Supplementary Material for "Bayesian Group Regularization in Generalized Linear Models with a Continuous Spike-and-Slab Prior"

Ray Bai
Department of Statistics, George Mason University

September 21, 2025

Abstract

Section A and Section B give the proofs for Section 3 and Section 4 of the main manuscript respectively. Section C gives additional details for the expectation-maximization (EM) algorithm introduced in Section 5.1 of the main manuscript. Section D gives the complete Gibbs sampler for the MCMC algorithms introduced in Section 5.3 of the main manuscript. Finally, Section E presents simulation studies for our model in negative binomial regression with a log link.

A Proofs for Section 3

Before proving Theorem 1, it will be necessary to define some terminology. Letting $\pi(\boldsymbol{\beta})$ denote the spike-and-slab group lasso (SSGL) prior (10), we can express the SSGL log prior as

$$\log \pi(\boldsymbol{\beta}) = \sum_{g=1}^{G} \text{pen}(\boldsymbol{\beta}_g),$$

where

$$\text{pen}(\boldsymbol{\beta}_g) = \log \left[(1-\theta)\boldsymbol{\Psi}(\boldsymbol{\beta}_g \mid \lambda_0) + \theta \boldsymbol{\Psi}(\boldsymbol{\beta}_g \mid \lambda_1) \right]. \tag{A.1}$$

Recall that

$$p^{\star}_{\theta}(\boldsymbol{\beta}_g) = \frac{\theta \boldsymbol{\Psi}(\boldsymbol{\beta}_g \mid \lambda_1)}{\theta \boldsymbol{\Psi}(\boldsymbol{\beta}_g \mid \lambda_1) + (1-\theta)\boldsymbol{\Psi}(\boldsymbol{\beta}_g \mid \lambda_0)}. \tag{A.2}$$

The derivative of $\text{pen}(\boldsymbol{\beta}_g)$ with respect to $\|\boldsymbol{\beta}_g\|_2$ is

$$\frac{\partial \text{pen}(\boldsymbol{\beta}_g)}{\partial \|\boldsymbol{\beta}_g\|_2} = -\lambda^{\star}_{\theta}(\boldsymbol{\beta}_g), \tag{A.3}$$

where

$$\lambda^{\star}_{\theta}(\boldsymbol{\beta}_g) = \lambda_1 p^{\star}_{\theta}(\boldsymbol{\beta}_g) + \lambda_0 \left[1 - p^{\star}_{\theta}(\boldsymbol{\beta}_g) \right]. \tag{A.4}$$

Under the objective function (13), it can then be seen from (5) and (A.3) that any local maximizer $\widehat{\boldsymbol{\beta}}$ of (13) must satisfy

$$\mathbf{X}_g^\top \mathrm{diag}\{\xi'(\mathbf{X}\widehat{\boldsymbol{\beta}})\}\left(\mathbf{Y} - b'(\xi(\mathbf{X}\widehat{\boldsymbol{\beta}}))\right) - \lambda_\theta^\star(\widehat{\boldsymbol{\beta}}_g)\partial\|\boldsymbol{\beta}_g\|_2 = \mathbf{0}_{m_g} \text{for all } g \in \{1,\ldots,G\}, \tag{A.5}$$

where ∂f denotes the subdifferential of f. From (A.5), we can obtain the necessary first-order Karush-Kuhn-Tucker (KKT) conditions for $\widehat{\boldsymbol{\beta}}$ to be a local maximizer of (13),

$$\begin{aligned} &\mathbf{X}_g^\top \mathrm{diag}\{\xi'(\mathbf{X}\widehat{\boldsymbol{\beta}})\}(\mathbf{Y} - b'(\xi(\mathbf{X}\widehat{\boldsymbol{\beta}}))) - \lambda_\theta^\star(\widehat{\boldsymbol{\beta}}_g)\tfrac{\widehat{\boldsymbol{\beta}}_g}{\|\widehat{\boldsymbol{\beta}}_g\|_2} = \mathbf{0}_{m_g}, && \text{for } \widehat{\boldsymbol{\beta}}_g \neq \mathbf{0}_{m_g}, \\ &\left\|\mathbf{X}_g^\top \mathrm{diag}\{\xi'(\mathbf{X}\widehat{\boldsymbol{\beta}})\}(\mathbf{Y} - b'(\xi(\mathbf{X}\widehat{\boldsymbol{\beta}})))\right\|_2 \leq \lambda_\theta^\star(\mathbf{0}_{m_g}) && \text{for } \widehat{\boldsymbol{\beta}}_g = \mathbf{0}_{m_g}. \end{aligned} \tag{A.6}$$

Proof of Theorem 1 Recall that $\boldsymbol{\beta}_0 = (\boldsymbol{\beta}_{0S_0}^\top, \mathbf{0}^\top)^\top$. To prove Theorem 1, we do the following two steps. First, we consider the subspace of vectors $\mathcal{T} = \{\boldsymbol{\beta} \in \mathbb{R}^p : \boldsymbol{\beta}_{S_0^c} = \mathbf{0}\}$. We show that on $\mathcal{T}$, there exists a local maximizer $\widehat{\boldsymbol{\beta}} = (\widehat{\boldsymbol{\beta}}_{S_0}^\top, \mathbf{0}^\top)^\top$ of (13) under the SSGL prior (10) such that $\|\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0\|_2 = O_p(\epsilon_n)$. Next, we show that $\widehat{\boldsymbol{\beta}}$ is a strict local maximizer of (13) under the SSGL prior (10) on the *entire* parameter space $\mathbb{R}^p$.

Step 1 (local consistency). First we solve the constrained penalized likelihood estimation problem on the subspace $\mathcal{T} = \{\boldsymbol{\beta} \in \mathbb{R}^p : \boldsymbol{\beta}_{S_0^c} = \mathbf{0}\}$. Then on $\mathcal{T}$, any estimator of $\boldsymbol{\beta}_0$ must be of the form $\widehat{\boldsymbol{\beta}} = (\widehat{\boldsymbol{\beta}}_{S_0}^\top, \mathbf{0}^\top)^\top$. We first show the existence of such a $\widehat{\boldsymbol{\beta}}$ to the optimization problem,

$$\widehat{\boldsymbol{\beta}} = \arg\max_{\boldsymbol{\beta}} Q_1(\boldsymbol{\beta}) = \arg\max_{\boldsymbol{\beta}} \left\{ \ell_n(\boldsymbol{\beta}) + \sum_{g=1}^{s_0} \mathrm{pen}(\boldsymbol{\beta}_g), \;\; \boldsymbol{\beta}_{S_0^c} = \mathbf{0} \right\}, \tag{A.7}$$

where $\mathrm{pen}(\boldsymbol{\beta}_g)$ is defined as in (A.1) and $\widehat{\boldsymbol{\beta}}$ satisfies

$$\|\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0\|_2 = O_p(\epsilon_n), \tag{A.8}$$

with $\epsilon_n = (s_0 \log G/n)^{1/2}$ and $\boldsymbol{\beta}_0 = (\boldsymbol{\beta}_{0S_0}^\top, \mathbf{0}^\top)^\top$. Let $\boldsymbol{\Delta} = (\boldsymbol{\Delta}_{S_0}^\top, \mathbf{0}^\top)^\top$, where $\|\boldsymbol{\Delta}_{S_0}\|_2 = C$. In order to establish (A.8), it suffices to show that, for some small $\varepsilon > 0$ and sufficiently large $C > 0$,

$$P\left\{ \sup_{\|\boldsymbol{\Delta}_{S_0}\|_2 = C} Q_1(\boldsymbol{\beta}_0 + \epsilon_n \boldsymbol{\Delta}) < Q_1(\boldsymbol{\beta}_0) \right\} \geq 1 - \varepsilon, \tag{A.9}$$

as $n \to \infty$. If (A.9) holds, this will imply that with probability at least $1 - \varepsilon$, there exists a local maximum inside the ball $\{\boldsymbol{\beta}_0 + \epsilon_n \boldsymbol{\Delta} : \|\boldsymbol{\Delta}_{S_0}\|_2 \leq C\}$ as $n \to \infty$. In order to establish (A.9), we must show that with probability tending

to one and sufficiently large $C > 0$,

$$\begin{aligned} & Q_1(\boldsymbol{\beta}_0 + \epsilon_n \boldsymbol{\Delta}) - Q_1(\boldsymbol{\beta}_0) \\ & \quad = [\ell_n(\boldsymbol{\beta}_0 + \epsilon_n \boldsymbol{\Delta}) - \ell_n(\boldsymbol{\beta}_0)] + \sum_{g=1}^{s_0} [\mathrm{pen}(\boldsymbol{\beta}_{0g} + \epsilon_n \boldsymbol{\Delta}_g) - \mathrm{pen}(\boldsymbol{\beta}_{0g})] \\ & \quad \triangleq \mathrm{(I)} + \mathrm{(II)} < 0, \end{aligned} \tag{A.10}$$

