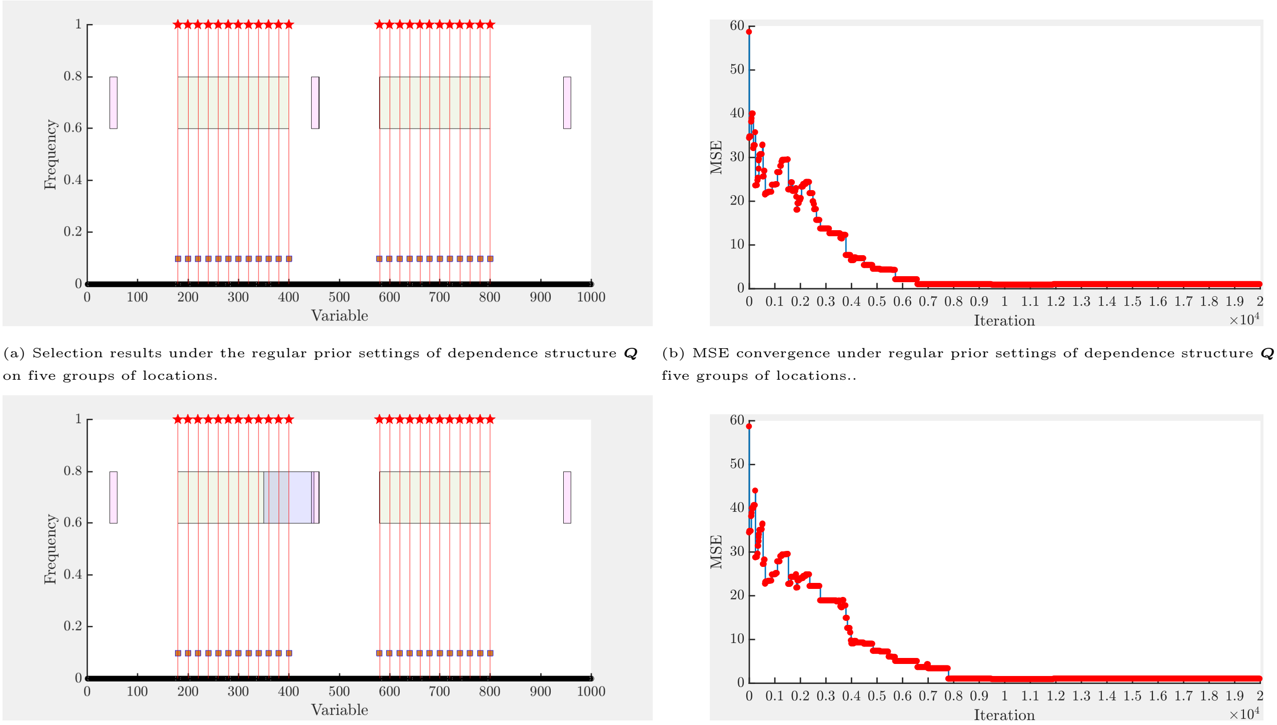
$\rho = 1.05$. They have same the 88% quantile 0.5 and similar shapes with mode around 0.2. Therefore, the Euclidean correlation structure can be considered as a special case of the exponential correlation structure, because larger $\rho$ (smaller $C_{ij}$) groups OTUs into clusters at a lower phylogenetic depth (a cluster is defined as a group of highly correlated OTUs). We included these options in our codes, and in this paper we used the Euclidean correlation structure to complete the analysis.

[1] C. G. Khatri, "Some results for the singular normal multivariate regression models," *Sankhyā: The Indian Journal of Statistics, Series A*, pp. 267–280, 1968.

[2] K.-S. Kwong and B. Iglewicz, "On singular multivariate normal distribution and its applications," *Computational statistics & data analysis*, vol. 22, no. 3, pp. 271–285, 1996.

[3] V. Pawlowsky-Glahn, J. J. Egozcue, and R. Tolosana Delgado, "Lecture notes on compositional data analysis," 2007.

[4] F. Li and N. R. Zhang, "Bayesian variable selection in structured high-dimensional covariate spaces with applications in



(a) Selection results under the regular prior settings of dependence structure $\boldsymbol{Q}$ on five groups of locations.

(b) MSE convergence under regular prior settings of dependence structure $\boldsymbol{Q}$ on five groups of locations..

