Supplementary material for ‘An approximate Bayesian approach to covariate-dependent graphical modeling’

Appendices

A Assumptions for the theoretical results

A.1 Assumptions for continuous covariate-dependent model

Assumptions on true data generating distribution:

Assumption T1 (Sparsity in β). Assume that $\beta_j^*(\mathrm{z})$ has at most s_j^* non-zero elements for any z in its support, and let $s^* = \max_{j=1}^p s_j^*$ with $s^* \geq 1$. Suppose $n \geq c \max\{s^* \log p, s^* \log n\}$ for some constant $c > 1$. In addition, we assume $n = o(p^C)$ for some positive constant $C > 0$.

Assumption T2 (Sparsity in derivatives). Assume that for any z, $\dot{\beta}_j^*(\mathrm{z})$ and $\ddot{\beta}_j^*(\mathrm{z})$ have at most $c_5 s_j^*$ non-zero elements for $j = 1, ..., p$, for some constant $c_5 > 0$.

Assumption T3 (Smoothness). We assume that up to the second order derivatives for all the components of the graph coefficient with respect to the covariates are uniformly bounded by a constant. That is, $\|\beta_j^*(\mathrm{z})\|_\infty$,$\|\dot{\beta}_j^*(\mathrm{z})\|_\infty$, $\|\ddot{\beta}_j^*(\mathrm{z})\|_\infty$ are uniformly bounded above by constants for any z, j.

Assumption T4 (Random design). Suppose $\mathbf{z}_1, ..., \mathbf{z}_n$ are i.i.d. samples from a distribution with density $f(\mathrm{z})$ on a compact support, where $|f(\mathrm{z})|$, $\dot{f}(\mathrm{z})$ and $\ddot{f}(\mathrm{z})$ are all bounded below and above by constants.

Assumption T5 (Eigenvalue Conditions). All eigenvalues of $\Sigma(\mathrm{z})$ and $\dot{\Sigma}(\mathrm{z})$ are uniformly upper and lower bounded by constants for any z. In addition, suppose that the marginal distribution of x_i with density $\int f(x \mid \Sigma(\mathrm{z}))f(\mathrm{z})dz$ is a sub-Gaussian random vector with covariance Σ. All eigenvalues of Σ are also uniformly upper and lower bounded by constants.

Assumptions on the model and prior:

Assumption K (Kernel property). Suppose the kernel function K used to fit weights satisfies: $\sup_x |K(x)| \leq c_3 < \infty$, $\int K(x)dx = 1$, $\int K^2(x)dx = c_0 < \infty$, $\int xK(x)dx = 0$,$\int x^2K(x)dx = c_2 < \infty$.

Assumption P. We assume a spike-and-slab prior $p_{\beta_j|\gamma_j}(\beta_j)p_{\gamma_j}(\gamma_j)$ for the parameter $\theta_j = (\beta_j, \gamma_j)$ with $p_{\beta_j|\gamma_j} = \prod_{k=1,k\neq j}^{p} \mathcal{N}(\beta_{jk}; 0, \sigma_*^2)^{\gamma_{jk}} \delta_0(\beta_{jk})^{1-\gamma_{jk}}$, and $p_{\gamma_j}(\gamma_j) \geq \exp\{-C\|\gamma_j\|_0 \log p\}$, where $\|\gamma_j\|_0$ is the number of non-zero elements of γ_j.

Assumption T1 describes a relationship between n, p and s. A constraint of $n \geq cs^* \log p$ is assumed to ensure that the restricted eigenvalue conditions hold for sample covariances, see Raskutti et al. (2010); Zhou (2009). In addition, $n \geq cs^* \log n$ is assumed to guarantee that the error rate is $O(1)$ in the case when $n > p$. Finally, $n = o(p^C)$ ensures the consistency holds with a high probability when a bound for the maximum of risks across subjects is considered. Since the risk bound for a single subject holds with probability $1 - p^{-C}$, the maximum of the risk can be bounded with probability $1 - np^{-C}$ using the union bound which requires $n = O(p^C)$. **Assumption T2** posits sparsity in the first and second derivatives of $\beta(\mathrm{z})$ with z. Essentially, the assumption implies that $\beta_{jk}(\mathrm{z})$ satisfies $\beta_{jk}(\mathrm{z}) = 0$ for $\mathrm{z} \in [a_{jk}, b_{jk}]$, where $[a_{jk}, b_{jk}]$ is a constant length interval in the support of $\beta_{jk}(\mathrm{z})$ for $j = 1, ..., p$, $k = 1, ..., (p-1)$. Here, $\beta_{jk}(\mathrm{z})$ denotes the k-th coordinate of β_j as a function of z. One such example is $(1 - \mathrm{z}^2)^2 I(|\mathrm{z}| \leq 1)$ for $\mathrm{z} \in [-2, 2]$, which is zero for $\mathrm{z} \in [-2, -1]$ and $\mathrm{z} \in [1, 2]$. **Assumption T3** ensures that the covariates carry information about the graph coefficients, together with some regularity conditions on the covariance matrix. **Assumption T4** asserts that the sampled covariates are representative in the sense that they are i.i.d. from some homogeneous distributions, e.g., uniform distributions on a bounded interval. **Assumption K** indicates that the kernel function should be smooth enough to capture the shared information across subjects. The **Assumption P** for priors encompasses a wide variety of prior distributions, as discussed in Castillo et al. (2012).

A.2 Assumptions for discrete covariate-dependent/covariate-independent graph models

Assumption W: Let $c_l = \min_{k:z_k \neq z_l} |\mathbf{z}_k - \mathbf{z}_l|$ and assume that c_l is lower bounded by positive constants for $l = 1, ..., K$. Suppose the Gaussian kernel $K(x) \propto e^{-x^2}$ is used and the tuning parameter τ for the kernel satisfies $\tau = c \min_l c_l / \sqrt{\log n}$ for some positive constant $c < 1$. Then we have the following result, for a graph estimate of an individual with covariate level $\mathbf{z}_l$.

Assumption T: We assume that the underlying data at each covariate level is generated from a homogeneous dependence structure, as described in (4). Given a covariate value z, $\Omega^*(\mathrm{z})$ has maximum and minimum eigenvalues bounded away from 0 and ∞. Assume that $\beta_j^*(\mathrm{z})$ has at most s_j^* non-zero elements, and let $s^* = \max_j \{s_j^*\}$.

Assumption A: Assume that β_j^* has at most s_j^* non-zero elements, and let $s^* = \max_j \{s_j^*\}$. Assume that $s^* \log(np)/n \to 0$ for $j = 1, ..., p$.

B Proofs of main theorems

Notations. We first define the following terms:

$$\underset{\sim}{\kappa}(s, \Omega^*(\mathrm{z})) = \inf \left\{ \frac{u^{\mathrm{T}} \Omega^* u(\mathrm{z})}{n \|u\|_2^2} : u \in \mathbb{R}^p, 1 \leq \|u\|_0 \leq s \right\},$$

$$\tilde{\kappa}(s, \Omega^*(\mathrm{z})) = \sup \left\{ \frac{u^{\mathrm{T}} \Omega^* u(\mathrm{z})}{n \|u\|_2^2} : u \in \mathbb{R}^p, 1 \leq \|u\|_0 \leq s \right\},$$

where $\underset{\sim}{\kappa}(s, \Omega^*(\mathrm{z}))$ is the minimum eigenvalue, and $\tilde{\kappa}(s, \Omega^*(\mathrm{z}))$ is the maximum eigenvalue of $\Omega^*(\mathrm{z})$ for s-sparse matrices. Define $\gamma_{jk}(\mathrm{z}) = \mathbb{I}\{\beta_{jk}(\mathrm{z}) \neq 0\}$ which denotes the number of non-zero entries of the coefficient parameter $\beta_{jk}(\mathrm{z})$. Also, define:

$$\underset{\sim}{\kappa}(s, \mathbf{X}) = \inf \left\{ \frac{u^{\mathrm{T}} \mathbf{X}^{\mathrm{T}} \mathbf{X} u}{\|u\|_2^2} : u \in \mathbb{R}^p, 1 \leq \|u\|_0 \leq s \right\}, \quad \tilde{\kappa}(s, \mathbf{X}) = \sup \left\{ \frac{u^{\mathrm{T}} \mathbf{X}^{\mathrm{T}} \mathbf{X} u}{\|u\|_2^2} : u \in \mathbb{R}^p, 1 \leq \|u\|_0 \leq s \right\}$$

where $\underset{\sim}{\kappa}(s, \mathbf{X})$is the minimum and $\tilde{\kappa}(s, \mathbf{X})$ is the maximum eigenvalue of $\mathbf{X}$. Let $s^* = \max\{s_j^*, j = 1, \ldots, p\}$, where s_j* be the true sparsity structure of the j-th row of $\Omega^*(\mathrm{z})$. Let

$$\underline{\kappa}(s, \mathbf{X}) = \inf_{\gamma_j, \|\gamma_j\|_0 \leq s} \inf \left\{ \frac{u^{\mathrm{T}} \mathbf{X}^{\mathrm{T}} \mathbf{X} u}{n \|u\|_2^2} : u \in \mathbb{R}^p, \sum_{k: \gamma_{jk}=0} |u| \leq 7 \sum_{k: \gamma_{jk}=1} |u| \right\}$$

and

$$\underline{\kappa}(\Omega^*(\mathrm{z})) = \inf \left\{ \frac{u^{\mathrm{T}} \Omega^*(\mathrm{z}) u}{\|u\|_2^2} : u \in \mathbb{R}^p, \sum_{k: \gamma_{jk}=0} |u| \leq 7 \sum_{k: \gamma_{jk}=1} |u| \right\}$$

Define the set $\mathcal{G}_{n,p} = \{\mathrm{z} \in \mathbb{R}^{n \times p} : \tilde{\kappa}(s^*, \mathrm{z}) \leq c_1 \tilde{\kappa}(s^*, \Omega^*(\mathrm{z})), \tilde{\kappa}(1, \mathrm{z}) \leq c_1 \tilde{\kappa}(1, \Omega^*(\mathrm{z})), \underline{\kappa}(s^*, \mathrm{z}) \geq c_2 \underline{\kappa}(\Omega^*(\mathrm{z}))\}$. Then, following Atchadé (2019) we have that if $\mathbf{X} \in \mathcal{G}_{n,p}$, then $\mathbf{X}_{-j} \in \mathcal{G}_{n,p-1}$ for any $j \in \{1, 2, \ldots, p\}$.

Let A_{-j} be the submatrix of matrix A except the j-th row and $A_{-j,-j}$ be the submatrix of matrix A except the j-th row and column. We denote

$$\Psi^w(q_{\theta_j^l}(\mathbf{z})) = \int \log \frac{p^w(x_j \mid \tilde{\theta}_j^l(\mathbf{z}), \mathbf{X}_{-j}, \mathbf{z})}{p^w(x_j \mid \theta_j^l(\mathbf{z}), \mathbf{X}_{-j}, \mathbf{z})} q_{\theta_j^l}(\theta_j^l(\mathbf{z})) d\theta_j^l(\mathbf{z}) + \alpha^{-1} \mathrm{D}_{\mathrm{KL}}(q_{\theta_j^l} \| p_{\theta_j^l}).$$

Because subject index l do not change during the proof of Theorem 1, Theorem 3, and Corollary 1, we ignore them in these proofs.

B.1 Proof of Lemma 1

Proof. For any observation l, we have the marginalized KL divergence:

$$\begin{aligned}
&\int p(\mathbf{X} \mid \Theta^*(\mathbf{z}_1), ..., \Theta^*(\mathbf{z}_n)) \log \frac{p(\mathbf{X} \mid \Theta^*(\mathbf{z}_1), ..., \Theta^*(\mathbf{z}_n))}{\prod_{j=1}^p p^{w_l}(x_j \mid \mathbf{X}_{-j}, \Theta_{-j}^l(\mathbf{z}))} d\mathbf{X} \\
&= c - \sum_{j=1}^p \int p(\mathbf{X} \mid \Theta^*(\mathbf{z}_1), ..., \Theta^*(\mathbf{z}_n)) \log p^{w_l}(x_j \mid \mathbf{X}_{-j}, \Theta_{-j}^l(\mathbf{z})) d\mathbf{X} \\
&= c - \sum_{j=1}^p \int p(\mathbf{X}_{-j} \mid \Theta_{-j,-j}^*(\mathbf{z}_1), ..., \Theta_{-j,-j}^*(\mathbf{z}_n)) p(x_j \mid \mathbf{X}_{-j}, \Theta_{-j}^*(\mathbf{z}_1), ..., \Theta_{-j}^*(\mathbf{z}_n)) \\
&\log p^{w_l}(x_j \mid \mathbf{X}_{-j}, \Theta_{-j}^l(\mathbf{z})) dx_j d\mathbf{X}_{-j},
\end{aligned}$$

where $\Theta^{l}_{-j}(\mathbf{z})$ is the targeted coefficient with l-th covariate.

Therefore, $\tilde{\beta}^{l}_{j}(\mathbf{z})$ should be the minimizer of the following objective function

$$\begin{aligned}
&-\int p(\mathbf{X}_{-j} \mid \Theta^{*}_{-j,-j}(\mathbf{z}_1), ..., \Theta^{*}_{-j,-j}(\mathbf{z}_n))p(x_j \mid \mathbf{X}_{-j}, \Theta^{*}_{-j}(\mathbf{z}_1), ..., \Theta^{*}_{-j}(\mathbf{z}_n)) \\
&\times \log p^{w_l}(x_j \mid \mathbf{X}_{-j}, \beta^{l}_{j}(\mathbf{z}), \gamma^{l}_{j}(\mathbf{z}))dx_j d\mathbf{X}_{-j} \\
&= \mathbf{E}_{\mathbf{X}_{-j},x_j}\left\{\sum_{k=1}^{n}(x_{kj} - \mathbf{x}^{\mathrm{T}}_{k,-j}\beta^{l}_{j}(\mathbf{z}))^{\mathrm{T}}\frac{\mathrm{w}_l(\mathbf{z}_k)}{2\sigma^2_*}(x_{kj} - \mathbf{x}^{\mathrm{T}}_{k,-j}\beta^{l}_{j}(\mathbf{z}))\right\} \\
&= \mathbf{E}_{\mathbf{X}_{-j}}\left\{\sum_{k=1}^{n}(\beta^{*}_{j}(\mathbf{z}_k) - \beta^{l}_{j}(\mathbf{z}))^{\mathrm{T}}\frac{\mathrm{w}_l(\mathbf{z}_k)}{2\sigma^2_*}\mathbf{x}_{k,-j}\mathbf{x}^{\mathrm{T}}_{k,-j}(\beta^{*}_{j}(\mathbf{z}_k) - \beta^{l}_{j}(\mathbf{z}))\right\} \\
&= \sum_{k=1}^{n}(\beta^{*}_{j}(\mathbf{z}_k) - \beta^{l}_{j}(\mathbf{z}))^{\mathrm{T}}\frac{\mathrm{w}_l(\mathbf{z}_k)}{2\sigma^2_*}\Sigma^{*}_{-j,-j}(\mathbf{z}_k)(\beta^{*}_{j}(\mathbf{z}_k) - \beta^{l}_{j}(\mathbf{z})),
\end{aligned}$$

under the constraint $\|\beta^{l}_{j}(\mathbf{z})\|_0 \leq C_0 s^{*}_{j}$ for $C_0 \geq 1$.