as $n \to \infty$. We bound the terms in (A.10) separately. We first focus on bounding $\mathrm{(I)} := \ell_n(\boldsymbol{\beta}_0 + \epsilon_n \boldsymbol{\Delta}) - \ell_n(\boldsymbol{\beta}_0)$. Let ∇_{S_0} and $\nabla^2_{S_0}$ denote the partial derivatives with respect to $\boldsymbol{\beta}_{0S_0}$. By a Taylor expansion and the fact that $\boldsymbol{\beta}_0 = (\boldsymbol{\beta}_{0S_0}^\top, \mathbf{0}^\top)^\top$, we have

$$\begin{aligned} \mathrm{(I)} &= \epsilon_n \nabla_{S_0} \ell_n(\boldsymbol{\beta}_0)^\top \boldsymbol{\Delta}_{S_0} + \frac{\epsilon_n^2}{2} \boldsymbol{\Delta}_{S_0}^\top \nabla^2_{S_0} \ell_n(\widetilde{\boldsymbol{\beta}}) \boldsymbol{\Delta}_{S_0} \\ &\triangleq J_1 + J_2, \end{aligned} \tag{A.11}$$

where $\widetilde{\boldsymbol{\beta}}$ lies on the line segment between $\boldsymbol{\beta}_0$ and $\boldsymbol{\beta}_0 + \epsilon_n \boldsymbol{\Delta}$. Next, we bound the terms J_1 and J_2 in (A.11) separately. We have

$$\nabla_{S_0} \ell_n(\boldsymbol{\beta}_0) = \mathbf{X}_{S_0}^\top \mathrm{diag}\{\xi'(\mathbf{X}\boldsymbol{\beta}_0)\}(\mathbf{Y} - b'(\xi(\mathbf{X}\boldsymbol{\beta}_0))).$$

Let d_2 denote the largest diagonal entry in $\mathrm{diag}(\xi'(\mathbf{X}\boldsymbol{\beta}_0))$, which we know is finite and positive due to Assumption (A4)(ii). Then we have, for some $A > 0$,

$$\begin{aligned} P\left(\|\nabla_{S_0} \ell_n(\boldsymbol{\beta}_0)\|_2 > An\epsilon_n\right) &\le P\left(\left\|\frac{1}{n}\mathbf{X}_{S_0}^\top(\mathbf{Y} - b'(\xi(\mathbf{X}\boldsymbol{\beta}_0)))\right\|_2^2 > A^2\epsilon_n^2/d_2^2\right) \\ &\le \frac{d_2^2 n}{A^2 s_0 \log G} \mathbb{E}\left\{\sum_{g=1}^{s_0}\sum_{k=1}^{m_g}\left[\frac{1}{n}\sum_{i=1}^{n} x_{igk}\left(y_i - b'(\xi(\mathbf{x}_i^\top \boldsymbol{\beta}_0))\right)\right]^2\right\} \\ &= \frac{d_2^2}{A^2 s_0 \log G} \mathrm{tr}\left(\frac{1}{n}\mathbf{X}_{S_0}^\top \boldsymbol{\Omega}(\boldsymbol{\beta}_0)\mathbf{X}_{S_0}\right) \\ &\le \frac{d_2^2}{A^2 s_0 \log G}\left(\sum_{g=1}^{s_0} m_g\right)\lambda_{\max}\left(\frac{1}{n}\mathbf{X}_{S_0}^\top \boldsymbol{\Omega}(\boldsymbol{\beta}_0)\mathbf{X}_{S_0}\right) \\ &\le \frac{d_2^2 s_0 m_{\max}}{A^2 s_0 \log G}\lambda_{\max}\left(\frac{1}{n}\mathbf{X}_{S_0}^\top \boldsymbol{\Omega}(\boldsymbol{\beta}_0)\mathbf{X}_{S_0}\right) \\ &\lesssim \frac{d_2^2 \log n}{A^2 \log G}\lambda_{\max}\left(\frac{1}{n}\mathbf{X}_{S_0}^\top \boldsymbol{\Omega}(\boldsymbol{\beta}_0)\mathbf{X}_{S_0}\right) \to 0 \quad \text{as} \quad n \to \infty, \end{aligned} \tag{A.12}$$

where the last inequality of the display uses Assumption (A2) that $m_{\max} = O(\log n \wedge (\log G / \log n))$ and Assumption (A3)(ii) that $\lambda_{\max}\left(n^{-1}\mathbf{X}_{S_0}^\top \boldsymbol{\Omega}(\boldsymbol{\beta}_0)\mathbf{X}_{S_0}\right) = o(\log G / \log n)$. Therefore, from (A.12), we have that $\|\nabla_{S_0} \ell_n(\boldsymbol{\beta}_0)\|_2 = O_p(n\epsilon_n)$, and so an upper bound for J_1 in (A.11) is

$$J_1 \le |J_1| \le \epsilon_n \|\nabla_{S_0} \ell_n(\boldsymbol{\beta}_0)\|_2 \|\boldsymbol{\Delta}_{S_0}\|_2 = An\epsilon_n^2 \|\boldsymbol{\Delta}_{S_0}\|_2.$$

for some $A > 0$. We also have by Assumption (A3)(ii) that

$$\begin{aligned} J_2 &= -\frac{\epsilon_n^2}{2}\boldsymbol{\Delta}_{S_0}^\top \left\{\mathbf{X}_{S_0}^\top \boldsymbol{\Sigma}(\widetilde{\boldsymbol{\beta}})\mathbf{X}_{S_0}\right\}\boldsymbol{\Delta}_{S_0} \\ &\leq -\frac{\epsilon_n^2}{2}\lambda_{\min}\left(\mathbf{X}_{S_0}^\top \boldsymbol{\Sigma}(\widetilde{\boldsymbol{\beta}})\mathbf{X}_{S_0}\right)\|\boldsymbol{\Delta}_{S_0}\|_2^2 \\ &\lesssim -\frac{n\epsilon_n^2}{2}\|\boldsymbol{\Delta}_{S_0}\|_2^2. \end{aligned}$$

Therefore, an upper bound for (I) in (A.10) is

$$\text{(I)} \lesssim An\epsilon_n^2\|\boldsymbol{\Delta}_{S_0}\|_2 - \frac{n\epsilon_n^2}{2}\|\boldsymbol{\Delta}_{S_0}\|_2^2 \tag{A.13}$$

Now we turn our attention to upper-bounding (II) in (A.10). We first focus on bounding each of the summands in (II) from above. We first rewrite $\text{pen}(\boldsymbol{\beta}_g)$ in (A.1) as

$$\text{pen}(\boldsymbol{\beta}_g) = -\lambda_1\|\boldsymbol{\beta}_g\|_2 - \log\left[p_\theta^\star(\boldsymbol{\beta}_g)\right],$$

where $p_\theta^\star(\boldsymbol{\beta}_g)$ is defined as in (A.2). Therefore, for any one group g,

$$\begin{aligned} &\text{pen}(\boldsymbol{\beta}_{0g} + \epsilon_n\boldsymbol{\Delta}_g) - \text{pen}(\boldsymbol{\beta}_{0g}) \\ &\quad= \lambda_1\left\{\|\boldsymbol{\beta}_{0g}\|_2 - \|\boldsymbol{\beta}_{0g} + \epsilon_n\boldsymbol{\Delta}_g\|_2\right\} + \left[\log\left(p_\theta^\star(\boldsymbol{\beta}_{0g})\right) - \log\left(p_\theta^\star(\boldsymbol{\beta}_{0g} + \epsilon_n\boldsymbol{\Delta}_g)\right)\right] \\ &\quad\leq \lambda_1\epsilon_n\|\boldsymbol{\Delta}_g\|_2 - \left[\log(p_\theta^\star(\boldsymbol{\beta}_{0g} + \epsilon_n\boldsymbol{\Delta}_g)) - \log(p_\theta^\star(\boldsymbol{\beta}_{0g}))\right] \\ &\quad\leq \lambda_1\epsilon_n\|\boldsymbol{\Delta}_g\|_2 - \epsilon_n\left(\frac{\partial p_\theta^\star(\widetilde{\boldsymbol{\beta}}_g)}{\partial\widetilde{\boldsymbol{\beta}}_g}\right)^\top\boldsymbol{\Delta}_g, \\ &\qquad\text{where } \widetilde{\boldsymbol{\beta}}_g \text{ lies on line segment between } \boldsymbol{\beta}_{0g} \text{ and } \boldsymbol{\beta}_{0g} + \epsilon_n\boldsymbol{\Delta}_g, \\ &\quad\leq \epsilon_n\|\boldsymbol{\Delta}_g\|_2\left\{\lambda_1 + \left\|\frac{\partial p_\theta^\star(\widetilde{\boldsymbol{\beta}}_g)}{\partial\widetilde{\boldsymbol{\beta}}_g}\right\|_2\right\}, \end{aligned} \tag{A.14}$$