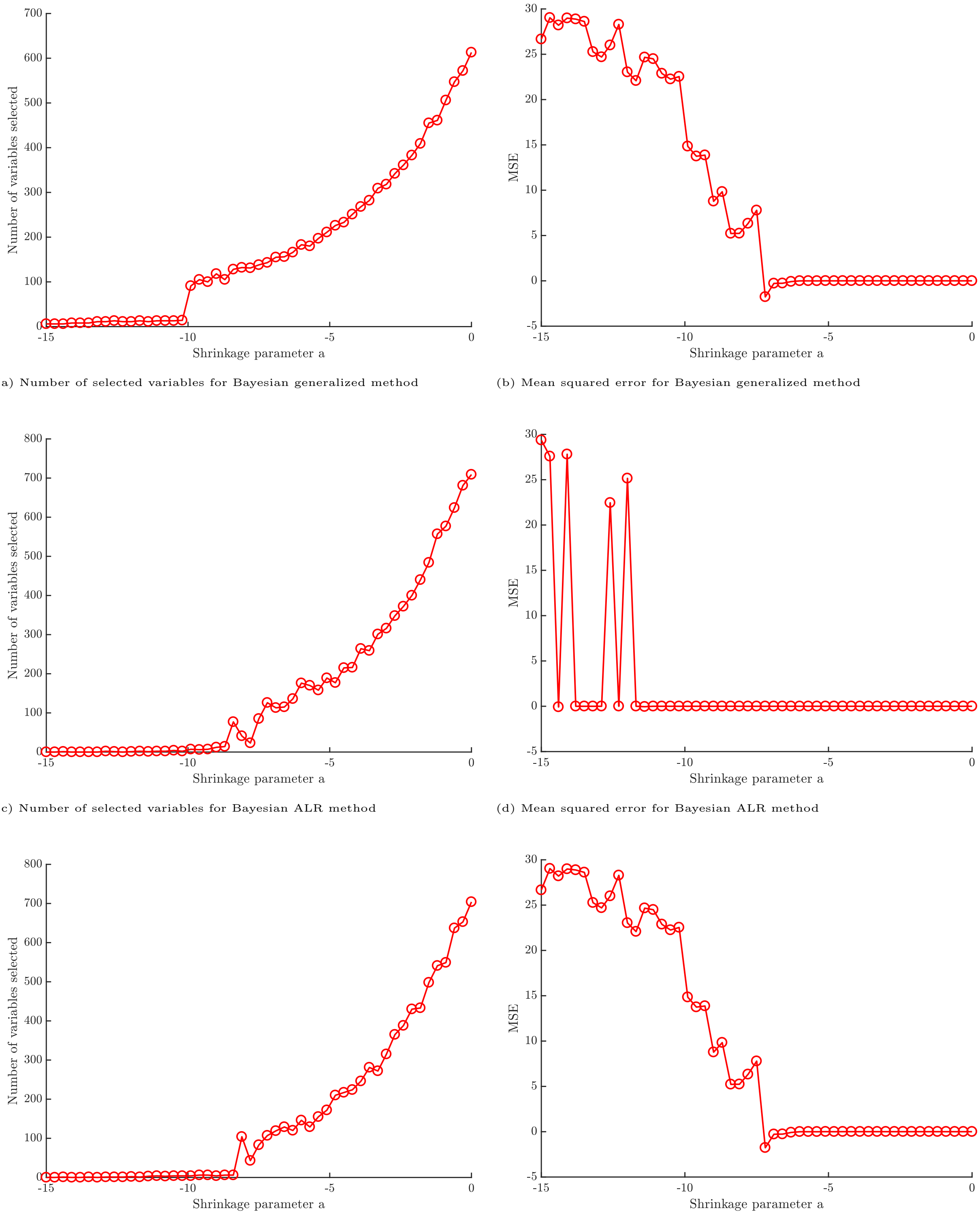
(c) Selection results when allowing $Q$ to be one for false variables $\boldsymbol{j} = \{340 + 10l\}_{l=1}^{11}$ besides the five groups of locations.

(d) MSE convergence when allowing $Q$ to be one for false variables $\boldsymbol{j} = \{340 + 10l\}_{l=1}^{11}$ besides the five groups of locations.