Since $\|\beta^{*}_{j}(\mathbf{z}_l)\|_0 \leq s^{*}_{j}$ is in the constrained region, by basic inequality, we have

$$\begin{aligned}
&\sum_{k=1}^{n}(\beta^{*}_{j}(\mathbf{z}_k) - \tilde{\beta}^{l}_{j}(\mathbf{z}))^{\mathrm{T}}\frac{\mathrm{w}_l(\mathbf{z}_k)}{2\sigma^2_*}\Sigma^{*}_{-j,-j}(\mathbf{z}_k)(\beta^{*}_{j}(\mathbf{z}_k) - \tilde{\beta}^{l}_{j}(\mathbf{z})) \\
\leq &\sum_{k=1}^{n}(\beta^{*}_{j}(\mathbf{z}_k) - \beta^{*}_{j}(\mathbf{z}_l))^{\mathrm{T}}\frac{\mathrm{w}_l(\mathbf{z}_k)}{2\sigma^2_*}\Sigma^{*}_{-j,-j}(\mathbf{z}_k)(\beta^{*}_{j}(\mathbf{z}_k) - \beta^{*}_{j}(\mathbf{z}_l)),
\end{aligned}$$

where $\Sigma^{*}_{-j,-j}(\mathbf{z}_k)$ is the submatrix of the k-th true covariance except the j-th row and column. After some algebra, we have

$$(\beta^{l}_{j}(\mathbf{z})^{*} - \tilde{\beta}^{l}_{j}(\mathbf{z}))^{\mathrm{T}}\sum_{k=1}^{n}\mathrm{w}_l(\mathbf{z}_k)\Sigma^{*}_{-j,-j}(\mathbf{z}_k)(\beta^{l}_{j}(\mathbf{z})^{*} - \tilde{\beta}^{l}_{j}(\mathbf{z})) \leq 2(\tilde{\beta}^{l}_{j}(\mathbf{z}) - \beta^{*}_{j}(\mathbf{z}_l))(\sum_{k=1}^{n}\mathrm{w}_l(\mathbf{z}_k)\Sigma^{*}_{-j,-j}(\mathbf{z}_k)\beta^{*}_{j}(\mathbf{z}_k)) \tag{1}$$

Since the eigenvalues of $\Sigma^{*}_{-j,-j}(\mathbf{z}_k)$ are all lower bounded by constant, by Weyl's inequality, we have $\lambda_{\min}(\sum_{k=1}^{n}\mathrm{w}_l(\mathbf{z}_k)\Sigma^{*}_{-j,-j}(\mathbf{z}_k))$ lower bounded by constant multiplied by $\sum_{k=1}^{n}\mathrm{w}_k$.

Note that

$$\mathbf{E}\left\{\sum_{k=1}^{n} \mathrm{w}_l(\mathbf{z}_k)\right\} = \mathbf{E}\left\{\frac{1}{\tau}\sum_{k=1}^{n} K\left(\frac{\mathbf{z}_k - \mathbf{z}_l}{\tau}\right)\right\} = n\int K\left(u\right) f(\mathbf{z}_l + \tau u)du$$

$$= c_0 n f(\mathbf{z}_l) + o(n).$$

Therefore, we have $\mathbf{E}(\sum_{k=1}^{n} \mathrm{w}_l(\mathbf{z}_k)) \geq cn$ for some positive constant c. Applying the above lower eigenvalues and Cauchy-Schwartz in equality on equation (1) after taking expectation to z, we have

$$\mathbf{E}_z \|\beta_j^*(\mathbf{z}_l) - \tilde{\beta}_j^l(\mathbf{z})\|_2^2 \lesssim \mathbf{E}_z \|\frac{1}{n}\sum_{k=1}^{n} \mathrm{w}_l(\mathbf{z}_k)\Sigma_{-j,-j}^*(\mathbf{z}_k)(\beta_j^*(\mathbf{z}_k) - \beta_j^*(\mathbf{z}_l))\|_2^2.$$

$$\frac{1}{n}\sum_{k=1}^{n} \mathrm{w}_l(\mathbf{z}_k)\Sigma_{-j,-j}^*(\mathbf{z}_k)(\beta_j^*(\mathbf{z}_k) - \beta_j^*(\mathbf{z}_l)) = \frac{c_l}{n\tau}\sum_{k=1}^{n} K\left(\frac{\mathbf{z}_k - \mathbf{z}_l}{\tau}\right)\Sigma_{-j,-j}^*(\mathbf{z}_k)(\beta_j^*(\mathbf{z}_k) - \beta_j^*(\mathbf{z}_l)).$$

Then we have

$$\mathbf{E}_{\mathbf{z}_k}(\frac{1}{n}\sum_{k=1}^{n} \mathrm{w}_l(\mathbf{z}_k)\Sigma_{-j,-j}^*(\mathbf{z}_k)(\beta_j^*(\mathbf{z}_k) - \beta_j^*(\mathbf{z}_l))) = \frac{c_l}{\tau}\int K\left(\frac{\mathbf{z} - \mathbf{z}_l}{\tau}\right)\Sigma_{-j,-j}^*(\mathbf{z})\left(\beta_j^*(\mathbf{z}) - \beta_j^*(\mathbf{z}_l)\right) f(\mathbf{z})dz$$

$$= c_l\int K(u)\Sigma_{-j,-j}^*(\mathbf{z}_l + \tau u)(\beta_j^*(\mathbf{z}_l + \tau u) - \beta_j^*(\mathbf{z}_l))f(\mathbf{z}_l + \tau u)du.$$

Expanding $\Sigma_{-j,-j}^*(\mathbf{z}_l + \tau u)$, $\beta_j^*(\mathbf{z}_l + \tau u)$ and $f(\mathbf{z}_l + \tau u)$ component-wisely in Taylor expansion, we have

$$
\begin{aligned}
&\mathbf{E}(\sum_{k=1}^{n} \mathrm{w}_l(\mathbf{z}_k)\Sigma^*_{-j,-j}(\mathbf{z}_k)(\beta^*_j(\mathbf{z}_k)-\beta^*_j(\mathbf{z}_l))) \\
&= c_l \int K(u)\left\{\Sigma^*_{-j,-j}(\mathbf{z}_l)+\tau\mu\dot{\Sigma}^*_{-j,-j}(\mathbf{z}_l^{(1)})\right\}\left\{\tau u\dot{\beta}^*_j(\mathbf{z}_l)+\frac{\tau^2}{2}u^2\ddot{\beta}^*_j(\mathbf{z}_l^{(2)})\right\}\times \\
&\left\{f(\mathbf{z}_l)+u\tau\dot{f}(\mathbf{z}_l^{(3)})\right\}du \\
&= c_l\left(\int uK(u)du\right)\tau\Sigma^*_{-j,-j}(\mathbf{z}_l)\dot{\beta}^*_j(\mathbf{z}_l)f(\mathbf{z}_l)+ \\
&\left(\int u^2K(u)du\right)\tau^2\left(\Sigma^*_{-j,-j}(\mathbf{z}_l)\frac{1}{2}\ddot{\beta}^*_j(\mathbf{z}_l^{(2)})f(\mathbf{z}_l)+\Sigma^*_{-j,-j}(\mathbf{z}_l)\dot{\beta}^*_j(\mathbf{z}_l)\dot{f}(\mathbf{z}_l^{(3)})+\dot{\Sigma}^*_{-j,-j}(\mathbf{z}_l^{(1)})\dot{\beta}^*_j(\mathbf{z}_l)f(\mathbf{z}_l)\right)+o(\tau^2) \\
&= c_lc_2\tau^2\left(\dot{\Sigma}^*_{-j,-j}(\mathbf{z}_l^{(1)})\dot{\beta}^*_j(\mathbf{z}_l)f(\mathbf{z}_l)+\frac{1}{2}\Sigma^*_{-j,-j}(\mathbf{z}_l)\ddot{\beta}^*_j(\mathbf{z}_l^{(2)})f(\mathbf{z}_l)+\Sigma^*_{-j,-j}(\mathbf{z}_l)\dot{\beta}^*_j(\mathbf{z}_l)\dot{f}(\mathbf{z}_l^{(3)})\right)+o(\tau^2).
\end{aligned}
$$

where $\mathbf{z}_l^{(1)}, \mathbf{z}_l^{(2)}, \mathbf{z}_l^{(3)}$ in the first equation are between $\mathbf{z}_l$ and $\mathbf{z}_l+\tau\mu$. Note that the ℓ_2 norm of the reminder term is no larger than s^*_j up to some constant factor given that $\dot{\beta}^*_j(\mathbf{z}_l)$ and $\ddot{\beta}^*_j(\mathbf{z}_l^{(2)})$ are s^*_j sparse and $\|\Sigma^*_{-j,-j}(\mathbf{z}_l)\|_2$ and $\|\dot{\Sigma}^*_{-j,-j}(\mathbf{z}_l^{(1)})\|_2$ are upper bounded by some constant.

In addition, denote a^2 as the element-wise square for a vector a, the variance can also be similarly calculated:

$$
\begin{aligned}
&\mathrm{Var}_{\mathbf{z}}(\frac{1}{n}\sum_{k=1}^{n}\mathrm{w}_l(\mathbf{z}_k)\Sigma^*_{-j,-j}(\mathbf{z}_k)(\beta^*_j(\mathbf{z}_k)-\beta^*_j(\mathbf{z}_l))) \\
&= \mathrm{Var}_{\mathbf{z}}\left(\frac{c_1}{n\tau}\sum_{k=1}^{n}K\left(\frac{\mathbf{z}_k-\mathbf{z}_l}{\tau}\right)\Sigma^*_{-j,-j}(\mathbf{z}_k)(\beta^*_j(\mathbf{z}_k))\right) \\
&= \frac{c_l^2}{n\tau^2}\mathbf{E}\left[\left(K\left(\frac{\mathbf{z}_k-\mathbf{z}_l}{\tau}\right)\Sigma^*_{-j,-j}(\mathbf{z}_k)\beta^*_j(\mathbf{z}_k)\right)^2\right]-\frac{1}{n^2}\{\mathbf{E}(\sum_{k=1}^{n}\mathrm{w}_l(\mathbf{z}_k)\Sigma^*_{-j,-j}(\mathbf{z}_k)(\beta^*_j(\mathbf{z}_k)))\}^2 \\
&= \frac{c_l^2}{n\tau^2}\int K\left(\frac{\mathbf{z}-\mathbf{z}_l}{\tau}\right)^2(\Sigma^*_{-j,-j}(\mathbf{z}_k)\beta^*_j(\mathbf{z}_k))^2f(\mathbf{z})dz+o(\frac{1}{n\tau}) \\
&= \frac{c_l^2}{n\tau}\int K(u)^2(\Sigma^*_{-j,-j}(\mathbf{z}_l+u\tau)\beta^*_j(\mathbf{z}_l+u\tau))^2f(\mathbf{z}_l+u\tau)du+o(\frac{1}{n\tau}) \\
&= c_l^2c_0\frac{(\Sigma^*_{-j,-j}(\mathbf{z}_l)\beta^*_j(\mathbf{z}_l))^2f(\mathbf{z}_l)}{n\tau}+o(\frac{1}{n\tau}),
\end{aligned}
$$

where we use component-wisely Taylor expansion again in the last equation. Since each component of $\|\Sigma^*_{-j,-j}(\mathbf{z}_l)\|_2$ is bounded by constant and $\beta^*_j(\mathbf{z}_l), \dot{\beta}^*_j(\mathbf{z}_l), \ddot{\beta}^*_j(\mathbf{z}_l)$ are all s^*_j sparse, we have

$\|\Sigma^*_{-j,-j}(\mathbf{z}_l)\ddot{\beta}^*_j(\mathbf{z}_l)\|_2^2 \lesssim s^*_j$, $\|\Sigma^*_{-j,-j}(\mathbf{z}_l)\dot{\beta}^*_j(\mathbf{z}_l)\|_2^2 \lesssim s^*_j$ and $\|\Sigma^*_{-j,-j}(\mathbf{z}_l)\beta^*_j(\mathbf{z}_l)\|_2^2 \lesssim s^*_j$. Therefore, the final conclusion holds by aggregating the bias and variance. □

B.2 Proof of Theorem 1

Proof. We first prove that given a single subject (the index is omitted for notation simplicity), and a single component j, we have with probability at least $1 - c_2\exp(-c_3 n) - c_4/p^{c_0+1} - \xi$,

$$\int \frac{1}{n} d_\alpha(\theta_j(\mathbf{z}), \tilde{\theta}_j(\mathbf{z}))\hat{q}_{\theta_j}(\theta_j(\mathbf{z}))d\theta_j(\mathbf{z}) \leq C\frac{\alpha}{1-\alpha}\left(\frac{s^*_j\log(np)}{n} + \frac{s^*_j}{n^{\frac{3}{5}}}\right) + \frac{\log(1/\xi)}{n(1-\alpha)}, \tag{2}$$

for positive constants $c_0, c_1, c_2, c_3, c_4, C > 0$.

Define the density function $\tilde{q}_\theta$ as the restriction of the prior $p_{\theta_j(\mathbf{z})}$ restricted in the neighborhood

$$\begin{gathered}\mathcal{N}(\tilde{\theta}_j(\mathbf{z}), \epsilon) := \{\theta_j(\mathbf{z}) = (\beta_j(\mathbf{z}), \gamma_j(\mathbf{z})) : \beta_{j,k}(\mathbf{z}) = 0, \\ \text{for } \tilde{\beta}_{j,k}(\mathbf{z}) = 0, \text{ and } |\beta_{j,k}(\mathbf{z}) - \tilde{\beta}_{j,k}(\mathbf{z})| \leq c_0\tau\epsilon/\sqrt{s^*_j} \text{ for } \tilde{\beta}_{j,k}(\mathbf{z}) \neq 0\}\end{gathered} \tag{3}$$

with $\epsilon = \sqrt{s^*_j\log(np)/n} + \sqrt{s^*_j n^{-3/5}}$ for small enough constant $c_0 > 0$ and $\tau = n^{-1/5}$. Then the measure $\tilde{q}_{\beta_j(\mathbf{z}),\gamma_j(\mathbf{z})}$ belongs to the specified variational family. The choice of ϵ is decided by the rate of the misspecified KL ball, which is upper bounded by $\|\mathrm{W}_l^{1/2}\mathbf{X}_{-j}(\tilde{\beta}^l_j(\mathbf{z}) - \beta^*_j(\mathbf{z}_l))\|_2^2$ as shown in Lemma C.2.

First, by Lemma C.1, it follows with probability at least $1 - \xi$, we have

$$\int \frac{1}{n} d_{\alpha,\tilde{\theta}_j(\mathbf{z})}(\theta_j(\mathbf{z}), \tilde{\theta}_j(\mathbf{z}))\hat{q}_{\theta_j}(\theta_j(\mathbf{z}))d\theta_j(\mathbf{z}) \leq \frac{\alpha}{n(1-\alpha)}\Psi^w(q_{\theta_j(\mathbf{z})}) + \frac{\log(1/\xi)}{n(1-\alpha)},$$

for any measure $\hat{q}_{\theta_j} \ll p_{\theta_j(\mathbf{z})}$.