We examine the second term in brackets in (A.14). Noting that

$$p_\theta^\star(\boldsymbol{\beta}_g) = \frac{1}{1 + \left(\frac{1-\theta}{\theta}\right)\left(\frac{\lambda_0}{\lambda_1}\right)^{m_g}\exp\left[-(\lambda_0 - \lambda_1)\|\boldsymbol{\beta}_g\|_2\right]} := \frac{1}{1 + c(\boldsymbol{\beta}_g)}, \tag{A.15}$$

we can see that when $\boldsymbol{\beta}_g \neq \mathbf{0}_{m_g}$,

$$\frac{\partial p_\theta^\star(\boldsymbol{\beta}_g)}{\partial\boldsymbol{\beta}_g} = \frac{(\lambda_0 - \lambda_1)c(\boldsymbol{\beta}_g)}{(1 + c(\boldsymbol{\beta}_g))^2} \times \frac{\boldsymbol{\beta}_g}{\|\boldsymbol{\beta}_g\|_2}.$$

Note that

$$\frac{(\lambda_0 - \lambda_1)c(\boldsymbol{\beta}_g)}{(1 + c(\boldsymbol{\beta}_g))^2} < (\lambda_0 - \lambda_1)c(\boldsymbol{\beta}_g) \prec 1,$$

since the exponent term $\exp[-(\lambda_0 - \lambda_1)\|\boldsymbol{\beta}_g\|_2]$ in $c(\boldsymbol{\beta}_g)$ decays faster than $(\lambda_0 - \lambda_1)$ multiplied by the other terms in $c(\boldsymbol{\beta}_g)$ as $n \to \infty$. Since it must be that $\widetilde{\boldsymbol{\beta}}_g$ in (A.14) is a nonzero vector, we have that for large n,

$$\left\| \frac{\partial p_\theta^\star(\widetilde{\boldsymbol{\beta}}_g)}{\partial \widetilde{\boldsymbol{\beta}}_g} \right\|_2 < \left\| \frac{\widetilde{\boldsymbol{\beta}}_g}{\|\widetilde{\boldsymbol{\beta}}_g\|_2} \right\|_2 = 1,$$

and so we can further bound (A.14) from above by $\epsilon_n(1+\lambda_1)\|\boldsymbol{\Delta}_g\|_2$. Therefore, an upper bound for (II) in (A.10) is

$$\begin{aligned} \text{(II)} \lesssim \epsilon_n \sum_{g=1}^{s_0} (1+\lambda_1)\|\boldsymbol{\Delta}_g\|_2 &\leq s_0 \epsilon_n (1+\lambda_1) \max_{1 \leq g \leq s_0} \|\boldsymbol{\Delta}_g\|_2 \\ &\leq s_0 \epsilon_n (1+\lambda_1)\|\boldsymbol{\Delta}_{S_0}\|_2. \end{aligned} \tag{A.16}$$

Combining the upper bounds in (A.13) and (A.16), it is clear that

$$\text{(I)} + \text{(II)} \lesssim An\epsilon_n^2 \|\boldsymbol{\Delta}_{S_0}\|_2 - \frac{n\epsilon_n^2}{2}\|\boldsymbol{\Delta}_{S_0}\|_2^2 + s_0\epsilon_n(1+\lambda_1)\|\boldsymbol{\Delta}_{S_0}\|_2,$$

and the right-hand side can be made negative by choosing $C := \|\boldsymbol{\Delta}_{S_0}\|_2$ large enough. For large $C > 0$, the negative term above is the dominating term. This proves (A.10), and thus, $\widehat{\boldsymbol{\beta}} = (\widehat{\boldsymbol{\beta}}_{S_0}^\top, \mathbf{0}^\top)^\top$ is a local maximum over the subspace $\mathcal{T} = \{\boldsymbol{\beta} \in \mathbb{R}^p : \boldsymbol{\beta}_{S_0^c} = \mathbf{0}\}$ as $n \to \infty$.

Step 2 (sparsity). Since $\log \pi(\boldsymbol{\beta})$ is a nonconvex function, the KKT conditions (A.6) only give necessary conditions for $\widehat{\boldsymbol{\beta}}$ from Step 1 to be a local maximum on $\mathcal{T}$. However, by suitably modifying Theorem 1 of Fan and Lv (2011) to the grouped GLM setting with the log SSGL prior as the penalty function, we can also obtain *sufficient* conditions for $\widehat{\boldsymbol{\beta}} \in \mathcal{T}$ from Step 1 to be a strict local maximum over *all* of $\mathbb{R}^p$. This will be the case if, for sufficiently large n, $\widehat{\boldsymbol{\beta}}$ from Step 1 satisfies

$$\lambda_{\max}\left\{\nabla_g^2 \ell_n(\widehat{\boldsymbol{\beta}}) + \nabla^2 \text{pen}(\widehat{\boldsymbol{\beta}}_g)\right\} < 0 \;\; \text{for all } \widehat{\boldsymbol{\beta}}_g \neq \mathbf{0}_{m_g}, g \in S_0, \tag{A.17}$$

where ∇_g^2 denotes the second partial derivative with respect to $\widehat{\boldsymbol{\beta}}_g$, $\text{pen}(\boldsymbol{\beta}_g)$ is the penalty function in (A.1), and

$$\|\mathbf{X}_g^\top \text{diag}(\xi'(\mathbf{X}\widehat{\boldsymbol{\beta}}))[\mathbf{Y} - b'(\xi(\mathbf{X}\widehat{\boldsymbol{\beta}}))]\|_2 < \lambda_\theta^\star(\mathbf{0}_{m_g}) \;\; \text{for all } g \in S_0^c. \tag{A.18}$$

That is, we require the KKT second order sufficiency condition to hold as $n \to \infty$, and the second inequality in (A.6) has to hold with a *strict* inequality.

We first prove (A.17). For $\widehat{\boldsymbol{\beta}}_g \neq \mathbf{0}_{m_g}$, we have by (6) and the chain rule for

second derivatives (with $f(\mathbf{u}) = \text{pen}(\mathbf{u})$ and $\mathbf{u} = g(\mathbf{v}) = \|\mathbf{v}\|_2$) that

$$\begin{aligned}
&\nabla_g^2 \ell_n(\widehat{\boldsymbol{\beta}}) + \nabla^2 \text{pen}(\widehat{\boldsymbol{\beta}}_g) \\
&\quad = -\mathbf{X}_g^\top \boldsymbol{\Sigma}(\widehat{\boldsymbol{\beta}}) \mathbf{X}_g + p_\theta^\star(\widehat{\boldsymbol{\beta}}_g)[1 - p_\theta^\star(\widehat{\boldsymbol{\beta}}_g)](\lambda_0 - \lambda_1)^2 \frac{\widehat{\boldsymbol{\beta}}_g \widehat{\boldsymbol{\beta}}_g^\top}{\|\widehat{\boldsymbol{\beta}}_g\|_2^2} \\
&\qquad - \lambda_\theta^\star(\widehat{\boldsymbol{\beta}}_g) \left[\frac{1}{\|\widehat{\boldsymbol{\beta}}_g\|_2} \mathbf{I}_{m_g} - \frac{\widehat{\boldsymbol{\beta}}_g \widehat{\boldsymbol{\beta}}_g^\top}{\|\widehat{\boldsymbol{\beta}}_g\|_2^3} \right]. \qquad \text{(A.19)}
\end{aligned}$$

Note that $\lambda_{\min}(\|\widehat{\boldsymbol{\beta}}_g\|_2^{-1} \mathbf{I}_{m_g} - \|\widehat{\boldsymbol{\beta}}_g\|_2^{-3} \widehat{\boldsymbol{\beta}}_g \widehat{\boldsymbol{\beta}}_g^\top) = 0$, since $\mathbf{I}_{m_g} - \|\widehat{\boldsymbol{\beta}}_g\|_2^{-2} \widehat{\boldsymbol{\beta}}_g \widehat{\boldsymbol{\beta}}_g^\top$ is the annihilator matrix. Using the fact that $\lambda_{\max}(-\mathbf{A}) = -\lambda_{\min}(\mathbf{A})$ for a square matrix $\mathbf{A}$, it follows from (A.19) that

$$\begin{aligned}
&\lambda_{\max} \left\{ \nabla_g^2 \ell_n(\widehat{\boldsymbol{\beta}}) + \nabla^2 \text{pen}(\widehat{\boldsymbol{\beta}}_g) \right\} \\
&\leq -\lambda_{\min} \left(\mathbf{X}_g^\top \boldsymbol{\Sigma}(\widehat{\boldsymbol{\beta}}) \mathbf{X}_g \right) + \ p_\theta^\star(\widehat{\boldsymbol{\beta}}_g)[1 - p_\theta^\star(\widehat{\boldsymbol{\beta}}_g)](\lambda_0 - \lambda_1)^2 \\
&\lesssim -n + \frac{c(\widehat{\boldsymbol{\beta}}_g)(\lambda_0 - \lambda_1)^2}{(1 + c(\widehat{\boldsymbol{\beta}}_g))^2}, \quad \text{where } c(\boldsymbol{\beta}_g) \text{ is as in (A.15)} \\
&\leq -n + c(\widehat{\boldsymbol{\beta}}_g)(\lambda_0 - \lambda_1)^2 \\
&\lesssim -n + 1 < 0,
\end{aligned}$$

for sufficiently large n, where we used Assumption (A3)(ii) in the third line and the fact that given our choices of hyperparameters for $(\theta, \lambda_0, \lambda_1)$ in Theorem 1, $c(\boldsymbol{\beta}_g)(\lambda_0 - \lambda_1)^2 \prec 1$ for any $\boldsymbol{\beta}_g \neq \mathbf{0}_{m_g}$ in the last line. Thus, we have shown that (A.17) holds for large n.