FIG. S5: Variable selection and model convergence of proposed Bayesian generalized method in different scenarios. In the left panels, the x-axis denotes the variables, and y-axis denotes the frequency of each variable being selected. The shaded green rectangles describes the structural prior Q put on the true signal region. The shaded purple rectangles describe the structural prior Q put on the false signal region. The red cubes on the bottom indicate the variables with true non-zero coefficients. The sticks with a star head indicate that the variable is selected with selection frequency more than 0.5. In the right panels, the x-axis denotes the number of iterations, and the y-axis denotes the mean squared estimation error. The sample size $n$ is 100, and the number of covariates $p$ is 1000. The SNR is 1.

genomics," *Journal of the American Statistical Association*, vol. 105, no. 491, pp. 1202–1214, 2010.

[5] A. Gelman *et al.*, "Prior distributions for variance parameters in hierarchical models (comment on article by browne and draper)," *Bayesian analysis*, vol. 1, no. 3, pp. 515–534, 2006.



(a) Number of selected variables for Bayesian generalized method

(b) Mean squared error for Bayesian generalized method

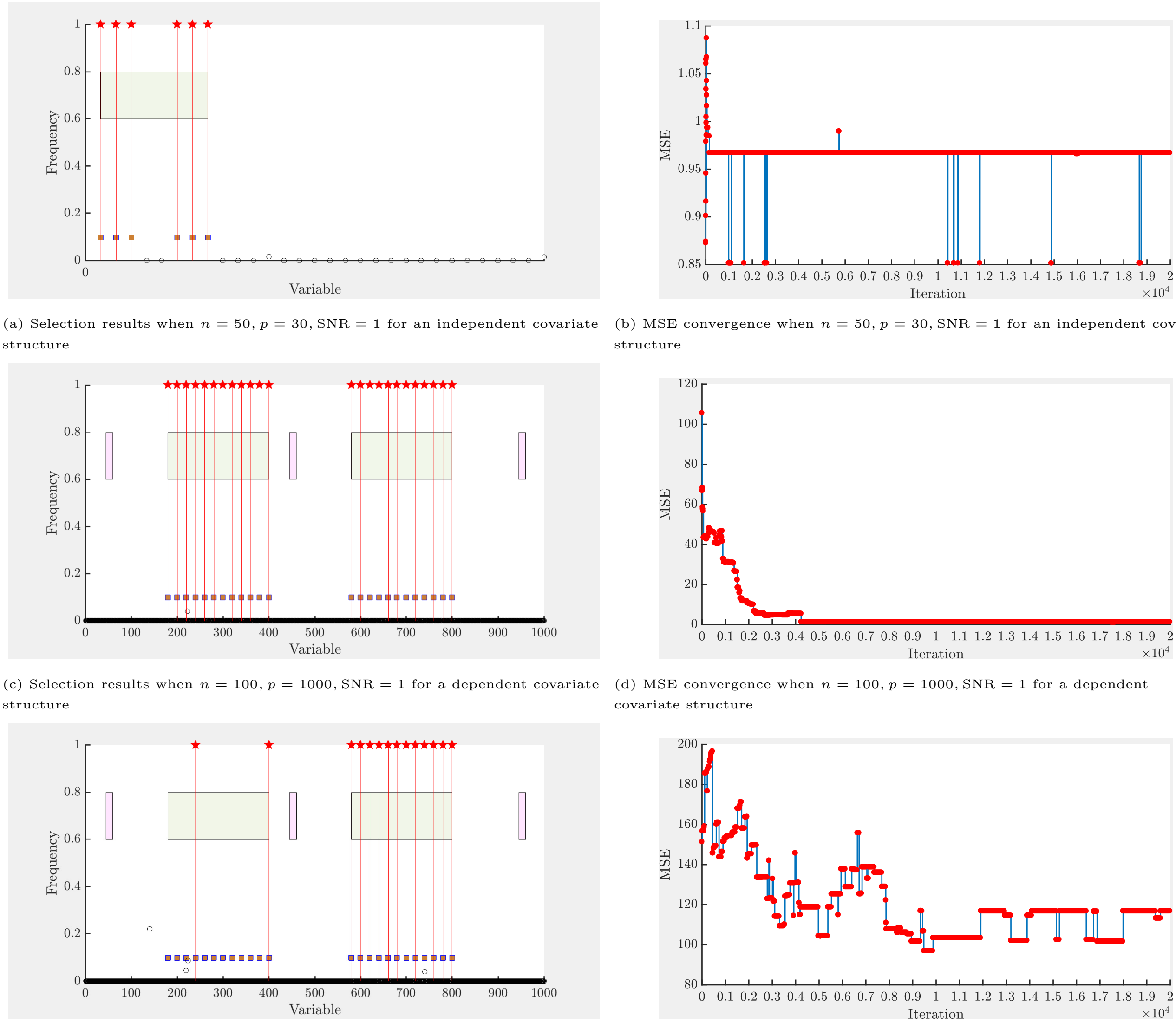
(c) Number of selected variables for Bayesian ALR method