Then, by Lemma C.2, we have with probability $1 - c_1\exp(-c_2 n) - c_3/p^{c_0+1}$,

$$-\int \log\left\{\frac{p^w(x_j \mid \theta_j(\mathbf{z}), \mathbf{X}_{-j}, \mathbf{z})}{p^w(x_j \mid \tilde{\theta}_j(\mathbf{z}), \mathbf{X}_{-j}, \mathbf{z})}\right\}\tilde{q}_{\theta_j}(\theta_j(\mathbf{z}))d\theta_j(\mathbf{z}) < Dn\epsilon^2.$$

Finally, by the KL divergence of restricted measure vs. original measure, we have $\mathrm{D}_{\mathrm{KL}}(\tilde{q}_{\theta_j}\|p_{\theta_j(\mathbf{z})}) = -\log(p(\theta \in \mathcal{N}(\tilde{\theta}_j(\mathbf{z}),\epsilon))) \lesssim s_j^* \log p + s_j^* \log((s_j^*)^{1/2}\tau^{-1}/\epsilon) \lesssim n\epsilon^2$. Then equation (2) holds given that $\Psi^w(\hat{q}_{\theta_j}) \leq \Psi^w(\tilde{q}_{\theta_j})$.

Given conclusion in equation (2), for each j, we choose $\xi = (np)^{-c_5 s^*}$ such that $\log(1/\xi) = c_5 s^* \log(np)$. Then by union bound for $\theta_j(\mathbf{z})$, $j = 1,...,p$, we have with probability at least, $1 - c_2 p\exp(-c_3 n) - c_4/p^{c_0} - pe^{-c_5 s^* \log(np)}$,

$$\max_{j=1,...,p} \int \frac{1}{n} d_\alpha(\theta_j(\mathbf{z}), \tilde{\theta}_j(\mathbf{z}))\hat{q}_{\theta_j}(\theta_j(\mathbf{z}))d\theta_j(\mathbf{z}) \leq C\frac{1+\alpha}{1-\alpha}\left(\frac{s^*\log(np)}{n} + \frac{s^*}{n^{\frac{3}{5}}}\right) \tag{4}$$

for positive constants $c_0, c_1, c_2, c_3, c_4, C > 0$. Finally, by the union bound applying across subject $l = 1,...,n$, we have the conclusion of the theorem. □

B.3 Proof of Lemma 2

Proof. Similarly, for any observation l, we have the marginalized KL divergence:

$$\int p(\mathbf{X} \mid \Theta^*(\mathbf{z}_1),...,\Theta^*(\mathbf{z}_n)) \log \frac{p(\mathbf{X} \mid \Theta^*(\mathbf{z}_1),...,\Theta^*(\mathbf{z}_n))}{\prod_{j=1}^p p^{w_l}(x_j \mid \mathbf{X}_{-j}, \Theta_{-j}^l(\mathbf{z}))} d\mathbf{X}$$
$$= c - \sum_{j=1}^p \int p(\mathbf{X} \mid \Theta^*(\mathbf{z}_1),...,\Theta^*(\mathbf{z}_n)) \log p^{w_l}(x_j \mid \mathbf{X}_{-j}, \Theta_{-j}^l(\mathbf{z}))d\mathbf{X}$$
$$= c - \sum_{j=1}^p \int p(\mathbf{X}_{-j} \mid \Theta_{-j,-j}^*(\mathbf{z}_1),...,\Theta_{-j,-j}^*(\mathbf{z}_n))p(x_j \mid \mathbf{X}_{-j}, \Theta_{-j}^*(\mathbf{z}_1),...,\Theta_{-j}^*(\mathbf{z}_n))$$
$$\log p^{w_l}(x_j \mid \mathbf{X}_{-j}, \Theta_{-j}^l(\mathbf{z}))dx_j d\mathbf{X}_{-j}.$$

Therefore, $\tilde{\beta}_j^l(\mathbf{z})$ should be the minimizer of the following objective function

$$\begin{aligned}
&-\int p(\mathbf{X}_{-j} \mid \Theta^*_{-j,-j}(\mathbf{z}_1), ..., \Theta^*_{-j,-j}(\mathbf{z}_n))p(x_j \mid \mathbf{X}_{-j}, \Theta^*_{-j}(\mathbf{z}_1), ..., \Theta^*_{-j}(\mathbf{z}_n)) \\
&\times \log p^{w_l}(x_j \mid \mathbf{X}_{-j}, \beta_j^l(\mathbf{z}), \gamma_j^l(\mathbf{z}))dx_j d\mathbf{X}_{-j} \\
&= \mathbf{E}_{\mathbf{X}_{-j},x_j}\left\{\sum_{k=1}^n (x_{kj} - \mathbf{x}_{k,-j}^{\mathrm{T}}\beta_j^l(\mathbf{z}))^{\mathrm{T}}\frac{\mathrm{w}_l(\mathbf{z}_k)}{2\sigma_*^2}(x_{kj} - \mathbf{x}_{k,-j}^{\mathrm{T}}\beta_j^l(\mathbf{z}))\right\} \\
&= \mathbf{E}_{\mathbf{X}_{-j}}\left\{\sum_{k=1}^n (\beta_j^*(\mathbf{z}_k) - \beta_j^l(\mathbf{z}))^{\mathrm{T}}\frac{\mathrm{w}_l(\mathbf{z}_k)}{2\sigma_*^2}\mathbf{x}_{k,-j}\mathbf{x}_{k,-j}^{\mathrm{T}}(\beta_j^*(\mathbf{z}_k) - \beta_j^l(\mathbf{z}))\right\} \\
&= \sum_{k=1}^n (\beta_j^*(\mathbf{z}_k) - \beta_j^l(\mathbf{z}))^{\mathrm{T}}\frac{\mathrm{w}_l(\mathbf{z}_k)}{2\sigma_*^2}\Sigma^*_{-j,-j}(\mathbf{z}_k)(\beta_j^*(\mathbf{z}_k) - \beta_j^l(\mathbf{z})),
\end{aligned}$$

under the constraint $\|\beta_j^l(\mathbf{z})\|_0 \leq C_0 s_j^*$ for $C_0 \geq 1$.

Since $\|\beta_j^*(\mathbf{z}_l)\|_0 \leq s_j^*$ is in the constrained region, by basic inequality, we have

$$\begin{aligned}
&\sum_{k=1}^n (\beta_j^*(\mathbf{z}_k) - \tilde{\beta}_j^l(\mathbf{z}))^{\mathrm{T}}\frac{\mathrm{w}_l(\mathbf{z}_k)}{2\sigma_*^2}\Sigma^*_{-j,-j}(\mathbf{z}_k)(\beta_j^*(\mathbf{z}_k) - \beta_j^l(\mathbf{z})) \\
\times \leq &\sum_{k=1}^n (\beta_j^*(\mathbf{z}_k) - \beta_j^*(\mathbf{z}_l))^{\mathrm{T}}\frac{\mathrm{w}_l(\mathbf{z}_k)}{2\sigma_*^2}\Sigma^*_{-j,-j}(\mathbf{z}_k)(\beta_j^*(\mathbf{z}_k) - \beta_j^*(\mathbf{z}_l)),
\end{aligned}$$

where $\Sigma^*_{-j,-j}(\mathbf{z}_k)$ is the submatrix of kth true covariance except jth row and column. After some algebra, we have

$$(\beta_j^l(\mathbf{z})^* - \tilde{\beta}_j^l(\mathbf{z}))^{\mathrm{T}}\sum_{k=1}^n \mathrm{w}_l(\mathbf{z}_k)\Sigma^*_{-j,-j}(\mathbf{z}_k)(\beta_j^l(\mathbf{z})^* - \tilde{\beta}_j^l(\mathbf{z})) \leq 2(\beta_j^l(\mathbf{z}) - \beta_j^*(\mathbf{z}_l))(\sum_{k=1}^n \mathrm{w}_l(\mathbf{z}_k)\Sigma^*_{-j,-j}(\mathbf{z}_k)\beta_j^*(\mathbf{z}_k))$$

Since the eigenvalues of $\Sigma^*_{-j,-j}(\mathbf{z}_k)$ are all lower bounded by constant, by Weyl's inequality, we have $\lambda_{\min}(\sum_{k=1}^n \mathrm{w}_l(\mathbf{z}_k)\Sigma^*_{-j,-j}(\mathbf{z}_k))$ lower bounded by constant multiplied by $\sum_{k=1}^n \mathrm{w}_k$.

Suppose that $\mathbf{z}$ takes K distinct values $z_0^1, z_0^2, \ldots, z_0^K$.

Note that

$$\left\{\sum_{k=1}^n \mathrm{w}_l(\mathbf{z}_k)\right\} = \left\{\frac{1}{\tau}\sum_{k=1}^n K\left(\frac{\mathbf{z}_k - \mathbf{z}_l}{\tau}\right)\right\}.$$

For a Gaussian kernel, the terms $K((\mathbf{z}_l - \mathbf{z}_k)/\tau)$ are bounded away from zero for $\mathbf{z}_k = \mathbf{z}_l$, and hence $\sum_{k=1}^{n} \mathrm{w}_k(\mathbf{z}_l) \geq c_0 n_l/\tau$ for some positive constant c_0. Applying the above lower eigenvalues and Cauchy-Schwartz in equality on equation (1) after taking expectation to $\mathbf{z}$, we have

$$\|\beta_j^*(\mathbf{z}_k) - \tilde{\beta}_j^l(\mathbf{z})\|_2^2 \lesssim \|\frac{\tau}{n_l}\sum_{k=1}^{n} \mathrm{w}_l(\mathbf{z}_k)\Sigma_{-j,-j}^*(\mathbf{z}_k)(\beta_j^*(\mathbf{z}_k) - \beta_j^*(\mathbf{z}_l))\|^2.$$

Now, we have

$$\frac{\tau}{n_l}\sum_{k=1}^{n} \mathrm{w}_l(\mathbf{z}_k)\Sigma_{-j,-j}^*(\mathbf{z}_k)(\beta_j^*(\mathbf{z}_k) - \beta_j^*(\mathbf{z}_l)) = \frac{1}{n_l}\sum_{k=1}^{n} K\left(\frac{\mathbf{z}_k - \mathbf{z}_l}{\tau}\right)\Sigma_{k,-j,-j}^*(\beta_j^*(\mathbf{z}_k) - \beta_j^*(\mathbf{z}_l)).$$

Let $c_l = \min_{k:z_k \neq z_l} |\mathbf{z}_k - \mathbf{z}_l|$. Then we have, using the fact that the kernel is Gaussian,

$$\begin{aligned}\frac{\tau}{n_l}\sum_{k=1}^{n} \mathrm{w}_l(\mathbf{z}_k)\Sigma_{-j,-j}^*(\mathbf{z}_k)(\beta_j^*(\mathbf{z}_k) - \beta_j^*(\mathbf{z}_l)) &\leq \frac{c}{n_l}\sum_{l:\mathbf{z}_k \neq \mathbf{z}_l} K\left(\frac{\mathbf{z}_l - \mathbf{z}_k}{\tau}\right)\Sigma_{-j,-j}^*(\mathbf{z}_k)(\beta_j^*(\mathbf{z}_k) - \beta_j^*(\mathbf{z}_l)) \\ &\leq \frac{c}{n_l}\exp\left(-c_l^2/\tau^2\right)\sum_{l:\mathbf{z}_k \neq \mathbf{z}_l} \Sigma_{-j,-j}^*(\mathbf{z}_k)(\beta_j^*(\mathbf{z}_k) - \beta_j^*(\mathbf{z}_l)).\end{aligned}$$

where c is a positive constant that changes between steps but does not affect the overall rate. Also, note that $\|\Sigma_{-j,-j}^*(\mathbf{z}_l)\|_2$ is upper bounded by some constant. Therefore, given the sparsity of $\beta_j^*(\mathbf{z}_k) - \beta_j^*(\mathbf{z}_l)$ and bounded eigenvalues of $\Sigma_{-j,-j}^*(\mathbf{z}_k)$, we have the ℓ_2 norm of right hand side of the above inequality is bounded by $c(n - n_l)\sqrt{s_j^*}$, therefore

$$\|\tilde{\beta}_j^l(\mathbf{z}) - \beta_j^*(\mathbf{z}_l)\|_2^2 \leq c\exp\left(-2c_l^2/\tau^2 + 2\log(n/n_l - 1)\right)s_j^*,$$

which converges to zero faster than s_j^*/n as long as $c_l^2/\tau^2 > \log(n)$.

□

B.4 Proof of Lemma 3

Proof. Based on the proof of Lemma 2, we have $\tilde{\beta}_j^l(\mathbf{z})$ should be the minimizer of the following objective function

$$\sum_{k=1}^{n}(\beta_j^*(\mathbf{z}_k) - \beta_j^l(\mathbf{z}))^{\mathrm{T}}\frac{\mathrm{w}_l(\mathbf{z}_k)}{2\sigma_*^2}\Sigma^*_{-j,-j}(\mathbf{z}_k)(\beta_j^*(\mathbf{z}_k) - \beta_j^l(\mathbf{z})).$$

Note that under the homogeneous assumption $\beta^{1*} = \beta^{2*} = ...\beta^{n*} = \beta^*$, the objective function becomes

$$(\beta_j{}^* - \beta_j^l(\mathbf{z}))^{\mathrm{T}}\frac{n}{2\sigma_*^2}\Sigma^*_{-j,-j}(\beta_j{}^* - \beta_j^l(\mathbf{z})).$$

Given that $\Sigma^*_{-j,-j}$ is positive definite, the Kullback-Leibler minimizer satisfies $\tilde{\beta}_j = \beta_j^*$. □

B.5 Proof of Theorem 3

We have,

$$\mathbf{E}_{-j}\,\mathbf{E}_j\exp\left\{\alpha\frac{p^w(x_j \mid \mathbf{X}_{-j}, \tilde{\theta}_j(\mathbf{z}))}{p^w(x_j \mid \mathbf{X}_{-j}, \theta_j(\mathbf{z}))}\right\} = \exp\left\{-(1-\alpha)d_{\alpha,\theta_j^*}(\theta_j(\mathbf{z}), \tilde{\theta}_j(\mathbf{z}) \mid \mathbf{X}_{-j})\right\}.$$

Following the steps of Lemma C.4, we have

$$\begin{aligned}\mathbf{E}_{-j}\Bigg[P\bigg(\int(1-\alpha)d_{\alpha,\theta_j^*}(\theta_j(\mathbf{z}), \tilde{\theta}_j(\mathbf{z}))\hat{q}_{\theta_j}(\theta_j(\mathbf{z}))d\theta_j(\mathbf{z}) \leq & -\alpha\int\log\frac{p^w(x_j \mid \tilde{\theta}_j(\mathbf{z}), \mathbf{X}_{-j})}{p^w(x_j \mid \theta_j(\mathbf{z}), \mathbf{X}_{-j})}\hat{q}_{\theta_j}(\theta_j(\mathbf{z}))d\theta_j(\mathbf{z}) \\ & + \mathrm{D}_{\mathrm{KL}}(\hat{q}_{\theta_j}\|p_{\theta_j}) + \log\left(\frac{1}{\zeta}\right)\bigg)\Bigg] \geq 1-\zeta.\end{aligned} \tag{5}$$

Define a specific Kullback-Leibler ball around $\tilde{\theta}_j(\mathbf{z})$ as

$$\mathcal{B}_{n,\theta_j^*}(\tilde{\theta}_j, \epsilon, \mathbf{X}_{-j}) = \Bigg\{\theta_j : \int \log \frac{p^w(x_j \mid \mathbf{X}_{-j}, \tilde{\theta}_j(\mathbf{z}))}{p^w(x_j \mid \mathbf{X}_{-j}, \theta_j)} p(x_j \mid \mathbf{X}_{-j}, \theta_j^*) dx_j \leq n\epsilon^2, \\ \int \log^2 \frac{p^w(x_j \mid \mathbf{X}_{-j}, \tilde{\theta}_j(\mathbf{z}))}{p^w(x_j \mid \mathbf{X}_{-j}, \theta_j)} p(x_j \mid \mathbf{X}_{-j}, \theta_j^*) dx_j \leq n\epsilon^2 \Bigg\}.$$