Next, we prove (A.18). It suffices to show that as $n \to \infty$,

$$P \left(\exists \ g \in S_0^c, \quad \left\| \frac{\partial \ell_n(\widehat{\boldsymbol{\beta}})}{\partial \boldsymbol{\beta}_g} \right\|_2 \geq \lambda_\theta^\star(\mathbf{0}_{m_g}) \right) \to 0.$$

To ease the notation, let $h_g(\widehat{\boldsymbol{\beta}}) = \partial \ell_n(\widehat{\boldsymbol{\beta}}) / \partial \boldsymbol{\beta}_g$. By a Taylor expansion around $\boldsymbol{\beta}_0$ and the fact that $\widehat{\boldsymbol{\beta}}_{0 S_0^c} = \mathbf{0}$, we have

$$h_g(\widehat{\boldsymbol{\beta}}) = h_g(\boldsymbol{\beta}_0) + \nabla_{S_0} h_g(\widetilde{\boldsymbol{\beta}})^\top (\widehat{\boldsymbol{\beta}}_{S_0} - \boldsymbol{\beta}_{0 S_0}), \qquad \text{(A.20)}$$

where $\widetilde{\boldsymbol{\beta}}$ lies on the line segment between $\boldsymbol{\beta}_0$ and $\widehat{\boldsymbol{\beta}}$. By (A.20), we have that

$$
\begin{aligned}
&P\left(\exists\ g \in S_0^c,\quad \|h_g(\widehat{\boldsymbol{\beta}})\|_2 \geq \lambda_\theta^\star(\mathbf{0}_{m_g})\right) \\
&\quad \leq P\left(\exists\ g \in S_0^c,\quad \|h_g(\boldsymbol{\beta}_0)\|_2 \geq \frac{\lambda_\theta^\star(\mathbf{0}_{m_g})}{2}\right) \\
&\qquad + P\left(\exists\ g \in S_0^c,\quad \|\nabla_{S_0} h_g(\widetilde{\boldsymbol{\beta}})^\top(\widehat{\boldsymbol{\beta}}_{S_0} - \boldsymbol{\beta}_{0S_0})\|_2 \geq \frac{\lambda_\theta^\star(\mathbf{0}_{m_g})}{2}\right) \\
&\quad \leq \sum_{g=s_0+1}^{G} P\left(\|h_g(\boldsymbol{\beta}_0)\|_2 \geq \frac{\lambda_\theta^\star(\mathbf{0}_{m_g})}{2}\right) \\
&\qquad + \sum_{g=s_0+1}^{G} P\left(\|\nabla_{S_0} h_g(\widetilde{\boldsymbol{\beta}})^\top(\widehat{\boldsymbol{\beta}}_{S_0} - \boldsymbol{\beta}_{0S_0})\|_2 \geq \frac{\lambda_\theta^\star(\mathbf{0}_{m_g})}{2}\right) \\
&\quad \triangleq T_1 + T_2. \qquad \text{(A.21)}
\end{aligned}
$$

We will bound each of the terms in (A.21) from above separately. First note that since $\lambda_0 = (1-\theta)/\theta \asymp G^c, c > 2$, and $\lambda_1 \asymp n^{-1}$, we have that for some constant $M > 0$,

$$
p_\theta^\star(\mathbf{0}_{m_g}) = \frac{1}{1 + \left(\frac{1-\theta}{\theta}\right)\left(\frac{\lambda_0}{\lambda_1}\right)^{m_g}} = \frac{1}{1 + Mn\lambda_0^{m_g+1}},
$$

and thus,

$$
\begin{aligned}
\lambda_\theta^\star(\mathbf{0}_{m_g}) &= \lambda_1 p_\theta^\star(\mathbf{0}_{m_g}) + \lambda_0\left[1 - p_\theta^\star(\mathbf{0}_{m_g})\right] \\
&= \frac{\lambda_1 + \lambda_0(Mn\lambda_0^{m_g+1})}{1 + Mn\lambda_0^{m_g+1}} > \lambda_0\left(\frac{Mn\lambda_0^{m_g+1}}{1 + Mn\lambda_0^{m_g+1}}\right) > \frac{\lambda_0}{2}, \qquad \text{(A.22)}
\end{aligned}
$$

for large n.

We first examine each of the summands in T_1 in (A.21). Since all of the entries of $\mathbf{X}$ satisfy $|x_{ij}| = O(\log G)$ (by Assumption (A3)(i)), we have that for some constant $D > 0$ and any $g \in \{1, \ldots, G\}$, $\|\mathbf{X}_g\|_{2,\infty} \leq D\log G\sqrt{m_g} \leq D\log G\sqrt{m_{\max}}$. Thus, in light of (A.22), we have that for sufficiently large n,

$$
\begin{aligned}
P\left(\|h_g(\boldsymbol{\beta}_0)\|_2 \geq \frac{\lambda_\theta^\star(\mathbf{0}_{m_g})}{2}\right) &< P\left(\|\mathbf{X}_g\|_{2,\infty}\|\mathbf{Y} - b'(\xi(\mathbf{X}\boldsymbol{\beta}_0))\|_2 \geq \frac{\lambda_0}{2}\right) \\
&\leq P\left(\|\mathbf{Y} - b'(\xi(\mathbf{X}\boldsymbol{\beta}_0))\|_2 \geq \frac{\lambda_0}{2D\log G\sqrt{m_{\max}}}\right) \\
&\leq 2\exp\left(-\frac{\lambda_0^2/(4D^2(\log G)^2 m_{\max})}{2[2nG + L\lambda_0/(D\log G\sqrt{m_{\max}})]}\right) \\
&= 2\exp\left(-\frac{\lambda_0^2}{16D^2(\log G)^2 m_{\max} nG + 8L\lambda_0 D\log G\sqrt{m_{\max}}}\right) \qquad \text{(A.23)}
\end{aligned}
$$

where we used Assumption (A4)(i) and the Bernstein inequality (specifically, Lemma 2.2.11 of van der Vaart and Wellner (2006)). Thus, from (A.23), we have as an upper bound for T_1 in (A.21),

$$T_1 < 2\exp\left(\log G - \frac{\lambda_0^2}{16D^2(\log G)^2 m_{\max} nG + 8L\lambda_0 D \log G\sqrt{m_{\max}}}\right) \to 0, \tag{A.24}$$

where we used the fact that $\lambda_0 \asymp G^c, c > 2$, and Assumptions (A1) and (A2) that $G \gg n$ and $m_{\max} = O(\log n \wedge (\log G/\log n))$. Thus, the second term in the exponent of (A.24) dominates the first term as $n \to \infty$.

We next examine each of the summands in T_2 in (A.21). In light of (A.22) that $\lambda_\theta^\star(\mathbf{0}_{m_g}) > \lambda_0/2$, we have that for each $g \in S_0^c$ and sufficiently large n,

$$\begin{aligned}
&P\left(\|\nabla_{S_0} h_g(\widetilde{\boldsymbol{\beta}})^\top(\widehat{\boldsymbol{\beta}}_{S_0} - \boldsymbol{\beta}_{0S_0})\|_2 \geq \frac{\lambda_\theta^\star(\mathbf{0}_{m_g})}{2}\right) \\
&\quad < P\left(\|\nabla_{S_0} h_g(\widetilde{\boldsymbol{\beta}})^\top(\widehat{\boldsymbol{\beta}}_{S_0} - \boldsymbol{\beta}_{0S_0})\|_2 \geq \frac{\lambda_0}{4}\right) \\
&\quad \leq P\left(\|\nabla_{S_0} h_g(\widetilde{\boldsymbol{\beta}})\|_{2,\infty}\|\widehat{\boldsymbol{\beta}}_{S_0} - \boldsymbol{\beta}_{0S_0}\|_2 \geq \frac{\lambda_0}{4}\right) \\
&\quad \leq P\left(\|\mathbf{X}_{S_0}^\top \boldsymbol{\Sigma}(\widetilde{\boldsymbol{\beta}})\mathbf{X}_g\|_{2,\infty}\|\widehat{\boldsymbol{\beta}}_{S_0} - \boldsymbol{\beta}_{0S_0}\|_2 \geq \frac{\lambda_0}{4}\right) \\
&\quad \lesssim \frac{4n\epsilon_n}{\lambda_0},
\end{aligned} \tag{A.25}$$