(d) Mean squared error for Bayesian ALR method

(e) Number of selected variables for Bayesian CLR method

(f) Mean squared error for Bayesian CLR method

FIG. S6: Trace plot of sensitivity analysis of shrinkage parameter $a$ for the real application.



(a) Selection results when $n = 50, p = 30, \mathrm{SNR} = 1$ for an independent covariate structure

(b) MSE convergence when $n = 50, p = 30, \mathrm{SNR} = 1$ for an independent covariate structure

(c) Selection results when $n = 100, p = 1000, \mathrm{SNR} = 1$ for a dependent covariate structure

(d) MSE convergence when $n = 100, p = 1000, \mathrm{SNR} = 1$ for a dependent covariate structure

(e) Selection results when $n = 100, p = 1000, \mathrm{SNR} = 0.1$ for a dependent covariate structure

(f) MSE convergence when $n = 100, p = 1000, \mathrm{SNR} = 0.1$ for a dependent covariate structure

FIG. S7: Variable selection and model convergence of proposed Bayesian generalized method in different scenarios. In the left panels, the x-axis denotes the variables, and y-axis denotes the frequency of each variable being selected. The shaded green rectangles describes the structural prior $Q$ put on the true signal region. The shaded purple rectangles indicate the structural prior $Q$ put on the false signal region. The red cubes on the bottom indicate the variables with true non-zero coefficients. The sticks with a star head indicate that the variable is selected with selection frequency more than 0.5. In the right panels, the x-axis denotes the number of iterations, and the y-axis denotes the mean squared estimation error.

TABLE S1: Independent data structure with sample size $n = 50$ and number of covariates $p = 30$