Next, define the set

$$\mathcal{A}^w(\mathbf{X}_{-j}, \epsilon) = \Bigg\{x_j : \int \log \frac{p^w(x_j \mid \tilde{\theta}_j(\mathbf{z}), \mathbf{X}_{-j}, \mathbf{z})}{p^w(x_j \mid \theta_j(\mathbf{z}), \mathbf{X}_{-j}, \mathbf{z})} q_{\theta_j}(\theta_j(\mathbf{z})) d\theta_j(\mathbf{z}) \leq Dn\epsilon^2 \Bigg\}$$

for some positive constant D. Following the steps of Lemma C.5, we have

$$P(x_j \notin \mathcal{A}^w(\mathbf{X}_{-j}, \epsilon)) \leq \frac{c_2}{p^{c_1+1}}.$$

Let the precision matrix of the p-variate data generating distribution be s^*-sparse and have eigenvalues bounded away from 0 and ∞. Then, it follows that for any $\zeta \in (0,1)$ and $n \geq a_1 s^* \log p$, and any measure $q_{\theta_j} \in \Gamma$ such that $q_{\theta_j} \ll p_{\theta_j}$, we have

$$P\Bigg(\int \frac{1}{n} d_{\alpha,\theta_j^*}(\theta_j(\mathbf{z}), \tilde{\theta}_j(\mathbf{z})) \hat{q}_{\theta_j}(\theta_j(\mathbf{z})) d\theta_j(\mathbf{z}) \leq \frac{\alpha}{n(1-\alpha)} \Psi(q_{\theta_j}) \\ + \frac{1}{n(1-\alpha)} \log(1/\zeta) \Bigg) \geq 1 - \zeta - \frac{c_2}{p^{c_1+1}} - \exp\{-a_2 n\}.$$

for some positive constants a_1, D and a_2. In the covariate-independent setup, since $\theta_j^* = \tilde{\theta}_j(\mathbf{z})$, the variational estimate $\hat{q}_{\theta_j}(\theta_j(\mathbf{z}))$ assumes the following form:

$$\hat{q}_{\theta_j}(\theta_j(\mathbf{z})) = \underset{q_{\beta_j,\gamma_j} = \prod_{k=1}^p q_{\beta_{jk},\gamma_{jk}}}{\operatorname{argmin}} \Bigg\{ - \int \sum_{\delta \in \{0,1\}^{p-1}} \log \frac{p^w(x_j \mid \beta_j(\mathbf{z}), \gamma_j(\mathbf{z}), \mathbf{X}_{-j})}{p^w(x_j \mid \beta_j^*, \gamma_j^*, \mathbf{X}_{-j})} q_{\theta_j}(\theta_j(\mathbf{z})) d\theta_j(\mathbf{z}) + \\ \alpha^{-1} \mathrm{D}_{\mathrm{KL}}(q_{\theta_j}(\mathbf{z}) \| p_{\theta_j}) \Bigg\}.$$

Since we have $x_j \in \mathcal{A}(\mathbf{X}_{-j}, \epsilon)$, we have $\int \log \left\{ p(x_j \mid \theta_j(\mathbf{z}), \mathbf{X}_{-j}) / p(x_j \mid \theta_j^*, \mathbf{X}_{-j}) \right\} q_{\theta_j}(\theta_j(\mathbf{z})) d\theta_j(\mathbf{z}) < Dn\epsilon^2$.

Define the density function $q^*_{\beta_j, \gamma_j}$ as the restriction of the prior in the neighborhood $\mathcal{N}_n(\theta_j^*, \epsilon) = \{\theta_j = (\beta_j, \gamma_j) : \beta_{j, \gamma_j^* = 0} = 0; |\beta_{j,\gamma_j} - \beta^*_{j,\gamma_j}| < c_0 \tau \epsilon / \sqrt{s_j^*}, \text{for } \gamma_j^* \neq 0\}$, where c_0 is a sufficiently small constant. Then the measure $q_{\beta_j^*, \gamma_j^*}$ belongs to the variational family. Following the steps of the proof of Corollary 1, we have the statement of Theorem 3.

B.6 Proof of Corollary 1

When we model the covariate levels independently, the covariate values themselves have no effect on the analysis. This scenario results in Kullback-Leibler balls around the true parameter since the models are well-specified, corresponding to the weights being one for all observations. That is, $p(x_j \mid \theta_j(\mathbf{z}), \mathbf{X}_{-j})$ corresponds to (9) with W as the identity matrix. Consider the following term

$$\mathbf{E}_{\theta_j^*} \left[\exp \left(\alpha \log \frac{p(\mathbf{X} \mid \theta_j(\mathbf{z}))}{p(\mathbf{X} \mid \theta_j^*)} \right) \right] = \sum_{j=1}^{n} \mathbf{E}_{-j} \left[\mathbf{E}_j \left\{ \exp \left(\alpha \log \frac{p(x_j \mid \theta_j(\mathbf{z}))}{p(x_j \mid \theta_j^*)} \right) \Bigm| \mathbf{X}_{-j} \right\} \right]. \qquad (6)$$

Note that the expectation on the left hand side of (6) is with respect to the original data distribution (multivariate Gaussian), whereas the expression within the j-th expectation is with respect to the conditional distribution $p(x_j \mid \mathbf{X}_{-j}, \theta_j^*)$. We focus on the j-th term on the right hand side, given $\mathbf{X}_{-j}$. Thus,

$$\mathbf{E}_j \left[\exp \left(\alpha \log \frac{p(x_j \mid \theta_j(\mathbf{z}))}{p(x_j \mid \theta_j^*)} \right) \Bigm| \mathbf{X}_{-j} \right] = \exp \left\{ -(1-\alpha) \mathrm{D}_\alpha(\theta_j(\mathbf{z}), \theta_j^* | \mathbf{X}_{-j}) \right\}.$$

Next we have, for well-specified models,

$$\Psi(q_{\theta_j}(\mathbf{z})) = -\int \log \frac{p(x_j \mid \theta_j(\mathbf{z}), \mathbf{X}_{-j})}{p(x_j \mid \theta_j^*, \mathbf{X}_{-j})} q_{\theta_j}(\theta_j(\mathbf{z})) d\theta_j(\mathbf{z}) + \alpha^{-1} \mathrm{D}_{\mathrm{KL}}(q_{\theta_j}(\mathbf{z}) \| p_{\theta_j}). \qquad (7)$$

Then by Lemma C.4, define an ϵ-ball around the true parameter θ_j as

$$\mathcal{B}_n(\theta_j^*, \epsilon, \mathbf{X}_{-j}) = \big\{\theta_j : \mathrm{D}_{\mathrm{KL}}(p(x_j \mid \mathbf{X}_{-j}, \theta_j^*) \| p(x_j \mid \mathbf{X}_{-j}, \theta_j)) \leq n\epsilon^2,$$
$$V(p(x_j \mid \mathbf{X}_{-j}, \theta_j^*) \| p(x_j \mid \mathbf{X}_{-j}, \theta_j)) \leq n\epsilon^2\big\}.$$

Here $V(p \,\|\, q) = \int p \log^2(p/q) dx$ is a discrepency measure called V-divergence. Next, define the following set $\mathcal{A}(\mathbf{X}_{-j}, \epsilon)$ as

$$\mathcal{A}(\mathbf{X}_{-j}, \epsilon) = \left\{ x_j : \int \hat{q}_{\theta_j}(\theta_j(\mathbf{z})) \log \frac{p(x_j \mid \theta_j(\mathbf{z}), \mathbf{X}_{-j})}{p(x_j \mid \theta_j^*, \mathbf{X}_{-j})} d\theta_j(\mathbf{z}) \leq Dn\epsilon^2 \right\}$$

for some positive constant $D > 1$.

Next, define the set $\mathcal{C}(\mathbf{X}) = \{\mathbf{X} : \tilde{\mathrm{A}}(\mathbf{X}_{-j}, x_j) \leq 0, x_j \in \mathcal{A}(\mathbf{X}_{-j}, \epsilon), \mathbf{X}_{-j} \in \mathcal{G}_{n,p-1})\}$. Consider the current problem of high dimensional Bayesian linear regression with spike-and-slab priors for the coefficient parameters. Using the mean-field variational family for the parameter $\theta_j = (\beta_j, \gamma_j)$, we have the following form for $q_{\theta_j}(\theta_j(\mathbf{z}))$.

$$q_{\theta_j}(\theta_j(\mathbf{z})) = \prod_{k=1}^{p-1} q_{\beta_{jk}, \gamma_{jk}}(\beta_{jk}(\mathbf{z}), \gamma_{jk}(\mathbf{z}))$$

Now, consider the term $\Psi(q_{\theta_j})$ as in (7). It is combination of a model fit term and a regularization term. In the current setup, the variational estimate $\hat{q}_{\theta_j}(\theta_j)$ assumes the following form:

$$\hat{q}_{\theta_j}(\theta_j) = \underset{q_{\beta_j,\gamma_j} = \prod_{k=1}^{p} q_{\beta_{jk},\gamma_{jk}}}{\operatorname{argmin}} \left\{ -\int \sum_{\delta \in \{0,1\}^{p-1}} \log \frac{p(x_j \mid \beta_j(\mathbf{z}), \gamma_j(\mathbf{z}), \mathbf{X}_{-j})}{p(x_j \mid \beta_j^*, \gamma_j^*, \mathbf{X}_{-j})} q_{\theta_j}(\theta_j(\mathbf{z})) d\theta_j(\mathbf{z}) + \alpha^{-1} \mathrm{D}_{\mathrm{KL}}(q_{\theta_j} \| p_{\theta_j}) \right\}.$$

In the set $\mathcal{C}(\mathbf{X})$, we have $x_j \in \mathcal{A}(\mathbf{X}_{-j}, \epsilon)$. Therefore, $\int \log \left\{ p(x_j \mid \theta_j(\mathbf{z}), \mathbf{X}_{-j}) / p(x_j \mid \theta_j^*, \mathbf{X}_{-j}) \right\} q_{\theta_j}(\theta_j(\mathbf{z})) d\theta_j(\mathbf{z})$ $Dn\epsilon^2$.

Define the density function $q^*_{\beta_j,\gamma_j}$ as the restriction of the prior in the neighborhood $\mathcal{N}_n(\theta_j^*, \epsilon) = \{\theta_j = (\beta_j, \gamma_j) : \beta_{j,\gamma_j^*=0} = 0; |\beta_{j,\gamma_j} - \beta^*_{j,\gamma_j}| < c_0 \tau \epsilon / \sqrt{s_j^*}, \text{for}\, \gamma_j^* \neq 0\}$, where c_0 is a sufficiently small constant. Then the measure q_{β_j,γ_j^*} belongs to the variational family.

If $\mathbf{X}_{-j} \in \mathcal{G}_{n,p-1}$, we have

$$c_2\underset{\sim}{\kappa}(s^*, \mathbf{X}_{-j})(\beta_j^* - \beta_j(\mathbf{z}))^{\mathrm{T}}(\beta_j^* - \beta_j(\mathbf{z})) \leq (\beta_j^* - \beta_j(\mathbf{z}))^{\mathrm{T}}\mathbf{X}_{-j}^{\mathrm{T}}\mathbf{X}_{-j}(\beta_j^* - \beta_j(\mathbf{z}))$$
$$\leq c_1\tilde{\kappa}(s^*, \mathbf{X}_{-j})(\beta_j^* - \beta_j(\mathbf{z}))^{\mathrm{T}}(\beta_j^* - \beta_j(\mathbf{z})).$$

Then, $\mathcal{N}_n(\theta_j^*, \epsilon) \subset \mathcal{B}_n(\theta_j^*, \epsilon, \mathbf{X}_{-j})$. Note that since $\mathbf{X}_{-j} \in \mathcal{G}_{n,p-1}$, by the volume of the neighborhood $\mathcal{N}_n(\theta_j^*, \epsilon)$, we have $\mathrm{D}_{\mathrm{KL}}(q_{\theta_j} \| p_{\theta_j}) < -\log p_{\theta_j}[\mathcal{N}_n(\theta_j^*, \epsilon)]/n(1-\alpha) < \frac{s_j^*}{n(1-\alpha)}\log(s_j^*/\epsilon) + \log p_{\gamma_j}(\gamma_j^*)$ where s_j^* is the number of non-zero entries. Since, $-\log p_{\gamma_j}(\gamma_j^*) \leq s_j^* \log p$, based on Lemma C.5, it follows that with probability at least $1 - \zeta - \frac{c_2}{p^{c_1+1}} - \exp\{-a_2 n\}$, for some positive constants a_1, a_2 and D,

$$\int \frac{1}{n} d_\alpha(\theta_j(\mathbf{z}), \theta_j^*)\hat{q}_{\theta_j}(\theta_j(\mathbf{z}))d\theta_j(\mathbf{z}) \leq \frac{\alpha\epsilon^2}{(1-\alpha)} + \frac{s_j^*}{n(1-\alpha)}\log(\frac{s_j^*}{\epsilon}) + \frac{1}{n(1-\alpha)}s_j^*\log p + \frac{1}{n(1-\alpha)}\log\left(\frac{1}{\zeta}\right).$$

for all $j \in \{1, 2, \ldots, p\}$. The statement of the Corollary follows from noting that $d_\alpha(\Theta(\mathbf{z}), \Theta^*) = \max\limits_j d_\alpha(\theta_j(\mathbf{z}), \theta_j^*)$, and that $\hat{q}_\Theta(\Theta(\mathbf{z})) = \prod_{j=1}^p \hat{q}_{\theta_j}(\theta_j(\mathbf{z}))$, and replacing n with n_l.