To arrive at (A.25), we used the fact that $\nabla_{S_0} h_g(\widetilde{\boldsymbol{\beta}}) = \mathbf{X}_{S_0}^\top \boldsymbol{\Sigma}(\widetilde{\boldsymbol{\beta}})\mathbf{X}_g$ in the second line. In the last line, we used Assumption (A3)(iii) that for all $g \in \mathcal{S}_0^c$, $\|\mathbf{X}_{S_0}^\top \boldsymbol{\Sigma}(\widetilde{\boldsymbol{\beta}})\mathbf{X}_g\|_{2,\infty} = O(n)$, the fact that $\|\widehat{\boldsymbol{\beta}}_{S_0} - \boldsymbol{\beta}_{0S_0}\|_2 = O_p(\epsilon_n)$, and Markov's inequality. From (A.25), the fact that $\lambda_0 \asymp G^c, c > 2$, and $G \gg n$, we have as an upper bound for T_2 in (A.21),

$$T_2 \prec \frac{4Gn\epsilon_n}{\lambda_0} \prec \frac{Gn}{G^c} = \frac{n}{G^{c-1}} \to 0 \text{ as } n \to \infty. \tag{A.26}$$

Combining (A.21), (A.24), and (A.26) allows us to conclude that (A.18) holds for sufficiently large n, and the proof is finished. □

B Proofs for Section 4

To prove Theorems 2 and 3, we follow the proof strategy of Jeong and Ghosal (2021). However, a crucial difference is that we study an *absolutely continuous* spike-and-slab prior in this paper, whereas Jeong and Ghosal (2021) prove their results for a *point-mass* spike-and-slab prior. The continuous nature of the SSGL requires it to be handled quite differently from the point mass prior of Jeong and Ghosal (2021). In particular, since SSGL (10) puts zero probability

on exactly sparse configurations of $\boldsymbol{\beta}$, we need to instead rely on notions of "approximate" sparsity to appropriately bound the approximation error.

We first define some necessary notation. Let f_i be the density of y_i belonging to the exponential family (1), where the natural parameter θ_i is related to the grouped covariates $\mathbf{x}_i = (\mathbf{x}_{i1}^\top, \ldots, \mathbf{x}_{iG}^\top)^\top$ through (3). Denote f_{0i} analogously with true natural parameter θ_{0i}. Then the joint densities are $f = \prod_{i=1}^n f_i$ and $f_0 = \prod_{i=1}^n f_{0i}$. We define the average squared Hellinger metric between f and f_0 as $H_n^2(f, f_0) = n^{-1} \sum_{i=1}^n H^2(f_i, f_{0i})$, where $H(f_i, f_{0i}) = (\int (\sqrt{f_i} - \sqrt{f_{0i}})^2)^{1/2}$ is the Hellinger distance between f_i and f_{0i}. The Kullback-Leibler (KL) divergence and KL variation between f_i and f_{0i} are given by $K(f_{0i}, f_i) = \int f_{0i} \log(f_{0i}/f_i)$ and $V(f_{0i}, f_i) = \int f_{0i}(\log(f_{0i}/f_i) - K(f_{0i}, f_i))^2$ respectively. For a set Ω with semimetric d, the ε-covering number $N(\varepsilon, \Omega, d)$ is the minimal number of d-balls of radius ε needed to cover Ω. Meanwhile, the ε-packing number for Ω, denoted by $D(\varepsilon, \Omega, d)$, is the maximal number of d balls of radius ε needed to cover Ω.

B.1 Proof of Theorem 2

We first prove a lemma that is needed to prove Theorem 2.

Lemma 1 (evidence lower bound)**.** *Assume the same setup as Theorem 2. Then* $\sup \mathbb{P}_0(\mathcal{E}_n^c) \to 0$*, where the set* $\mathcal{E}_n$ *is defined as*

$$\mathcal{E}_n = \left\{ \int \prod_{i=1}^n \frac{f_i(y_i)}{f_{0i}(y_i)} d\Pi(\boldsymbol{\beta}) \geq e^{-C_1 n \epsilon_n^2} \right\}, \tag{B.1}$$

for some constant $C_1 > 0$ *and* $\epsilon_n^2 = s_0 \log G/n$*.*

Proof. By Lemma 10 of Ghosal and van der Vaart (2007), this statement will be proven if we can show that

$$\Pi(\mathcal{B}_n) \gtrsim e^{-C_1 n \epsilon_n^2}, \tag{B.2}$$

where the set $\mathcal{B}_n$ is defined as the event,

$$\mathcal{B}_n = \left\{ \frac{1}{n} \sum_{i=1}^n K(f_{0i}, f_i) \leq \epsilon_n^2, \frac{1}{n} \sum_{i=1}^n V(f_{0i}, f_i) \leq \epsilon_n^2 \right\}. \tag{B.3}$$

As established in Lemma 1 of Jeong and Ghosal (2021), the KL divergence and KL variation for the exponential family (1) is given by

$$\begin{aligned} K(f_{0i}, f_i) &= (\theta_{0i} - \theta_i) b'(\theta_{0i}) - b(\theta_{0i}) + b(\theta_i), \\ V(f_{0i}, f_i) &= b''(\theta_{0i})(\theta_i - \theta_{0i})^2. \end{aligned}$$

Recalling that $\theta_i = \xi(\mathbf{x}_i^\top \boldsymbol{\beta})$ and $\theta_{0i} = \xi(\mathbf{x}_i^\top \boldsymbol{\beta}_0)$ where $\xi = (h \circ b')^{-1}$, we have by Taylor expansion in $\mathbf{x}_i^\top \boldsymbol{\beta}$ at $\mathbf{x}_i^\top \boldsymbol{\beta}_0$ that

$$\begin{aligned} &\max \{K(f_{0i}, f_i), V(f_{0i}, f_i)\} \\ &\quad \leq b''(\theta_{0i})(\xi'(\mathbf{x}_i^\top \boldsymbol{\beta}_0))^2 (\mathbf{x}_i^\top \boldsymbol{\beta} - \mathbf{x}_i^\top \boldsymbol{\beta}_0)^2 + o((\mathbf{x}_i^\top \boldsymbol{\beta} - \mathbf{x}_i^\top \boldsymbol{\beta}_0)^2) \\ &\quad = (h^{-1})'(\mathbf{x}_i^\top \boldsymbol{\beta}_0) \xi'(\mathbf{x}_i^\top \boldsymbol{\beta}_0)(\mathbf{x}_i^\top \boldsymbol{\beta} - \mathbf{x}_i^\top \boldsymbol{\beta}_0)^2 + o((\mathbf{x}_i^\top \boldsymbol{\beta} - \mathbf{x}_i^\top \boldsymbol{\beta}_0)^2). \end{aligned}$$

Note that h^{-1} and ξ are strictly increasing. Thus, it must be the case that if $\max_{1\le i\le n}\{(h^{-1})'(\mathbf{x}_i^\top\boldsymbol{\beta}_0)\xi'(\mathbf{x}_i^\top\boldsymbol{\beta}_0)(\mathbf{x}_i^\top\boldsymbol{\beta}-\mathbf{x}_i^\top\boldsymbol{\beta}_0)^2\}\le\epsilon_n^2\to 0$, then $|\mathbf{x}_i^\top\boldsymbol{\beta}-\mathbf{x}_i^\top\boldsymbol{\beta}_0|\to 0$ for all $1\le i\le n$. Recall $\boldsymbol{\Omega}(\boldsymbol{\beta}_0)=\mathrm{diag}\{(h^{-1})'(\mathbf{X}\boldsymbol{\beta}_0)\}\mathrm{diag}\{\xi'(\mathbf{X}\boldsymbol{\beta}_0)\}$. Then both $n^{-1}\sum_{i=1}^n K(f_{0i},f_i)$ and $n^{-1}\sum_{i=1}^n V(f_{0i},f_i)$ can be bounded above by a constant multiple of $n^{-1}\|\boldsymbol{\Omega}^{1/2}(\boldsymbol{\beta}_0)\mathbf{X}(\boldsymbol{\beta}-\boldsymbol{\beta}_0)\|_2^2$ on $\mathcal{B}_n$.