| SNR | Method | PE | L1 loss | L2 loss | Linf loss | FP | FN |
|---|---|---|---|---|---|---|---|
| 10 | lasso ref | 0.0025 (0.0002) | 0.1469 (0.0047) | 0.0040 (0.0003) | 0.0357 (0.0013) | 2.3200 (0.1476) | 0 (0) |
| | lasso std | 0.0022 (0.0002) | 0.1266 (0.0037) | 0.0032 (0.0002) | 0.0331 (0.0010) | 1.4600 (0.1424) | 0 (0) |
| | lasso comp | 0.0021 (0.0002) | 0.1205 (0.0035) | 0.0028 (0.0002) | 0.0306 (0.0009) | 1.6000 (0.1456) | 0 (0) |
| | group lasso | 0.0015 (0.0001) | 0.1138 (0.0047) | 0.0014 (0.0001) | 0.0205 (0.0006) | 9.5500 (0.4808) | 0 (0) |
| | Bayesian ALR | 0.0359 (0.0022) | 0.1284 (0) | 0.0676 (0) | 0.0588 (0) | 0 (0) | 0 (0) |
| | Bayesian CLR | 0.0129 (0.0009) | 0.0841 (0) | 0.0300 (0) | 0.0200 (0) | 0 (0) | 0 (0) |
| | Bayesian general | 0.0123 (0.0008) | 0.0609 (0) | 0.0270 (0) | 0.0183 (0) | 0 (0) | 0 (0) |
| 5 | lasso ref | 0.0065 (0.0005) | 0.2121 (0.0068) | 0.0072 (0.0004) | 0.0492 (0.0015) | 4.0400 (0.2146) | 0 (0) |
| | lasso std | 0.0065 (0.0004) | 0.1956 (0.0060) | 0.0068 (0.0004) | 0.0489 (0.0015) | 3.0000 (0.1917) | 0 (0) |
| | lasso comp | 0.0062 (0.0004) | 0.1873 (0.0059) | 0.0062 (0.0004) | 0.0460 (0.0014) | 3.0000 (0.1917) | 0 (0) |
| | group lasso | 0.0060 (0.0004) | 0.2267 (0.0091) | 0.0058 (0.0003) | 0.0412 (0.0011) | 9.4400 (0.4753) | 0 (0) |
| | Bayesian ALR | 0.0658 (0.0042) | 0.1382 (0) | 0.0589 (0) | 0.0419 (0) | 0 (0) | 0 (0) |
| | Bayesian CLR | 0.0472 (0.0030) | 0.1185 (0) | 0.0440 (0) | 0.0339 (0) | 0 (0) | 0 (0) |
| | Bayesian general | 0.0455 (0.0031) | 0.0816 (0) | 0.0393 (0) | 0.0297 (0) | 0 (0) | 0 (0) |
| 1 | lasso ref | 0.1520 (0.0110) | 1.0258 (0.0336) | 0.1486 (0.0096) | 0.2181 (0.0067) | 5.3500 (0.2993) | 0.01 (0.01) |
| | lasso std | 0.1588 (0.0110) | 0.9858 (0.0301) | 0.1546 (0.0097) | 0.2277 (0.0074) | 3.9900 (0.2787) | 0.01 (0.01) |
| | lasso comp | 0.1495 (0.0108) | 0.9335 (0.0292) | 0.1347 (0.0076) | 0.2100 (0.0064) | 4.1300 (0.2932) | 0 (0) |
| | group lasso | 0.1510 (0.0110) | 1.1264 (0.0450) | 0.1428 (0.0077) | 0.2057 (0.0056) | 9.3500 (0.4696) | 0 (0) |
| | Bayesian ALR | 1.0658 (0.0739) | 0.4112 (0) | 0.1911 (0) | 0.1274 (0) | 0 (0) | 0 (0) |
| | Bayesian CLR | 1.2030 (0.0744) | 0.3933 (0) | 0.1645 (0) | 0.1448 (0) | 0 (0) | 0 (0) |
| | Bayesian general | 1.1415 (0.0751) | 0.2801 (0) | 0.1447 (0) | 0.1207 (0) | 0 (0) | 0 (0) |

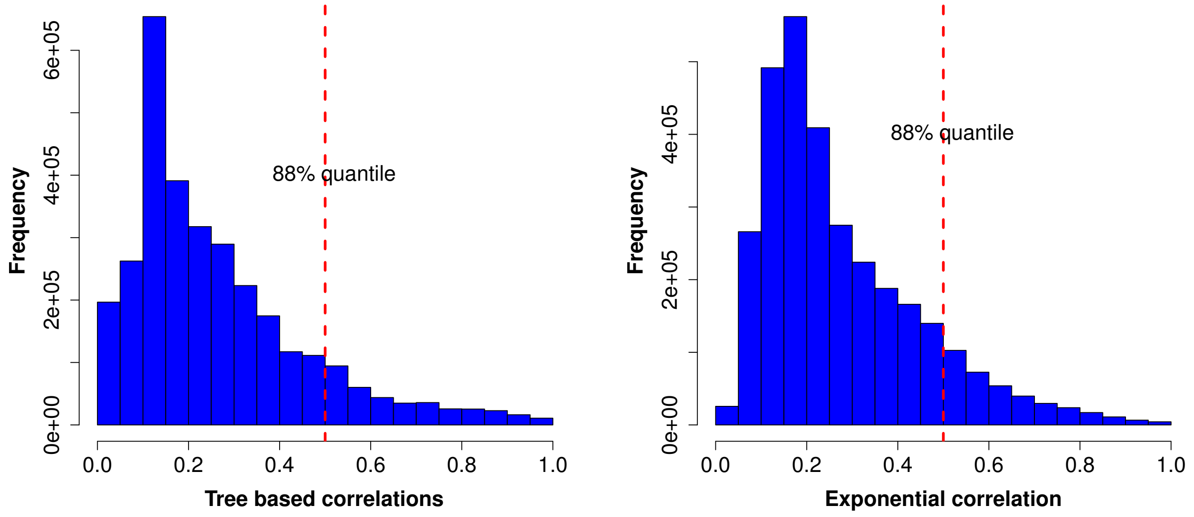


FIG. S8: Histograms of phylogeny-induced correlations conducted by two types of correlation structures.