C Auxiliary results

Lemma C.1. *Under Assumptions in Theorem 1, for any variational estimate $\hat{q}_{\theta_j}(\theta_j(\mathbf{z}))$ such that $\hat{q}_{\theta_j} \ll p_{\theta_j}$, we have*

$$P\bigg(\int (1-\alpha)d_\alpha(\theta_j(\mathbf{z}), \tilde{\theta}_j(\mathbf{z}))\hat{q}_{\theta_j}(\theta_j(\mathbf{z}))d\theta_j(\mathbf{z}) \leq -\alpha\int \log\frac{p^w(x_j \mid \theta_j(\mathbf{z}), \mathbf{X}_{-j}, \mathbf{z})}{p^w(x_j \mid \tilde{\theta}_j(\mathbf{z}), \mathbf{X}_{-j}, \mathbf{z})}\hat{q}_{\theta_j}(\theta_j(\mathbf{z}))d\theta_j(\mathbf{z}) +$$
$$\mathrm{D}_{\mathrm{KL}}(\hat{q}_{\theta_j} \| p_{\theta_j}) + \log\left(\frac{1}{\zeta}\right)\bigg) \geq 1 - \zeta.$$

Proof. First

$$\mathbf{E}_{\mathbf{z}}\, \mathbf{E}_{-j}\, \mathbf{E}_j \exp\left\{\alpha\log\frac{p^w(x_j \mid \theta_j(\mathbf{z}), \mathbf{X}_{-j}, \mathbf{z})}{p^w(x_j \mid \tilde{\theta}_j(\mathbf{z}), \mathbf{X}_{-j}, \mathbf{z})}\right\} = \exp\left\{-(1-\alpha)d_\alpha(\theta_j(\mathbf{z}), \tilde{\theta}_j(\mathbf{z}))\right\}.$$

Thus, for any $\zeta \in (0,1)$, we have

$$\mathbf{E}_{\mathbf{z}}\, \mathbf{E}_{-j}\, \mathbf{E}_j \left[\exp\left\{\alpha \log \frac{p^w(x_j \mid \theta_j(\mathbf{z}), \mathbf{X}_{-j}, \mathbf{z})}{p^w(x_j \mid \tilde{\theta}_j(\mathbf{z}), \mathbf{X}_{-j}, \mathbf{z})} + (1-\alpha) d_\alpha(\theta_j(\mathbf{z}), \tilde{\theta}_j(\mathbf{z})) - \log(1/\zeta)\right\}\right] \leq \zeta.$$

Integrating both sides of this inequality with respect to the prior distribution p_{θ_j} and interchanging the integrals using Fubini's theorem, we have

$$\mathbf{E}_{\mathbf{z}}\, \mathbf{E}_{-j}\, \mathbf{E}_j \int \exp\left\{\alpha \log \frac{p^w(x_j \mid \theta_j(\mathbf{z}), \mathbf{X}_{-j}, \mathbf{z})}{p^w(x_j \mid \tilde{\theta}_j(\mathbf{z}), \mathbf{X}_{-j}, \mathbf{z})} + (1-\alpha) d_\alpha(\theta_j(\mathbf{z}), \tilde{\theta}_j(\mathbf{z})) - \log(1/\zeta)\right\} p_{\theta_j}(\theta_j(\mathbf{z})) d\theta_j(\mathbf{z}) \leq \zeta.$$

Next we use the variational duality of the KL divergence. If μ is a probability measure and h is a measurable function such that $e^h \in L_1(\mu)$, then

$$\log \int e^h d\mu = \sup_{\rho \ll \mu} \left[\int h d\rho - \mathrm{D}_{\mathrm{KL}}(\rho \| \mu)\right]. \tag{8}$$

We set $h = \alpha \log \frac{p^w(x_j|\theta_j(\mathbf{z}), \mathbf{X}_{-j}, \mathbf{z})}{p^w(x_j|\tilde{\theta}_j(\mathbf{z}), \mathbf{X}_{-j}, \mathbf{z})} + (1-\alpha) d_\alpha(\theta_j(\mathbf{z}), \tilde{\theta}_j(\mathbf{z})) - \log(1/\zeta)$ and $\rho = \hat{q}_{\theta_j}(\theta_j(\mathbf{z}))$ in the above result where $\hat{q}_{\theta_j}(\theta_j)$ is the variational estimate of the fractional posterior distribution.

$$\begin{aligned}\mathbf{E}_{\mathbf{z}}\, \mathbf{E}_{-j}\, \mathbf{E}_j \exp\Bigg[\int \Bigg\{&\alpha \log \frac{p^w(x_j \mid \theta_j(\mathbf{z}), \mathbf{X}_{-j}, \mathbf{z})}{p^w(x_j \mid \tilde{\theta}_j(\mathbf{z}), \mathbf{X}_{-j}, \mathbf{z})} + (1-\alpha) d_\alpha(\theta_j(\mathbf{z}), \tilde{\theta}_j(\mathbf{z})). \\ &- \log(1/\zeta)\Bigg\} \hat{q}_{\theta_j}(\theta_j(\mathbf{z})) d\theta_j(\mathbf{z}) - \mathrm{D}_{\mathrm{KL}}(\hat{q}_{\theta_j} \| p_{\theta_j})\Bigg] \leq \zeta.\end{aligned}$$

Let

$$\tilde{\mathrm{A}}(\mathbf{X}_{-j}, \mathbf{z}) = \int \left\{\alpha \log \frac{p^w(x_j \mid \theta_j(\mathbf{z}), \mathbf{X}_{-j}, \mathbf{z})}{p^w(x_j \mid \tilde{\theta}_j(\mathbf{z}), \mathbf{X}_{-j}, \mathbf{z})} + (1-\alpha) d_\alpha(\theta_j(\mathbf{z}), \tilde{\theta}_j(\mathbf{z})) - \log \frac{1}{\zeta}\right\} \hat{q}_{\theta_j}(\theta_j(\mathbf{z})) d\theta_j - \mathrm{D}_{\mathrm{KL}}(\hat{q}_{\theta_j} \| p_{\theta_j}) \tag{9}$$

and

$$\mathrm{A}(\mathbf{X}_{-j}, \mathbf{z}) = \left\{x_j, \mathbf{z} : \tilde{\mathrm{A}}(\mathbf{X}_{-j}, \mathbf{z}) \leq 0\right\}. \tag{10}$$

Now we apply Markov's inequality to get a probability statement. Thus, the required statement follows from the following:

$$\mathbf{E}_{\mathbf{z}}\,\mathbf{E}_{-j}\,P\left[\exp\left\{\tilde{\mathrm{A}}(\mathbf{X}_{-j},\mathbf{z})\right\} > 1 \mid \mathbf{X}_{-j},\mathbf{z}\right] \leq \mathbf{E}_{\mathbf{z}}\,\mathbf{E}_{-j}\,\mathbf{E}_{j}\left[\exp\left\{\tilde{\mathrm{A}}(\mathbf{X}_{-j},\mathbf{z})\right\} \mid \mathbf{X}_{-j},\mathbf{z}\right] \leq \zeta$$

$$\mathbf{E}_{\mathbf{z}}\,\mathbf{E}_{-j}\left\{P(\mathrm{A}(\mathbf{X}_{-j},\mathbf{z}) \mid \mathbf{X}_{-j},\mathbf{z})\right\} = \int_{\mathbb{1}_{A(\mathbf{X}_{-j},\mathbf{z})}} f(x_j \mid \mathbf{X}_{-j},\mathbf{z})f(\mathbf{X}_{-j},\mathbf{z})dx_j d\mathbf{X}_{-j}dz \geq 1-\zeta,$$

which implies the conclusion. □

Lemma C.2. *Under Assumptions in Theorem 1. Suppose $\mathcal{N}(\tilde{\theta}_j,\epsilon)$ is defined in equation* (3) *and $\tilde{q}_{\theta_j}(\theta_j(\mathbf{z}))$ is a density restricted in $\mathcal{N}(\tilde{\theta}_j(\mathbf{z}),\epsilon)$, then we have*

$$P\left(-\int \log\left\{\frac{p^w(x_j \mid \tilde{\theta}_j(\mathbf{z}),\mathbf{X}_{-j},\mathbf{z})}{p^w(x_j \mid \tilde{\theta}_j(\mathbf{z}),\mathbf{X}_{-j},\mathbf{z})}\right\} q_{\tilde{\theta}_j}(\theta_j(\mathbf{z}))d\theta_j < Dn\epsilon^2\right) \geq 1 - c_1e^{-c_2n} - \frac{c_3}{p^{c_1+1}}.$$

for some positive constants c_1, c_2, c_3, D.

Proof. By the log-likelihood of Gaussian distribution, for some constant $c, C > 0$, we have

$$\log\left\{\frac{p^w(x_j \mid \tilde{\theta}_j(\mathbf{z}),\mathbf{X}_{-j},\mathbf{z})}{p^w(x_j \mid \theta_j(\mathbf{z}),\mathbf{X}_{-j},\mathbf{z})}\right\} = C(\|\mathrm{W}_l^{1/2}x_j - \mathrm{W}_l^{1/2}\mathbf{X}_{-j}\beta_j(\mathbf{z})\|^2 - \|x_j - \mathrm{W}_l^{1/2}\mathbf{X}_{-j}\tilde{\beta}_j(\mathbf{z})\|_2^2) + c.$$

Denote $x_j = \mathbf{X}_{-j}\beta_j^*(\mathbf{z}) + \varepsilon_j$. Note that

$$\begin{aligned}
&\|\mathrm{W}_l^{1/2}x_j - \mathrm{W}_l^{1/2}\mathbf{X}_{-j}\beta_j(\mathbf{z})\|^2 - \|\mathrm{W}_l^{1/2}x_j - \mathrm{W}_l^{1/2}\mathbf{X}_{-j}\tilde{\beta}_j(\mathbf{z})\|_2^2 \\
&= \|\mathrm{W}_l^{1/2}\mathbf{X}_{-j}\tilde{\beta}_j(\mathbf{z}) - \mathrm{W}_l^{1/2}\mathbf{X}_{-j}\beta_j(\mathbf{z})\|^2 + 2\langle \mathrm{W}_l^{1/2}\mathbf{X}_{-j}(\tilde{\beta}_j(\mathbf{z}) - \beta_j(\mathbf{z})), \mathrm{W}_l^{1/2}\mathbf{X}_{-j}\beta_j^*(\mathbf{z}_l) + \mathrm{W}_l^{1/2}\varepsilon_j - \mathrm{W}_l^{1/2}\mathbf{X}_{-j}\tilde{\beta} \\
&\leq 2\|\mathrm{W}_l^{1/2}\mathbf{X}_{-j}\tilde{\beta}_j(\mathbf{z}) - \mathrm{W}_l^{1/2}\mathbf{X}_{-j}\beta_j(\mathbf{z})\|^2 + \|\mathrm{W}_l^{1/2}\mathbf{X}_{-j}\tilde{\beta}_j(\mathbf{z}) - \mathrm{W}_l^{1/2}\mathbf{X}_{-j}\beta_j^*(\mathbf{z}_l)\|^2 + 2\langle \mathrm{W}_l\mathbf{X}_{-j}(\tilde{\beta}_j(\mathbf{z}) - \beta_j(\mathbf{z})), \varepsilon
\end{aligned}$$

Under Lemma C.3, with probability at least $1 - \exp(-a_1n)$, we have $\mathbf{X} \notin \mathcal{G}_{n,p}$, this gives us

$$\|\mathrm{W}_l^{1/2}\mathbf{X}_{-j}\tilde{\beta}_j(\mathbf{z}) - \mathrm{W}_l^{1/2}\mathbf{X}_{-j}\beta_j^*(\mathbf{z}_l)\|^2 \leq cn\|\mathrm{W}\|_2\|\tilde{\beta}_j(\mathbf{z}) - \beta_j^*(\mathbf{z}_l)\|_2^2 \leq cn\tau^{-1}n^{-4/5}s_j^* \leq cs_j^*n^{2/5} \leq cn\epsilon^2.$$

In addition, by definition of $\tilde{q}_{\theta_j}(\theta_j(\mathbf{z}))$, we have

$$2\|\mathrm{W}_l^{1/2}\mathbf{X}_{-j}\tilde{\beta}_j(\mathbf{z}) - \mathrm{W}_l^{1/2}\mathbf{X}_{-j}\beta_j(\mathbf{z})\|^2 \leq cn\tau^{-1}\|\tilde{\beta}_j(\mathbf{z}) - \beta_j^*(\mathbf{z}_l)\|_2^2 \leq cn\epsilon^2.$$

For the last term $2\langle \mathrm{W}_l \mathbf{X}_{-j}(\tilde{\beta}_j(\mathbf{z}) - \beta_j(\mathbf{z})), \varepsilon_j \rangle$, first we have

$$2\langle \mathrm{W}_l \mathbf{X}_{-j}(\tilde{\beta}_j(\mathbf{z}) - \beta_j(\mathbf{z})), \varepsilon_j \rangle \leq 2\|\tilde{\beta}_j(\mathbf{z}) - \beta_j(\mathbf{z})\|_1 \|\mathbf{X}_{-j}^T \mathrm{W}_l \varepsilon_j\|_\infty \leq c\sqrt{s_j^*}\tau\epsilon\|\mathbf{X}_{-j}^T \mathrm{W}_l \varepsilon_j\|_\infty.$$

Note that $\mathbf{X}_{-j}^T \mathrm{W}_l \varepsilon_j$ is a $p-1$ dimensional Gaussian vector, with and with probability great than $1 - \exp(-a_2 n)$ by Lemma C.3, the scale of each component of the Gaussian vector is bounded by $\sigma^*\sqrt{n}/\tau$ multiplied constant, by maximal inequality of Gaussian random vector, we have

$$P(\|\mathbf{X}_{-j}^T \mathrm{W}_l \varepsilon_j\|_\infty \geq t) \leq e^{-\frac{\tau^2 t^2}{2\sigma^{*2} n} + \log p},$$

and we can choose $t = \sqrt{(c_0+2)n\log p}/\tau$ for a constant $c_0 > 0$, then the probability upper bound becomes $p^{-(c_0+1)}$. Therefore, we have

$$2\langle \mathrm{W}_l \mathbf{X}_{-j}(\tilde{\beta}_j(\mathbf{z}) - \beta_j(\mathbf{z})), \varepsilon_j \rangle \leq c\sqrt{s_j^*}\tau\epsilon\sqrt{n\log p}/\tau \leq cn\epsilon^2 + s_j^* \log p \leq cn\epsilon^2,$$

where the in the second inequality we use $2ab \leq a^2 + b^2$. □

Lemma C.3. *Under* **Assumption T1, T3, T4, T5**, *we have* $P(\mathbf{X} \notin \mathcal{G}_{n,p}) \leq \exp\{-a_1 n\}$. *In addition, with probability at least* $1 - \exp(-a_2 n)$, *we also have the maximal of* ℓ_2 *norm of column vectors of* $\mathbf{X}$ *satisfies* $\max_{i=1,\ldots,p} \|X_i\|_2 \leq a_3\sqrt{n}$ *for some constant* $a_3 > 0$.

Proof. In order to bound $P(\mathbf{X} \notin \mathcal{G}_{n,p})$ we use the Theorem 1.6 in Zhou (2009): since the covariance function $\Sigma(\mathbf{z}_i)$ is homogeneous and $\mathbf{z}_i$, $i = 1, ..., n$ are i.i.d. samples from $f(\mathbf{z})$. Note that $x_1(z_1), ..., x_n(z_n)$ are i.i.d samples form distribution $\int f(x \mid \Sigma(\mathbf{z}))f(\mathbf{z})$, which is assumed to be sub-Gaussian by the Assumption T5. Thus we have $P(\mathbf{X} \notin \mathcal{G}_{n,p}) \leq \exp\{-a_1 n\}$, for n larger than $a_4 s_j^* \log p$, where a_4 and a_1 are positive constants, by the similar argument with Lemma 3 in Atchadé (2019). Therefore, the following restricted eigenvalue conditions hold: $\underset{\sim}{\kappa}(2s^*, \mathbf{X}_{-j}^T\mathbf{X}_{-j}/n)$ and $\tilde{\kappa}(2s^*, \mathbf{X}_{-j}^T\mathbf{X}_{-j}/n)$ are constants.