Thus, for sufficiently large n, we have for some constants $\widetilde{C}_2, C_2>0$,

$$\begin{aligned}
\Pi(\mathcal{B}_n) &\ge \Pi\left(\frac{1}{n}\|\boldsymbol{\Omega}^{1/2}(\boldsymbol{\beta}_0)\mathbf{X}(\boldsymbol{\beta}-\boldsymbol{\beta}_0)\|_2^2\le\widetilde{C}_2^2\epsilon_n^2\right)\\
&\ge \Pi\left(\lambda_{\max}\left(\frac{1}{n}\mathbf{X}_g^\top\boldsymbol{\Omega}(\boldsymbol{\beta}_0)\mathbf{X}_g\right)\left(\sum_{g=1}^G\|\boldsymbol{\beta}_g-\boldsymbol{\beta}_{0g}\|_2\right)^2\le\widetilde{C}_2^2\epsilon_n^2\right)\\
&\ge \Pi\left(\sum_{g=1}^G\|\boldsymbol{\beta}_g-\boldsymbol{\beta}_{0g}\|_2\le\frac{C_2\sqrt{s_0}}{\sqrt{n}}\right)\\
&\ge \Pi\left(\sum_{g\in S_0}\|\boldsymbol{\beta}_g-\boldsymbol{\beta}_{0g}\|_2\le\frac{C_2\sqrt{s_0}}{2\sqrt{n}}\right)\Pi\left(\sum_{g\in S_0^c}\|\boldsymbol{\beta}_g-\boldsymbol{\beta}_{0g}\|_2\le\frac{C_2\sqrt{s_0}}{2\sqrt{n}}\right)\\
&\ge \prod_{g\in S_0}\Pi\left(\|\boldsymbol{\beta}_g-\boldsymbol{\beta}_{0g}\|_2^2\le\frac{C_2^2}{4n}\right)\prod_{g\in S_0^c}\Pi\left(\|\boldsymbol{\beta}_{S_0^c}\|_2^2\le\frac{C_2^2 s_0}{4n(G-s_0)}\right), \quad \text{(B.4)}
\end{aligned}$$

where we used Assumption (B3)(ii) in the third inequality and the Cauchy-Schwarz inequality in the fifth inequality. Arguing similarly as in (D.17)-(D.20) in the proof of Theorem 2 of Bai et al. (2022), we have that

$$\prod_{g\in S_0}\Pi\left(\|\boldsymbol{\beta}_g-\boldsymbol{\beta}_{0g}\|_2^2\le\frac{C_2^2}{4n}\right)\ge\theta^{s_0}e^{-\lambda_1\|\boldsymbol{\beta}_{S_0}\|_2}e^{-\lambda_1 C_2/2\sqrt{n}}\prod_{g\in S_0}\frac{C_g}{m_g!}\left(\frac{\lambda_1 C_2}{2s_0\sqrt{n}}\right)^{m_g},$$

where $C_g=2^{-m_g}\pi^{-(m_g-1)/2}[\Gamma((m_g+1)/2)]^{-1}$ and

$$\prod_{g\in S_0^c}\Pi\left(\|\boldsymbol{\beta}_g\|_2^2\le\frac{C_2^2 s_0}{4n(G-s_0)}\right)\ge(1-\theta)^{G-s_0}\left[1-\frac{4nGm_{\max}(m_{\max}+1)}{\lambda_0^2C_2^2s_0}\right]^{G-s_0}.$$

Thus, a lower bound on (B.4) is

$$\begin{aligned}
\Pi(\mathcal{B}_n)\ge\theta^{s_0}(1-\theta)^{G-s_0}e^{-\lambda_1\|\boldsymbol{\beta}_{S_0}\|_2}e^{-\lambda_1C_2/2\sqrt{n}}&\left[1-\frac{4nGm_{\max}(m_{\max}+1)}{\lambda_0^2C_2^2s_0}\right]^{G-s_0}\\
&\times\prod_{g\in S_0}\frac{C_g}{m_g!}\left(\frac{\lambda_1C_2}{2s_0\sqrt{n}}\right)^{m_g}. \quad \text{(B.5)}
\end{aligned}$$

Now, since $\lambda_0=(1-\theta)/\theta\asymp G^c, c>2$, and $s_0=o(n^{1/2}/\log G)$ and $m_{\max}=O(\log n\wedge(\log G/\log n))$ by Assumption (A2), we have that

$$\left[1-\frac{4nGm_{\max}(m_{\max}+1)}{\lambda_0^2C_2^2s_0}\right]^{G-s_0}\gtrsim\left[1-\frac{1}{G-s_0}\right]^{G-s_0}\to e^{-1}\text{ as }n\to\infty.$$

Furthermore, $\theta \asymp \frac{1}{G^c+1}$, and thus,

$$(1-\theta)^{G-s_0} \gtrsim \left(1-\frac{1}{G-s_0}\right)^{G-s_0} \to e^{-1}.$$

Thus, using Assumption (B4) that $\|\boldsymbol{\beta}_0\|_\infty = O(\log G)$ and $\|\boldsymbol{\beta}_{S_0}\|_2 \le \sqrt{s_0}\|\boldsymbol{\beta}_0\|_\infty$, we can further lower bound (B.5) as

$$\Pi(\mathcal{B}_n) \gtrsim (G^c+1)^{-s_0} e^{-\lambda_1\left(\sqrt{s_0}\log G + C_1/2\sqrt{n}\right)} \prod_{g\in S_0} \frac{C_g}{m_g!}\left(\frac{\lambda_1 C_2}{2s_0\sqrt{n}}\right)^{m_g},$$

and since $\lambda_1 \asymp 1/n$, we have

$$\begin{aligned} &-\log\Pi(\mathcal{B}_n) \\ &\quad\lesssim 2cs_0\log G + \sqrt{s_0}\log G + \sum_{g\in S_0}\left[\log(m_g!) - \log(C_g) + m_g\log\left(\frac{2s_0\sqrt{n}}{\lambda_1 C_2}\right)\right]. \end{aligned} \tag{B.6}$$

Similar to (D.24) in the proof of Theorem 2 of Bai et al. (2022), the first term of (B.6) can be shown to be the dominating term, and thus, $-\log\Pi(\mathcal{B}_n) \lesssim n\epsilon_n^2$. This establishes (B.2) and the proof is finished. □ □

Proof of Theorem 2 We follow the proof strategy of Theorem 2 in Jeong and Ghosal (2021) but work with the generalized dimensionality $|\nu(\boldsymbol{\beta})|$ (20). Let $\mathcal{E}_n$ be the event defined in (B.1), where $\epsilon_n = (s_0\log G/n)^{1/2}$. Define the set $\mathcal{A}_n = \{\boldsymbol{\beta} : |\nu(\boldsymbol{\beta})| \le K_1 s_0\}$, where $K_1 \ge C_1 \vee 1$, and C_1 is the constant in Lemma 1. Then we have

$$\mathbb{E}_0\Pi(\mathcal{A}_n^c \mid \mathbf{Y}) \le \mathbb{E}_0\Pi\left(\mathcal{A}_n^c \mid \mathbf{Y}\right)\mathbb{1}_{\mathcal{E}_n} + \mathbb{P}_0(\mathcal{E}_n^c).$$

By Lemma 1, $\mathbb{P}_0(\mathcal{E}_n^c) \to 0$ as $n\to\infty$, so it suffices to show that $\mathbb{E}_0\Pi(\mathcal{A}_n^c \mid \mathbf{Y})\mathbb{1}_{\mathcal{E}_n} \to 0$. Note that

$$\Pi(\mathcal{A}_n^c \mid \mathbf{Y}) = \frac{\int_{\mathcal{A}_n^c}\prod_{i=1}^n \frac{f_i(y_i)}{f_{0i}(y_i)} d\Pi(\boldsymbol{\beta})}{\int\prod_{i=1}^n \frac{f_i(y_i)}{f_{0i}(y_i)} d\Pi(\boldsymbol{\beta})}. \tag{B.7}$$

On the set $\mathcal{E}_n$, the denominator in (B.7) is bounded below by $e^{-C_1 n\epsilon_n^2}$ by Lemma 1. On the other hand, an upper bound for the expected value of the numerator is

$$\mathbb{E}_0\left(\int_{\mathcal{A}_n^c}\prod_{i=1}^n \frac{f_i(y_i)}{f_{0i}(y_i)} d\Pi(\boldsymbol{\beta})\right) \le \int_{\mathcal{A}_n^c} d\Pi(\boldsymbol{\beta}) = \Pi(|\boldsymbol{\beta}| > K_1 s_0). \tag{B.8}$$

Now, since $\pi(\boldsymbol{\beta}_g \mid \theta) < 2\theta C_g \lambda_1^{m_g} e^{-\lambda_1 \|\boldsymbol{\beta}_g\|_2}$ for all $\|\boldsymbol{\beta}_g\|_2 > \omega_g$, we have (arguing as in (D.30) of Bai et al. (2022)) that

$$\begin{aligned}\Pi(|\nu(\boldsymbol{\beta})| > K_1 s_0) \leq \sum_{S:|S|>C_3 s_0} 2^{|S|}\theta^{|S|} &\left\{\prod_{g\in S}\int_{\|\boldsymbol{\beta}_g\|_2>\omega_g} C_g \lambda_1^{m_g} e^{-\lambda_1\|\boldsymbol{\beta}_g\|_2} d\boldsymbol{\beta}_g\right\} \\ &\times \left\{\prod_{g\in S^c}\int_{\|\boldsymbol{\beta}_g\|_2\leq w_g} \pi(\boldsymbol{\beta}_g) d\boldsymbol{\beta}_g\right\} \\ \lesssim \sum_{S:|S|>K_1 s_0} &\theta^{|S|},\end{aligned}$$

where we bounded the first bracketed term from above as $\prod_{g\in S}(1/n)^{m_g} \leq n^{-|S|}$ and the second bracketed term from above by one. Now, since $\theta < 1/G^2$, we have

$$\begin{aligned}\sum_{S:|S|>K_1 s_0} \theta^{|S|} \leq \sum_{k=\lfloor C_1 s_0\rfloor+1}^{G} \binom{G}{k}\left(\frac{1}{G^2}\right)^k &\leq \sum_{k=\lfloor K_1 s_0\rfloor+1}^{G}\left(\frac{e}{kG}\right)^k \\ &< \sum_{k=\lfloor K_1 s_0\rfloor+1}^{G}\left(\frac{e}{G\lfloor K_1 s_0\rfloor+1}\right)^k \\ &\leq G^{-(\lfloor K_1 s_0\rfloor+1)} \lesssim e^{-K_1 n\epsilon_n^2}. \quad \text{(B.9)}\end{aligned}$$