TABLE S2: Independent data structure with sample size $n = 100$ and number of covariates $p = 200$

| SNR | Method | PE | L1 loss | L2 loss | Linf loss | FP | FN |
|---|---|---|---|---|---|---|---|
| 10 | lasso ref | 0.0019 (0.0001) | 0.1205 (0.0025) | 0.0024 (0.0001) | 0.0266 (0.0006) | 2.9500 (0.1904) | 0 (0) |
| | lasso std | 0.0018 (0.0001) | 0.1104 (0.0022) | 0.0022 (0.0001) | 0.0264 (0.0006) | 1.3200 (0.1294) | 0 (0) |
| | lasso comp | 0.0018 (0.0001) | 0.1049 (0.0020) | 0.0020 (0.0001) | 0.0244 (0.0005) | 1.5800 (0.1545) | 0 (0) |
| | group lasso | 0.0191 (0.0010) | 0.5043 (0.0083) | 0.0467 (0.0017) | 0.1131 (0.0025) | 0.1000 (0.0389) | 0 (0) |
| | Bayesian ALR | 0.0592 (0.0443) | 0.0505 (0.0116) | 0.0434 (0.0076) | 0.1630 (0.0987) | 0.02 (0.0141) | 0.03 (0.03) |
| | Bayesian CLR | 0.0087 (0.0004) | 0.0056 (0) | 0.0027 (0) | 0.0021 (0) | 0 (0) | 0 (0) |
| | Bayesian general | 0.0098 (0.0014) | 0.0057 (0) | 0.0029 (0) | 0.0023 (0) | 0 (0) | 0 (0) |
| 5 | lasso ref | 0.0067 (0.0003) | 0.2141 (0.0049) | 0.0072 (0.0003) | 0.0465 (0.0011) | 3.9000 (0.2209) | 0 (0) |
| | lasso std | 0.0065 (0.0003) | 0.1958 (0.0044) | 0.0068 (0.0003) | 0.0466 (0.0011) | 2.2700 (0.1948) | 0 (0) |
| | lasso comp | 0.0063 (0.0003) | 0.1878 (0.0041) | 0.0061 (0.0003) | 0.0436 (0.0010) | 2.3100 (0.1947) | 0 (0) |
| | group lasso | 0.0226 (0.0011) | 0.5074 (0.0085) | 0.0476 (0.0017) | 0.1157(0.0026) | 0.2200 (0.0596) | 0 (0) |
| | Bayesian ALR | 0.0402 (0.0019) | 0.0782 (0) | 0.0426 (0) | 0.0391 (0) | 0 (0) | 0 (0) |
| | Bayesian CLR | 0.0344 (0.0013) | 0.0103 (0) | 0.0050 (0) | 0.0039 (0) | 0 (0) | 0 (0) |
| | Bayesian general | 0.0333 (0.0014) | 0.0108 (0) | 0.0053 (0) | 0.0040 (0) | 0.01 (0.01) | 0 (0) |
| 1 | lasso ref | 0.1682 (0.0084) | 1.0664 (0.0246) | 0.1789 (0.0079) | 0.2316 (0.0056) | 3.9600 (0.2183) | 0 (0) |
| | lasso std | 0.1630 (0.0077) | 0.9823 (0.0223) | 0.1714 (0.0077) | 0.2334 (0.0056) | 2.2800 (0.2099) | 0 (0)) |
| | lasso comp | 0.1570 (0.0071) | 0.9385 (0.0208) | 0.1518 (0.0065) | 0.2170 (0.0052) | 2.4400 (0.2081) | 0 (0) |
| | group lasso | 0.1410 (0.0062) | 0.9868 (0.0261) | 0.1096 (0.0045) | 0.1894 (0.0041) | 13.4800 (0.5584) | 0 (0) |
| | Bayesian ALR | 0.8698 (0.0428) | 0.1815 (0.0032) | 0.0873 (0.0016) | 0.0679 (0.0012) | 0.09 (0.0288) | 0 (0) |
| | Bayesian CLR | 0.8851 (0.0379) | 0.1063 (0.0037) | 0.0459 (0.0013) | 0.0279 (0.0014) | 0.01 (0.01) | 0 (0) |
| | Bayesian general | 0.8504 (0.0369) | 0.1049 (0.0019) | 0.0459 (0.0010) | 0.0278 (0.0010) | 0.02 (0.0141) | 0 (0) |