For the second conclusion, first fix i, note that all eigenvalues of Σ are uniformly upper and lower bounded by constants. Then by Hanson-Wright inequality (Rudelson & Vershynin, 2013), we have

$$P(\|X_i\|_2 \geq c_1\sqrt{n}) \leq 2e^{-c_2 n},$$

for some constants $c_1, c_2 > 0$. Then by the union bound and $n \geq cs^* \log p$, by choosing large enough constant c_1', c_2', we have

$$P(\|X_i\|_2 \geq c_1'\sqrt{n}) \leq 2e^{-c_2' n + \log p} \leq 2e^{-c_3 n}.$$

□

Lemma C.4. *Under assumptions of Theorem 3, we have*

$$\mathbf{E}_{-j}\left[P\bigg(\int (1-\alpha) d_\alpha(\theta_j(\mathbf{z}), \theta_j^*) \hat{q}_{\theta_j}(\theta_j(\mathbf{z})) d\theta_j(\mathbf{z}) \leq -\alpha \int \log \frac{p(x_j \mid \theta_j(\mathbf{z}), \mathbf{X}_{-j})}{p(x_j \mid \theta_j^*, \mathbf{X}_{-j})} \hat{q}_{\theta_j}(\theta_j(\mathbf{z})) d\theta_j(\mathbf{z}) + \mathrm{D_{KL}}(\hat{q}_{\theta_j} \| p_{\theta_j}) + \log\left(\frac{1}{\zeta}\right)\bigg)\right] \geq 1-\zeta. \tag{11}$$

Proof. From (11), we have

$$\mathbf{E}_{-j}\,\mathbf{E}_j \exp\left\{\alpha \log \frac{p(x_j \mid \theta_j(\mathbf{z}), \mathbf{X}_{-j})}{p(x_j \mid \theta_j^*, \mathbf{X}_{-j})}\right\} = \exp\left(-(1-\alpha) d_\alpha(\theta_j(\mathbf{z}), \theta_j^* \mid \mathbf{X}_{-j})\right).$$

Thus, for any $\zeta \in (0,1)$, we have

$$\mathbf{E}_{-j}\,\mathbf{E}_j\left[\exp\left\{\alpha \log \frac{p(x_j \mid \theta_j(\mathbf{z}), \mathbf{X}_{-j})}{p(x_j \mid \theta_j^*, \mathbf{X}_{-j})} + (1-\alpha) d_\alpha(\theta_j(\mathbf{z}), \theta_j^*) - \log(1/\zeta)\right\}\right] \leq \zeta.$$

Integrating both sides of this inequality with respect to the prior distribution p_{θ_j} and interchanging the integrals using Fubini's theorem, we have

$$\mathbf{E}_{-j}\,\mathbf{E}_j \int \exp\left\{\alpha \log \frac{p(x_j \mid \theta_j(\mathbf{z}), \mathbf{X}_{-j})}{p(x_j \mid \theta_j^*, \mathbf{X}_{-j})} + (1-\alpha) d_\alpha(\theta_j(\mathbf{z}), \theta_j^*) - \log(1/\zeta)\right\} p_{\theta_j}(\theta_j(\mathbf{z})) d\theta_j(\mathbf{z}) \leq \zeta.$$

Next we use the following result from Yang et al. (2020).If μ is a probability measure and h is a measurable function such that $e^h \in L_1(\mu)$, then

$$\log \int e^h d\mu = \sup_{\rho \ll \mu} \left[\int h d\rho - D_{KL}(\rho \| \mu) \right]. \tag{12}$$

We set $h = \alpha \log \frac{p(x_j | \theta_j(\mathbf{z}), \mathbf{X}_{-j})}{p(x_j | \theta_j^*, \mathbf{X}_{-j})} + (1-\alpha) d_\alpha(\theta_j(\mathbf{z}), \theta_j^*) - \log(1/\zeta)$ and $\rho = \hat{q}_{\theta_j}(\theta_j(\mathbf{z}))$ in the above result where $\hat{q}_{\theta_j}(\theta_j(\mathbf{z}))$ is the variational estimate of the fractional posterior distribution.

Thus, we get

$$\mathbf{E}_{-j}\, \mathbf{E}_j \exp \left[\int \left\{ \alpha \log \frac{p(x_j \mid \theta_j(\mathbf{z}), \mathbf{X}_{-j})}{p(x_j \mid \theta_j^*, \mathbf{X}_{-j})} + (1-\alpha) d_\alpha(\theta_j(\mathbf{z}), \theta_j^*) - \log(1/\zeta) \right\} \hat{q}_{\theta_j}(\theta_j(\mathbf{z})) d\theta_j - D_{KL}(\hat{q}_{\theta_j} \| p_{\theta_j}) \right] \leq$$

Let

$$\tilde{A}(\mathbf{X}_{-j}) = \int \left\{ \alpha \log \frac{p(x_j \mid \theta_j(\mathbf{z}), \mathbf{X}_{-j})}{p(x_j \mid \theta_j^*, \mathbf{X}_{-j})} + (1-\alpha) d_\alpha(\theta_j(\mathbf{z}), \theta_j^*) - \log \frac{1}{\zeta} \right\} \hat{q}_{\theta_j}(\theta_j(\mathbf{z})) d\theta_j(\mathbf{z}) - D_{KL}(\hat{q}_{\theta_j} \| p_{\theta_j}), \tag{13}$$

and

$$A(\mathbf{X}_{-j}) = \left\{ x_j : \tilde{A}(\mathbf{X}_{-j}) \leq 0 \right\}. \tag{14}$$

Now we apply Markov's inequality to get a probability statement. Thus, the required statement follows from the following:

$$\begin{aligned} \mathbf{E}_{-j}\, P(\exp\left(\tilde{A}(\mathbf{X}_{-j})\right) > 1 \mid \mathbf{X}_{-j}) &\leq \mathbf{E}_{-j}\, \mathbf{E}_j(\exp\left(A(X_{-j})\right) \mid \mathbf{X}_{-j}) = \zeta \\ \mathbf{E}_{-j}\, P(A(\mathbf{X}_{-j}) \mid \mathbf{X}_{-j}) &\geq 1 - \zeta. \end{aligned}$$

□

Lemma C.5. *Under the assumptions of Theorem 3, let the precision matrix of the p-variate data generating distribution be s^*-sparse and have eigenvalues bounded away from 0 and ∞. For any*

$\zeta \in (0,1)$ *and* $n \geq a_1 s^* \log p$*, and any measure* $q_{\theta_j} \in \Gamma$ *such that* $q_{\theta_j} \ll p_{\theta_j}$*, we have*

$$P\left(\int \frac{1}{n} d_\alpha(\theta_j(\mathbf{z}), \theta_j^*) \hat{q}_{\theta_j}(\theta_j(\mathbf{z})) d\theta_j(\mathbf{z}) \leq \frac{\alpha}{n(1-\alpha)}\Psi(q_{\theta_j}) + \frac{1}{n(1-\alpha)}\log(1/\zeta)\right) \geq 1-\zeta-\frac{c_2}{p^{c_1+1}}-\exp\{-a_2 n\}.$$

for some positive constants a_1*,* D *and* a_2 *,* c_1, c_2*.*

Proof. Similar with Lemma C.2, we need to provide upper bound for

$$2\|\mathrm{W}_l^{1/2}\mathbf{X}_{-j}\tilde{\beta}_j(\mathbf{z}) - \mathrm{W}_l^{1/2}\mathbf{X}_{-j}\beta_j(\mathbf{z})\|^2 + \|\mathrm{W}_l^{1/2}\mathbf{X}_{-j}\tilde{\beta}_j(\mathbf{z}) - \mathrm{W}_l^{1/2}\mathbf{X}_{-j}\beta_j^*(\mathbf{z}_l)\|^2 + 2\langle \mathrm{W}_l\mathbf{X}_{-j}(\tilde{\beta}_j(\mathbf{z}) - \beta_j(\mathbf{z})), \varepsilon_j\rangle$$

By the similar argument with Lemma 3 in Atchadé (2019), the following restricted eigenvalue conditions hold: $\underset{\sim}{\kappa}(2s^*, \mathbf{X}_{-j}^T\mathbf{X}_{-j}/n)$ and $\tilde{\kappa}(2s^*, \mathbf{X}_{-j}^T\mathbf{X}_{-j}/n)$ are constants. Therefore, with probability at least $1-\exp(-a_1 n)$, we have $\mathbf{X} \notin \mathcal{G}_{n,p}$, this gives us

$$\|\mathrm{W}_l^{1/2}\mathbf{X}_{-j}\tilde{\beta}_j(\mathbf{z}) - \mathrm{W}_l^{1/2}\mathbf{X}_{-j}\beta_j^*(\mathbf{z}_l)\|^2 \leq cn\|\mathrm{W}\|_2\|\tilde{\beta}_j(\mathbf{z}) - \beta_j^*(\mathbf{z}_l)\|_2^2 \leq cn\tau^{-1}n^{-1}s_j^* \leq cn\epsilon^2,$$

given that $\tau^{-1} \leq c\sqrt{\log n}$.

In addition, by definition of $\tilde{q}_{\theta_j}(\theta_j(\mathbf{z}))$, similarly we have

$$2\|\mathrm{W}_l^{1/2}\mathbf{X}_{-j}\tilde{\beta}_j(\mathbf{z}) - \mathrm{W}_l^{1/2}\mathbf{X}_{-j}\beta_j(\mathbf{z})\|^2 \leq cn\tau^{-1}\|\tilde{\beta}_j(\mathbf{z}) - \beta_j^*(\mathbf{z}_l)\|_2^2 \leq cn\epsilon^2.$$

For the last term $2\langle \mathrm{W}_l\mathbf{X}_{-j}(\tilde{\beta}_j(\mathbf{z}) - \beta_j(\mathbf{z})), \varepsilon_j\rangle$, first we have

$$2\langle \mathrm{W}_l\mathbf{X}_{-j}(\tilde{\beta}_j(\mathbf{z}) - \beta_j(\mathbf{z})), \varepsilon_j\rangle \leq 2\|\tilde{\beta}_j(\mathbf{z}) - \beta_j(\mathbf{z})\|_1\|\mathbf{X}_{-j}^T\mathrm{W}_l\varepsilon_j\|_\infty \leq c\sqrt{s_j^*}\tau\epsilon\|\mathbf{X}_{-j}^T\mathrm{W}_l\varepsilon_j\|_\infty.$$

Note that $\mathbf{X}_{-j}^T\mathrm{W}_l\varepsilon_j$ is a $p-1$ dimensional Gaussian vector, with and with probability great than $1-\exp(-a_2 n)$, the scale of each component of the Gaussian vector is bounded by $\sigma^*\sqrt{n}/\tau$ multiplied constant, by maximal inequality of Gaussian random vector, we have

$$P(\|\mathbf{X}_{-j}^T\mathrm{W}_l\varepsilon_j\|_\infty \geq t) \leq e^{-\frac{\tau^2 t^2}{2\sigma^{*2}n}+\log p},$$

and we can choose $t = \sqrt{(c_0+2)n \log p}/\tau$ for a constant $c_0 > 0$, then the probability upper bound becomes $p^{-(c_0+1)}$. Therefore, we have

$$2\langle \mathrm{W}_l \mathbf{X}_{-j}(\tilde{\beta}_j(\mathbf{z}) - \beta_j(\mathbf{z})), \varepsilon_j \rangle \leq c\sqrt{s_j^*}\tau\epsilon\sqrt{n \log p}/\tau \leq cn\epsilon^2 + s_j^* \log p \leq cn\epsilon^2,$$

where the in the second inequality we use $2ab \leq a^2 + b^2$. □

D Steps of the algorithm

The following algorithm provides the steps for covariate-dependent graph estimation of $\mathbf{X} \in \mathbb{R}^{n \times p}$.

We select the bandwidth hyperparameter $\tau \in \mathbb{R}^n$ using a 2-step approach for density estimation discussed in Dasgupta et al. (2020); Abramson (1982); Van Kerm (2003). Under this approach, bandwidths are initialized using Silverman's rule of thumb, and the density is subsequently refined by updating the bandwidth values. We follow this methodology to estimate the density of $\mathbf{z}$, and use the updated bandwidths from the second step for τ.

We next fix x_j as the response and consider the task of performing n weighted spike-and-slab regressions with $\mathbf{X}_{-j}$ as predictors using the weights calculated using the bandwidth τ and the covariates. Each of these regressions requires the specification of three hyperparameters: π, σ^2, and σ_θ^2. To select the hyperparameters, we use a hybrid of model averaging and grid search. We first generate candidate grids of π, σ^2, and σ_θ^2 values. We denote the grid of π candidates by Θ_π, and the Cartesian product between the grid of σ^2 and σ_θ^2 candidates as Θ.

Next, for each $\pi \in \Theta_\pi$, we fit a spike-and-slab regression weighted with respect to individual l for each $(\sigma^2, \sigma_\theta^2) \in \Theta, l \in 1, ..., n$. We make a global selection of σ^2 and σ_θ^2 for each of the $\pi \in \Theta_\pi$ such that the sum of the ELBO across all n weighted regressions is maximized. This grid search produces $|\Theta_\pi|$ models per individual. For each of these models, we calculate a model averaging weight by taking the softmax over the ELBOs and use these to average over the variational approximations to the posterior quantities to construct the final model. Finally, to obtain the graph estimate, we symmetrize the posterior inclusion probabilities from the final model and threshold them at 0.5.

E Discrete Covariate Simulation Study

For the discrete covariate, we perform experiments in which we vary the data dimensionality, the distribution of the covariate levels, and the strength of the signal in the ground-truth precision structures. As with the continuous covariate, we perform 50 trials per experiment. We compare the performance of W-PL to mgm (Haslbeck & Waldorp, 2020), as well as to the method of Carbonetto et al. (2012) applied in a pseudo-likelihood fashion (CS). That is, we fix each variable as the response in turn and perform a variational spike-and-slab regression. We obtain the final graphs using the same symmetrization and thresholding scheme as used for W-PL. To incorporate $\mathbf{z}$, we apply this estimation procedure for each of the covariate levels independently. Because no information is shared between levels, this allows us to evaluate the impact of the weighting scheme in W-PL. We use the implementation of the variational spike-and-slab from Carbonetto et al. (2017), which employs a hybrid hyperparameter specification scheme, wherein the π candidates are averaged and σ^2 and σ_θ^2 are selected via Empirical Bayes for each of the π candidates.

In each of the experiments, we assign individuals $1, ..., n_1$ to the first level of a binary discrete covariate $\mathbf{z}_i = 1$, and the remaining individuals to the second level $\mathbf{z}_i = 2$. We refer to the structure of the sample as balanced when $n_1 = n - n_1 := n_2$, and unbalanced when $n_1 \neq n_2$.

E.1 Covariate-Independent Setting

We first consider a covariate-independent setting where the ground truth dependence structure is independent of $\mathbf{z}_i$ and set $n = 100, p = 10$. To construct the precision matrices, we first define

$$\lambda_{\mathbf{z}_i} \quad = \quad [c\mathbf{1}_4 \ \ , \ \ \mathbf{0}_{p-3}]^{\mathrm{T}}, \text{ for both } \mathbf{z}_i = 1, 2,$$

where $\mathbf{1}_4$ is a four-dimensional vector of ones, and $\mathbf{0}_{p-3}$ is a $(p-3)$-dimensional vector of zeroes. We refer to $c = 15$ as high signal, and $c = 3$ as reduced signal. We next define the precision matrix

c	n_1	n_2	Method	Sensitivity(↑)	Specificity(↑)
3	50	50	W-PL	$\mathbf{0.8517}(0.1591)$	$0.9843(0.0168)$
			mgm	$0.1417(0.1229)$	$\mathbf{0.9988}(0.0033)$
			CS	$0.2950(0.1489)$	$0.9929(0.0081)$
15	50	50	W-PL	$\mathbf{1.0000}(0.0000)$	$0.9941(0.0111)$
			mgm	$\mathbf{1.0000}(0.0000)$	$\mathbf{0.9990}(0.0031)$
			CS	$\mathbf{1.0000}(0.0000)$	$0.9955(0.0080)$
15	80	20	W-PL	$\mathbf{1.0000}(0.0000)$	$\mathbf{1.0000}(0.0000)$
			mgm	$0.9013(0.0602)$	$0.9992(0.0033)$
			CS	$0.8147(0.0215)$	$0.9980(0.0052)$

Table 1: *Results for the covariate-independent setting, presented as mean*$($*standard deviation*$)$

for the i-th individual as $\Omega_i = (\lambda_{\mathbf{z}_i}\lambda_{\mathbf{z}_i}{}^{\mathrm{T}} + 10\mathbb{I}_{p+1})$. The corresponding dependence structure we aim to estimate is

$$\mathrm{G}^* = \begin{bmatrix} \mathbb{J}_4 - \mathbb{I}_4 & \mathbf{0}_{4,p-3} \\ \mathbf{0}_{p-3,4} & \mathbf{0}_{p-3,p-3} \end{bmatrix}$$

where $\mathbb{J}_k$ is a $k \times k$ matrix where all entries are 1.