Combining (B.7)-(B.9) and Lemma 1 gives $\mathbb{E}_0\Pi(\mathcal{A}_n^c \mid \mathbf{Y})\mathbb{1}_{\mathcal{E}_n} \prec e^{-(K_1-C_1)n\epsilon_n^2}$ $\to 0$, since $K_1 > C_1$. This completes the proof. □

B.2 Proof of Theorem 3

We first prove a lemma which verifies that the prior puts sufficient mass on the event that the generalized dimensionality $|\boldsymbol{\nu}(\boldsymbol{\beta})|$ (20) equals s_0, where s_0 is the true number of nonzero groups in $\boldsymbol{\beta}_0$.

Lemma 2. *Assume the same setup as Theorem 2. Then for some constant $C_3 > 0$,*

$$\Pi(|\boldsymbol{\nu}(\boldsymbol{\beta})| = s_0) \geq e^{-C_3 n\epsilon_n^2}$$

Proof. Note that $\pi(\boldsymbol{\beta}_g) \geq \Psi(\boldsymbol{\beta}_g \mid \lambda_0)$ for all $\boldsymbol{\beta}_g$. Further, for $\boldsymbol{\beta}_g \sim \Psi(\boldsymbol{\beta}_g \mid \lambda_0)$, we have that $\|\boldsymbol{\beta}_g\|_2$ follows a gamma distribution with shape parameter m_g and

scale parameter λ_0. Thus,

$$\begin{aligned}\Pi(\|\boldsymbol{\beta}_g\|_2 > \omega_g) &\geq \int_{\omega_g}^{\infty} \frac{1}{\Gamma(m_g)\lambda_0^{m_g}} x^{m_g-1} e^{-x/\lambda_0} dx \\ &\geq \frac{1}{\Gamma(m_g)\lambda_0^{m_g}} \int_{\lambda_0}^{\infty} x^{m_g-1} e^{-x/\lambda_0} dx \\ &\geq \frac{1}{\Gamma(m_g)\lambda_0} \int_{\lambda_0}^{\infty} e^{-x/\lambda_0} dx \\ &= \frac{e^{-1}}{\Gamma(m_g)} \geq \frac{e^{-1}}{\Gamma(m_{\max})} > \frac{e^{-1}}{m_{\max}!}. \end{aligned} \tag{B.10}$$

Meanwhile,

$$\begin{aligned}\Pi(\|\boldsymbol{\beta}_g\|_2 \leq \omega_g) &\geq \int_{0}^{\omega_g} \frac{1}{\Gamma(m_g)\lambda_0^{m_g}} x^{m_g-1} e^{-x/\lambda_0} dx \\ &= 1 - \int_{\omega_g}^{\infty} \frac{1}{\Gamma(m_g)\lambda_0^{m_g}} x^{m_g-1} e^{-x/\lambda_0} dx \\ &\geq 1 - \exp(-\lambda_0\omega_g + m_{\max}) \\ &\geq 1 - \frac{1}{G - s_0}, \end{aligned} \tag{B.11}$$

where we used the tail bound for the gamma density on p. 29 of Boucheron et al. (2013) and the inequality $1 + x - \sqrt{1+2x} \geq (x-1)/2$ for $x > 0$ (similarly as in Ning et al. (2020)) to bound the integral in the second line from above by $\exp(-\lambda_0\omega_g + m_{\max})$.

From (B.10)-(B.11), we can bound $\Pi(|\boldsymbol{\nu}(\boldsymbol{\beta})| = s_0)$ from below as

$$\begin{aligned}\Pi(|\boldsymbol{\nu}(\boldsymbol{\beta})| = s_0) &\geq \left(\frac{e^{-1}}{m_{\max}!}\right)^{s_0} \left(1 - \frac{1}{G - s_0}\right)^{G-s_0} \\ &\gtrsim \left(\frac{e^{-1}}{m_{\max}^{m_{\max}}}\right)^{s_0}, \end{aligned}$$

where we used the facts that $(1-1/(G-s_0))^{G-s_0} \to e^{-1}$ as $n \to \infty$, and $x! \leq x^x$ in the second line of the display. Therefore, we see that for sufficiently large n,

$$-\log \Pi(|\boldsymbol{\nu}(\boldsymbol{\beta})| = s_0) \leq s_0 m_{\max} \log m_{\max} + s_0 \lesssim s_0 \log G, \tag{B.12}$$

by Assumptions (A1) and (A2) that $G \gg n$ and $m_{\max} = O(\log n \wedge (\log G / \log n))$. This proves the lemma. □ □

Proof of Theorem 3 The proof of Theorem 3 is broken up into two parts and follows the proof strategy of Theorems 2 and 3 of Jeong and Ghosal (2021). In the first part, we establish the posterior contraction rate in terms of the average squared Hellinger metric. In the second part, we convert the Hellinger contraction rate result to a contraction rate for the regression coefficients themselves.

Step 1: Hellinger contraction rate. We first show that there exists $C_4 > 0$ such that

$$\mathbb{E}_0\Pi\left(\boldsymbol{\beta} \in \mathbb{R}^p : H_n(\boldsymbol{\beta}, \boldsymbol{\beta}_0) > C_4\epsilon_n \mid \mathbf{Y}\right) \to 0 \text{ as } n \to \infty. \tag{B.13}$$

Define the set $\mathcal{A}_n = \{\boldsymbol{\beta} \in \mathbb{R}^p : |\boldsymbol{\nu}(\boldsymbol{\beta})| \leq K_1 s_0\}$, where K_1 is the constant from Theorem 2. Then for every $\epsilon > 0$,

$$\begin{aligned}
&\mathbb{E}_0\Pi\left(\boldsymbol{\beta} \in \mathbb{R}^p : H_n(\boldsymbol{\beta}, \boldsymbol{\beta}_0) > \epsilon \mid \mathbf{Y}\right) \\
&\quad \leq \mathbb{E}_0\Pi\left(\boldsymbol{\beta} \in \mathcal{A}_n : H_n(\boldsymbol{\beta}, \boldsymbol{\beta}_0) > \epsilon \mid \mathbf{Y}\right) \mathbb{1}_{\mathcal{E}_n} + \mathbb{E}_0\Pi(\mathcal{A}_n^c \mid \mathbf{Y}) + \mathbb{P}_0\mathcal{E}_n^c,
\end{aligned}$$

where $\mathcal{E}_n$ is in the event in (B.1). By Lemma 1 and Theorem 2, the second and third terms on the right-hand side above tend to zero uniformly over $\boldsymbol{\beta}_0$. Hence, in order to prove (B.13), we only focus on the first term. That is, it suffices to show that for $\epsilon = C\epsilon_n$, where $C > 0$ and $\epsilon_n = (s_0 \log G/n)^{1/2}$,

$$\mathbb{E}_0\Pi\left(\boldsymbol{\beta} \in \mathcal{A}_n : H_n(\boldsymbol{\beta}, \boldsymbol{\beta}_0) > \epsilon \mid \mathbf{Y}\right) \text{ as } n \to \infty, \tag{B.14}$$

uniformly over $\boldsymbol{\beta}_0$. Define the sieve,

$$\mathcal{A}_n^\star = \left\{\boldsymbol{\beta} \in \mathcal{A}_n : \|\boldsymbol{\beta} - \boldsymbol{\beta}_0\|_2 \leq G^M\right\}, \tag{B.15}$$

where $M > 0$ is defined as in Assumption (B3)(ii). By Lemma 2 of Ghosal and van der Vaart (2007), there exists a test function φ_n such that for any $\boldsymbol{\beta}_1$ with $H_n(\boldsymbol{\beta}_0, \boldsymbol{\beta}_1) > \epsilon$,

$$\mathbb{E}_0\varphi_n \leq \exp(-n\epsilon^2/2), \sup_{\boldsymbol{\beta}: H_n(\boldsymbol{\beta}, \boldsymbol{\beta}_1) \leq \epsilon/18} \mathbb{E}_{\boldsymbol{\beta}}(1 - \varphi_n) \leq \exp(-n\epsilon^2/2).$$