We perform experiments in the covariate-independent setting on high signal with balanced structure, reduced signal with balanced structure, and high signal with unbalanced structure ($n_1 = 80, n_2 = 20$). We present results for each of the experiments in Table 1. When the signal strength is high and the sample is balanced, all three methods correctly detect all of the edges in the ground truth structure. However, the performance of both competitors suffers under the unbalanced sample structure, particularly for CS, while W-PL correctly detects all edges. In the reduced signal setting, the differential between W-PL and the competitors grows significantly.

E.2 Covariate-Free Setting

We next examine a setting identical to the covariate-independent one, again with $n = 100, p = 10$, however, this time, assume that no information on the covariates is available. In the absence of covariate information, W-PL selects all weights to be equal to one. Thus, the graph estimates are identical for all the individuals in this setting, akin to the usual graph selection algorithms. Because

c	Method	Sensitivity(↑)	Specificity(↑)
3	W-PL	$\mathbf{0.9533}(0.1168)$	$\mathbf{0.9996}(0.0029)$
	CS	$0.9233(0.1313)$	$0.9939(0.0103)$
15	W-PL	$\mathbf{1.0000}(0.0000)$	$\mathbf{1.0000}(0.0000)$
	CS	$\mathbf{1.0000}(0.0000)$	$0.9963(0.0079)$

Table 2: *Results for covariate-free setting*

mgm requires the timepoints to be specified, we omit it from these experiments.

We present the results for experiments in the covariate-free setting with high and low signal in Table 2. Unsurprisingly, results for both methods are similar. The minor differences in performance may be attributed to the differing hyperparameter specification schemes.

E.3 Covariate-Dependent Setting

We next consider the setting in which the precision matrix varies with the covariate level. We define the relationship as

$$\begin{aligned} \lambda_{\mathbf{z}_i} &= [c\mathbf{1}_4 \ , \ \mathbf{0}_{p-3}]^{\mathrm{T}}, \text{ if } \mathbf{z}_i = 1, \text{ and} \\ \lambda_{\mathbf{z}_i} &= [\mathbf{0}_{p-3} \ , \ c\mathbf{1}_4]^{\mathrm{T}}, \text{ if } \mathbf{z}_i = 2. \end{aligned}$$

As before, we define the precision matrices as $\Omega_i = (\lambda_{\mathbf{z}_i}{\lambda_{\mathbf{z}_i}}^{\mathrm{T}} + 10\mathbb{I}_{p+1})$, and thus, the true graph structure $\mathrm{G}^*(\mathbf{z})$ for an individual with covariate value $\mathbf{z}$ is

$$\mathrm{G}^*(1) = \begin{bmatrix} \mathbb{J}_4 - \mathbb{I}_4 & \mathbf{0}_{4,p-3} \\ \mathbf{0}_{p-3,4} & \mathbf{0}_{p-3,p-3} \end{bmatrix}, \quad \mathrm{G}^*(2) = \begin{bmatrix} \mathbf{0}_{p-3,p-3} & \mathbf{0}_{p-3,4} \\ \mathbf{0}_{4,p-3} & \mathbb{J}_4 - \mathbb{I}_4 \end{bmatrix}.$$

We visualize these precision matrices and the corresponding dependence structures for $p = 10$ in Figure 1.

In addition to varying signal strength and sample structure for $p = 10$, we additionally vary the dimension of the data to $p = 30$ and $p = 50$ with high signal strength and balanced samples. In all

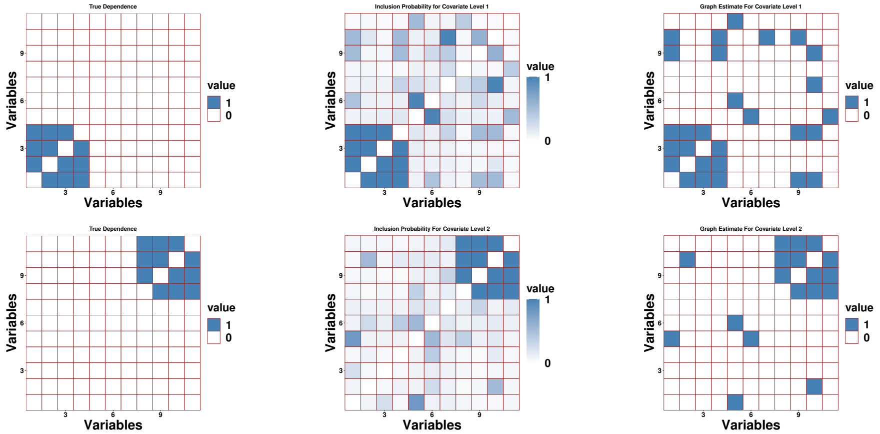


Figure 1: *Left to right: True dependence structures, estimated inclusion probabilities and estimated graphs for a sample simulation with two discrete covariate levels with p=10.*

experiments, we fix $n = 100$. We present results for these experiments in Table 3.

While the performance of W-PL and mgm are similar for $p = 10$, as p increases, the performance of mgm deteriorates. On the other hand, W-PL and CS demonstrate robustness to the increased sample size. As in the covariate-independent setting, the performance of both mgm and CS is significantly harmed relative to W-PL when faced with reduced signal.

E.4 High-Dimensional Setting

Our last series of experiments with the discrete covariate deals with a challenging high-dimensional setting where $p \geq n$. To handle the increased dimensionality, we found it necessary to modify our hyperparameter specification scheme for this experiment. We use Carbonetto et al. (2012) to obtain an Empirical Bayes estimate to the hyperparameter σ^2, and use a grid search to optimize over the hyperparameters π and σ_θ^2 using the ELBO as our objective function. For the bandwidth hyperparameter, we consider an ad-hoc choice of $\tau = 0.1$. We only perform 20 trials per experiment in this setting.

We maintain the relationship between the covariates and the ground truth structure as in Section E.3 and first consider an unbalanced setting with $p = 100, n_1 = 40, n_2 = 10$ and high signal. We present results from this setting in Figure 2, and exclude mgm from this experiment due to its

p	c	n_1	n_2	Method	Sensitivity(↑)	Specificity(↑)
10	3	50	50	W-PL	$\mathbf{0.5800}(0.1859)$	$0.9900(0.0133)$
				mgm	$0.0867(0.0937)$	$\mathbf{0.9980}(0.0046)$
				CS	$0.2950(0.1545)$	$0.9927(0.0090)$
10	15	50	50	W-PL	$\mathbf{1.0000}(0.0000)$	$0.9849(0.0187)$
				mgm	$0.9950(0.0354)$	$\mathbf{0.9982}(0.0053)$
				CS	$\mathbf{1.0000}(0.0000)$	$0.9953(0.0072)$
10	15	80	20	W-PL	$0.7973(0.0189)$	$0.9867(0.0106)$
				mgm	$\mathbf{0.8127}(0.0212)$	$\mathbf{0.9990}(0.0028)$
				CS	$0.8093(0.0166)$	$0.9983(0.0046)$
30	15	50	50	W-PL	$\mathbf{1.0000}(0.0000)$	$0.9926(0.0036)$
				mgm	$0.7567(0.1812)$	$\mathbf{0.9997}(0.0006)$
				CS	$\mathbf{1.0000}(0.0000)$	$0.9976(0.0018)$
50	15	50	50	W-PL	$0.9867(0.0425)$	$0.9958(0.0016)$
				mgm	$0.4550(0.2022)$	$\mathbf{0.9999}(0.0002)$
				CS	$\mathbf{0.9983}(0.0118)$	$0.9982(0.0009)$

Table 3: *Results for discrete covariate-dependent setting*

deteriorating performance with large p and high time-complexity. Note that as only 10 observations belong to level 2, separate estimation through CS suffers significantly compared to W-PL when estimating the graph for level 2.

Next, to demonstrate how the signal-to-noise ratio (SNR) influences the performance of our approach, we study several further experiments in the high-dimensional setting with $n = 50, p = 50$ and $n = 50, p = 100$, $n_1 = 20, n_2 = 30$, $\lambda = (c\mathbf{1}_4, \mathbf{0}_{p-3})^{\mathrm{T}}$ keeping other settings the same. Note that the SNR is controlled by c.

To assess performance under sparsity and with weak signal strength, we analyze the area under the receiver operating characteristic curve (AUC). By varying the threshold for a posterior inclusion probability to indicate an edge, we obtain a sequence of true and false positive ratios that we may use to calculate the corresponding AUC. AUC can also be defined by the fraction of pairs that the prediction ordered correctly: let $y_1, ..., y_n$ be the 0 and 1 responses and $p_1, ..., p_n$ be the

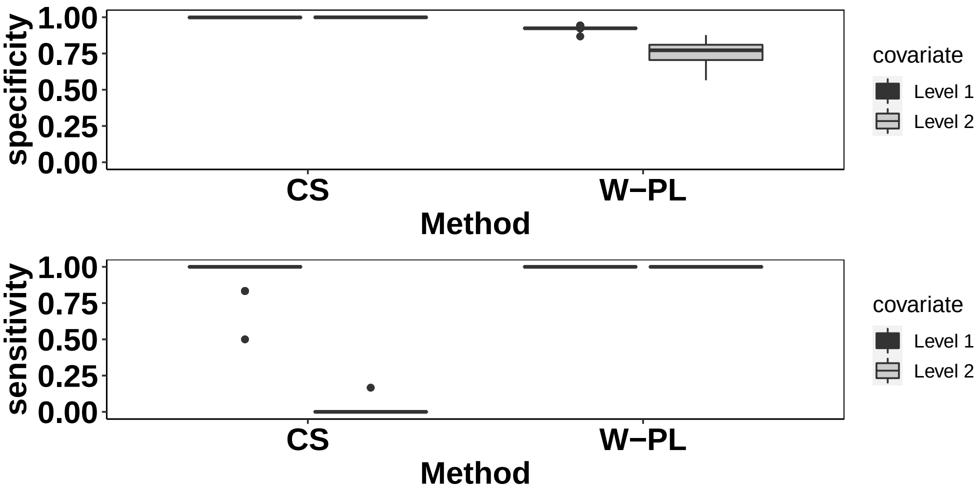


Figure 2: *Results for covariate-dependent setting with* $n = 50, p = 100$. *Top row: Specificity CS vs W-PL; Bottom row: sensitivity CS vs W-PL*

corresponding predicted probabilities. The AUC can then be calculated as

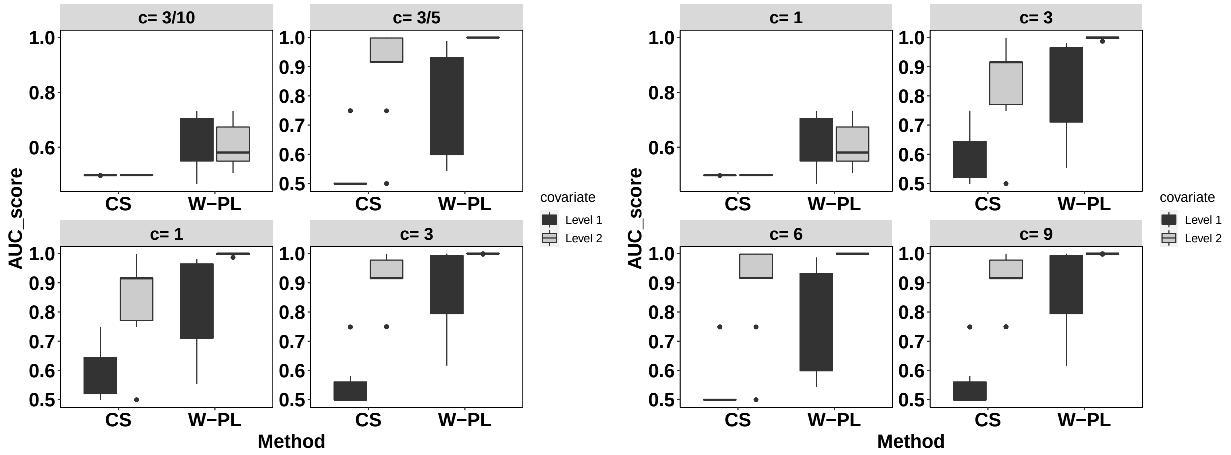
$$\sum_{i=1}^{n}\sum_{j=1}^{n}\mathbb{1}\{y_i < y_j\}\mathbb{1}\{p_i < p_j\}/\sum_{i=1}^{n}\sum_{j=1}^{n}\mathbb{1}_{\{y_i<y_j\}}$$

We present results from these experiments in Figure 3. As the signal strength (i.e., c) increases, the AUC also increases from around 0.5 to 1. When the SNR is low, W-PL does not work well and produces an AUC close to 0.5, which essentially is a random guess as to the presence of an edge. However, when there are sufficient observations and the SNR is high, the AUC for level 2 exceeds 0.9.

Because of the low level of observations in the first level of the covariate ($n_1 = 20$), separate estimation with CS does not perform well. W-PL consistently outperforms CS for both level 1 and level 2.

F Departure from Gaussian assumption

The method theoretically is built on the assumption that the true data generation is Gaussian, while the pseudo-likelihood approach is used mostly as a tool for estimation. To study the effects of departures from Gaussianity, we have investigated two scenarios. Firstly, we consider the situation



(a) $n = 50, p = 50$

(b) $n = 50, p = 100$

Figure 3: *CS versus W-PL method, measured by the AUC scores between the true graph structure and the calculated inclusion probabilities.*

where the data is contaminated, that is, the data comes from a Gaussian distribution with an independent structure, $c\%$ of which is contaminated by data coming from an unrelated independent Gaussian distribution. Figure 4 shows the results with 5% contamination. The results, however, get worse as the amount of contamination increases. Secondly, we consider the t-distribution with varying degrees of freedom. Figure 5 shows the sensitivity and specificity for varying degrees of freedom. The results indicate that for degrees of freedom greater than 6, the results are stable, and naturally shows improvement as the degrees of freedom increases. However, for degrees of freedom less than 6, the performance suffers, as shown in the left panel.

G Comparison to Qiu et al. (2016)

Here, we provide a brief comparison of W-PL to the method of Qiu et al. (2016). Although their method is applicable in the continuous covariate setting, their work focuses on the case when there are time replicates per subject. It is possible to extend their method to the case where there are not replicates by estimating the subject-level covariance matrices using a kernel-weighted average, however, the implementation provided by the authors does not include this functionality. Thus, we mainly focused on loggle and mgm as competitors for W-PL in our experiments, since the available

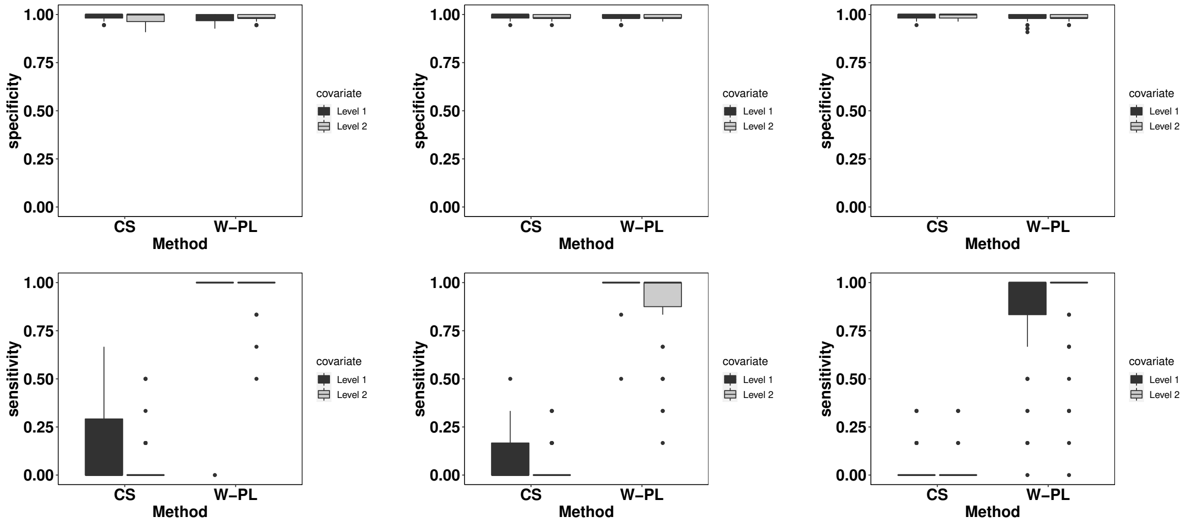


Figure 4: *Specificity and sensitivity comparisons between CS VS W-PL for data with* 2% *(left),* 5% *(middle) and* 10% *(right) contamination.*

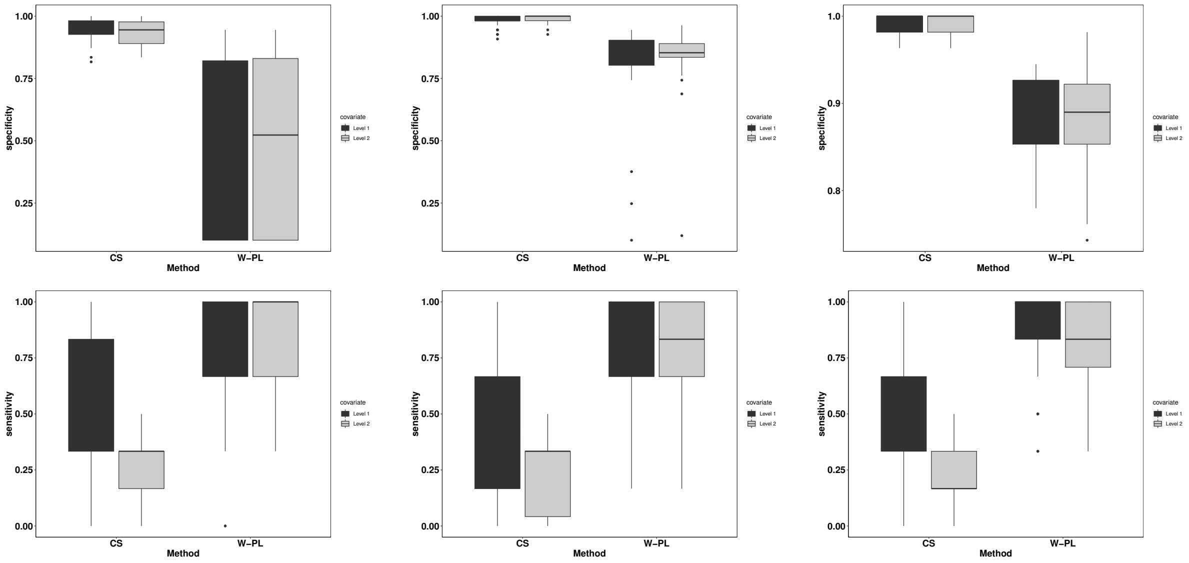


Figure 5: *Specificity and sensitivity comparisons between CS VS W-PL for t distributed data with df 3 (left), 6(middle) and 12(right), with* $n = 50$ *in both groups and* $p = 11$*.*

implementations of these models could directly handle data without replicates, and similar to Qiu et al. (2016), both are kernel-based models that frame the precision matrix as varying continuously with time.

We also note that although there are methods for modeling heterogeneous dependence structures other than those that model the dependence structure as varying with time, such as Ren et al. (2022), we are not aware of any that are directly applicable to a continuous covariate. For example, Ren et al. (2022) assumes that the data may be grouped into clusters such that the precision matrix is homogeneous within each of the clusters, and that the means of the clusters are sufficiently separable from one another. On the other hand, W-PL can model the precision matrices as varying continuously and does not place any restrictions on the mean structure of the data.

For comparison to Qiu et al. (2016), we used their simulation setting with 2 (the minimum allowed) time replicates for 100 individuals with 10 variables. We consider estimating the graph for subject 1. For W-PL, we only used the information at the first time point, whereas for Qiu et al. (2016), we used the full information on both time points. In this experiment, we revert to the hyperparameter specification scheme from Section E.4. We summarize results from this experiment in Figure 6. Notably, W-PL obtains superior results even when the method of Qiu et al. (2016) technically uses double the number of observations.

H Ground-Truth Dependence Structures

H.1 Unidimensional Covariate

In the unidimensional covariate setting, we can split the individuals in three clusters based on their covariates: $\mathcal{C}_1 = \{i : -3 < \mathbf{z}_i < -1\}, \mathcal{C}_2 = \{i : -1 < \mathbf{z}_i < 1\}$ and $\mathcal{C}_3 = \{i : 1 < \mathbf{z} < 3\}$. Then, the ground truth precision structures for each of the clusters is given by:

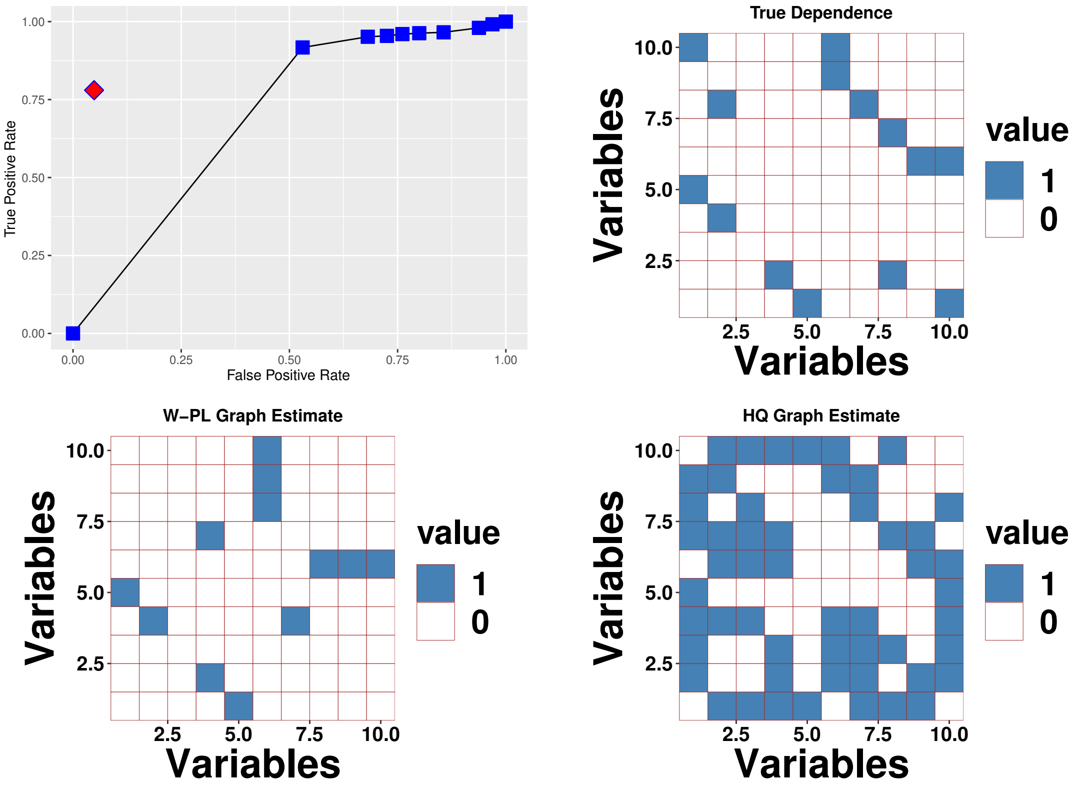


Figure 6: *Qiu's HQ method versus W-PL method. Top left shows the average TPR and FPR for HQ method (squares) versus our method (diamond) across 50 iterations. Top right shows the true dependence structure of the individual for a particular iteration, and the bottom left shows our estimate. Bottom right shows HQ estimate.*

$$\mathrm{G}^*(\mathcal{C}_1) = \begin{bmatrix} 0 & 1 & 0 & \\ 1 & 0 & 1 & \mathbf{0}_{3,p-2} \\ 0 & 1 & 0 & \\ & \mathbf{0}_{p-2,3} & & \mathbf{0}_{p-2,p-2} \end{bmatrix} \qquad \mathrm{G}^*(\mathcal{C}_2) = \begin{bmatrix} 0 & 1 & 1 & \\ 1 & 0 & 1 & \mathbf{0}_{3,p-2} \\ 1 & 1 & 0 & \\ & \mathbf{0}_{p-2,3} & & \mathbf{0}_{p-2,p-2} \end{bmatrix}$$

$$\mathrm{G}^*(\mathcal{C}_3) = \begin{bmatrix} 0 & 0 & 1 & \\ 0 & 0 & 1 & \mathbf{0}_{3,p-2} \\ 1 & 1 & 0 & \\ & \mathbf{0}_{p-2,3} & & \mathbf{0}_{p-2,p-2} \end{bmatrix}$$

H.2 Multidimensional Covariate

In the multidimensional covariate setting, we can split the individuals in nine clusters based on their covariates:

$$\mathcal{C}_1 = \{i : \mathbf{z}_i \in (-3,-1) \times (-3,-1)\} \qquad \mathcal{C}_2 = \{i : \mathbf{z}_i \in (-3,-1) \times (-1,1)\}$$

$$\mathcal{C}_3 = \{i : \mathbf{z}_i \in (-3,-1) \times (1,3)\} \qquad \mathcal{C}_4 = \{i : \mathbf{z}_i \in (-1,1) \times (-3,-1)\}$$

$$\mathcal{C}_5 = \{i : \mathbf{z}_i \in (-1,1) \times (-1,1)\} \qquad \mathcal{C}_6 = \{i : \mathbf{z}_i \in (-1,1) \times (1,3)\}$$

$$\mathcal{C}_7 = \{i : \mathbf{z}_i \in (1,3) \times (-3,-1)\} \qquad \mathcal{C}_8 = \{i : \mathbf{z}_i \in (1,3) \times (-1,1)\}$$

$$\mathcal{C}_9 = \{i : \mathbf{z}_i \in (1,3) \times (1,3)\}$$

Then, the ground truth precision structures for each of the clusters is given by:

$$\mathrm{G}^*(\mathcal{C}_1 \cup \mathcal{C}_4) = \begin{bmatrix} 0 & 1 & 0 & \\ 1 & 0 & 1 & \mathbf{0}_{3,p-2} \\ 0 & 1 & 0 & \\ & \mathbf{0}_{p-2,3} & & \mathbf{0}_{p-2,p-2} \end{bmatrix} \quad \mathrm{G}^*(\mathcal{C}_2 \cup \mathcal{C}_3 \cup \mathcal{C}_5 \cup \mathcal{C}_6) = \begin{bmatrix} 0 & 1 & 1 & \\ 1 & 0 & 1 & \mathbf{0}_{3,p-2} \\ 1 & 1 & 0 & \\ & \mathbf{0}_{p-2,3} & & \mathbf{0}_{p-2,p-2} \end{bmatrix}$$

$$\mathrm{G}^*(\mathcal{C}_7) = \begin{bmatrix} 0 & 0 & 0 & \\ 0 & 0 & 1 & \mathbf{0}_{3,p-2} \\ 0 & 1 & 0 & \\ & \mathbf{0}_{p-2,3} & & \mathbf{0}_{p-2,p-2} \end{bmatrix} \quad \mathrm{G}^*(\mathcal{C}_8 \cup \mathcal{C}_9) = \begin{bmatrix} 0 & 0 & 1 & \\ 0 & 0 & 1 & \mathbf{0}_{3,p-2} \\ 1 & 1 & 0 & \\ & \mathbf{0}_{p-2,3} & & \mathbf{0}_{p-2,p-2} \end{bmatrix}$$

References

ABRAMSON, I. S. (1982). On bandwidth variation in kernel estimates-a square root law. *The annals of Statistics* , 1217–1223.

ATCHADÉ, Y. F. (2019). Quasi-bayesian estimation of large gaussian graphical models. *Journal of Multivariate Analysis* **173**, 656–671.

CARBONETTO, P., STEPHENS, M. et al. (2012). Scalable variational inference for bayesian variable selection in regression, and its accuracy in genetic association studies. *Bayesian analysis* **7**, 73–108.

CARBONETTO, P., ZHOU, X. & STEPHENS, M. (2017). varbvs: Fast Variable Selection for Large-scale Regression.

CASTILLO, I., VAN DER VAART, A. et al. (2012). Needles and straw in a haystack: Posterior concentration for possibly sparse sequences. *The Annals of Statistics* **40**, 2069–2101.

DASGUPTA, S., PATI, D. & SRIVASTAVA, A. (2020). A two-step geometric framework for density modeling. *Statistica Sinica* **30**, 2155–2177.

HASLBECK, J. M. B. & WALDORP, L. J. (2020). mgm: Estimating Time-Varying Mixed Graphical Models in High-Dimensional Data. *Journal of Statistical Software* **93**, 1–46.

QIU, H., HAN, F., LIU, H. & CAFFO, B. (2016). Joint estimation of multiple graphical models from high dimensional time series. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **78**, 487–504.

RASKUTTI, G., WAINWRIGHT, M. J. & YU, B. (2010). Restricted eigenvalue properties for correlated gaussian designs. *The Journal of Machine Learning Research* **11**, 2241–2259.

REN, M., ZHANG, S., ZHANG, Q. & MA, S. (2022). Gaussian graphical model-based heterogeneity analysis via penalized fusion. *Biometrics* **78**, 524–535.

RUDELSON, M. & VERSHYNIN, R. (2013). Hanson-wright inequality and sub-gaussian concentration. *Electronic Communications in Probability* **18**, 1–9.

VAN KERM, P. (2003). Adaptive kernel density estimation. *The Stata Journal* **3**, 148–156.

YANG, Y., PATI, D., BHATTACHARYA, A. et al. (2020). α-variational inference with statistical guarantees. *Annals of Statistics* **48**, 886–905.

ZHOU, S. (2009). Restricted eigenvalue conditions on subgaussian random matrices. *arXiv preprint arXiv:0912.4045* .