We want to use Lemma 9 of Ghosal and van der Vaart (2007). In order to do so, we need to show that $\log N(\epsilon_n/36, \mathcal{A}_n^\star, H_n) \lesssim n\epsilon_n^2$ for $\epsilon_n = (s_0 \log G/n)^{1/2}$. Since the squared Hellinger distance is bounded by one half of the KL divergence, we have by a Taylor expansion that for any $\boldsymbol{\beta}_1, \boldsymbol{\beta}_2 \in \mathcal{A}_n^\star$,

$$\begin{aligned}
&H^2(f_{\boldsymbol{\beta}_1,i}, f_{\boldsymbol{\beta}_2,i}) \\
&\quad \leq \frac{K(f_{\boldsymbol{\beta}_1,i}, f_{\boldsymbol{\beta}_2,i})}{2} \\
&\quad = \frac{(h^{-1})'(\mathbf{x}_i^\top\boldsymbol{\beta}_1)\xi'(\mathbf{x}_i^\top\boldsymbol{\beta}_1)}{2}(\mathbf{x}_i^\top\boldsymbol{\beta}_1 - \mathbf{x}_i^\top\boldsymbol{\beta}_2)^2 + o((\mathbf{x}_i^\top\boldsymbol{\beta}_1 - \mathbf{x}_i^\top\boldsymbol{\beta}_2)^2).
\end{aligned}$$

It then follows that

$$\begin{aligned}
H_n^2(f_{\boldsymbol{\beta}_1}, f_{\boldsymbol{\beta}_2}) &= \frac{1}{2n}\|\boldsymbol{\Omega}^{1/2}(\boldsymbol{\beta}_1)\mathbf{X}(\boldsymbol{\beta}_1 - \boldsymbol{\beta}_2)\|_2^2 + o\left(\|\mathbf{X}(\boldsymbol{\beta}_1 - \boldsymbol{\beta}_2)\|_2^2\right) \\
&\leq \frac{1}{2}\max_{1\leq g\leq G}\left(\lambda_{\max}\left(\frac{1}{n}\mathbf{X}_g^\top\boldsymbol{\Omega}(\boldsymbol{\beta}_1)\mathbf{X}_g\right)\right)\left(\sum_{g=1}^G\|\boldsymbol{\beta}_{1g} - \boldsymbol{\beta}_{2g}\|_2\right)^2 \\
&\qquad + o\left(\|\mathbf{X}(\boldsymbol{\beta}_1 - \boldsymbol{\beta}_2)\|_2^2\right) \\
&\lesssim G\log G\|\boldsymbol{\beta}_1 - \boldsymbol{\beta}_2\|_2^2 + \|\mathbf{X}\|_{2,\infty}^2\|\boldsymbol{\beta}_1 - \boldsymbol{\beta}_2\|_2^2 \\
&\lesssim G^2\|\boldsymbol{\beta}_1 - \boldsymbol{\beta}_2\|_2^2,
\end{aligned}$$

where we used Assumption (B3)(ii) that $\lambda_{\max}(n^{-1}\mathbf{X}_g^\top \mathbf{\Omega}(\boldsymbol{\beta}_1)\mathbf{X}_g) \lesssim \log G$ and the fact that all of the entries of $\mathbf{X}$ are uniformly bounded, and thus, $\|\mathbf{X}\|_{2,\infty} \lesssim G^{1/2}$. Thus, for some $C_5 > 0$, we can bound $N(\epsilon_n/36, \mathcal{A}_n^\star, H_n)$ from above by

$$\begin{aligned} N(\epsilon_n/36, \mathcal{A}_n^\star, H_n) &\leq N\left(\frac{C_5\epsilon_n}{36G}, \mathcal{A}_n^\star, \|\cdot\|_2\right) \\ &= N\left(\frac{C_5\epsilon_n}{36G}, \left\{\boldsymbol{\beta} \in \mathcal{A}_n, \|\boldsymbol{\beta} - \boldsymbol{\beta}_0\|_2 \leq G^M\right\}, \|\cdot\|_2\right). \quad \text{(B.16)} \end{aligned}$$

Denote by $\widetilde{\omega}$,

$$\widetilde{\omega} = \frac{1}{\lambda_0 - \lambda_1} \log\left[\frac{1-\theta}{\theta}\frac{\lambda_0^{m_{\max}}}{\lambda_1^{m_{\max}}}\right],$$

It is clear that given our choice of hyperparameters for $(\theta, \lambda_0, \lambda_1)$, the threshold ω_g in (21) is an increasing function of m_g. Therefore,

$$\begin{aligned} &\{\boldsymbol{\beta} \in \mathcal{A}_n, \|\boldsymbol{\beta} - \boldsymbol{\beta}_0\|_2 \leq G^M\} \\ &\quad \subset \left\{\boldsymbol{\beta} \in \mathbb{R}^p : \|\boldsymbol{\beta}_{S_0} - \boldsymbol{\beta}_{0S_0}\|_2 \leq G^M, \text{ and } \|\boldsymbol{\beta}_g\|_2 \leq \widetilde{\omega} \text{ for all } g \in S_0^c\right\} \\ &\quad \subset \left\{\|\boldsymbol{\beta}_{S_0} - \boldsymbol{\beta}_{0S_0}\|_2 \leq G^M\right\} \times [-\widetilde{\omega}, \widetilde{\omega}]^{G - K_1 s_0}. \end{aligned}$$

By Lemma S4 of Shen et al. (2024), for any set of the form $E = A \times [-\delta, \delta]^{Q-s} \subset \mathbb{R}^Q$ where $A \subset \mathbb{R}^s$ and $s < Q$,

$$D(\epsilon, E, \|\cdot\|_2) \leq D((1 - T^{-1})^{1/2}\epsilon, A, \|\cdot\|_2),$$

if $\delta < \epsilon/(2[T(Q-s)]^{1/2})$, for some $T > 1$, where D denotes the packing number. In order to use Lemma A4 of Shen et al. (2024), we can check that for sufficiently large n and $T = 2$,

$$\widetilde{\omega} = \frac{1}{\lambda_0 - \lambda_1} \log\left[\frac{1-\theta}{\theta}\frac{\lambda_0^{m_{\max}}}{\lambda_1^{m_{\max}}}\right] \lesssim \frac{m_{\max}\log G}{G^2} < \frac{C_5\epsilon_n/36G}{2[2T(G-s_0)]^{1/2}},$$

where we used Assumptions (A1)-(A2) and the fact that $\lambda_0 = (1-\theta)/\theta \asymp p^c, c > 2$, and $\lambda_1 \asymp 1/n$. Hence, by Lemma S4 of Shen et al. (2024) and the fact that we can upper bound the covering number by the packing number, we can further upper bound (B.16) by

$$\begin{aligned} &\binom{G}{K_1 s_0} D\left(\frac{C_5\epsilon_n}{36G\sqrt{2}}, \left\{\|\boldsymbol{\beta}_S - \boldsymbol{\beta}_{0S}\|_2 \leq G^M\right\}, \|\cdot\|_2\right) \\ &\quad \leq \binom{G}{K_1 s_0}\left(\frac{108\sqrt{2}G^{M+1}}{C_5\epsilon_n}\right)^{K_1 s_0}, \quad \text{(B.17)} \end{aligned}$$

noting that on the set $\mathcal{A}_n$, $|S| \leq K_1 s_0$. From (B.16)-(B.17), we have that

$$\begin{aligned} \log N(\epsilon_n/36, \mathcal{A}_n^\star, H_n) &\leq K_1 s_0 \log G + K_1 s_0\left[(M+1)\log G + \log(108\sqrt{2n})\right] \\ &\lesssim n\epsilon_n^2, \quad \text{(B.18)} \end{aligned}$$

where we used the fact that $\binom{G}{K_1 s_0} \leq G^{K_1 s_0}$. Having established (B.18), we can now use Lemma 9 of Ghosal and van der Vaart (2007), which implies that for every $\epsilon > \epsilon_n$, there exists a test $\bar{\varphi}_n$ such that

$$\mathbb{E}_0 \bar{\varphi}_n \leq \frac{1}{2} \exp\left(C_6 n\epsilon_n^2 - n\epsilon/2\right), \sup_{\boldsymbol{\beta} \in \mathcal{A}_n^\star : H_n(\boldsymbol{\beta}, \boldsymbol{\beta}_0) > \epsilon} \mathbb{E}_{\boldsymbol{\beta}}(1 - \bar{\varphi}_n) \leq \exp(-n\epsilon^2/2).$$

for some $C_6 > 0$. By Lemma 2, $\Pi(|\boldsymbol{\nu}(\boldsymbol{\beta})| = s_0) \geq e^{-C_3 n\epsilon_n^2}$ for some $C_3 > 0$, and thus,

$$\begin{aligned}
&\mathbb{E}_0 \Pi\left(\boldsymbol{\beta} \in \mathcal{A}_n : H_n(\boldsymbol{\beta}, \boldsymbol{\beta}_0) > \epsilon \mid \mathbf{Y}\right) \\
&\quad \leq \mathbb{E}_0 \Pi\left(\boldsymbol{\beta} \in \mathcal{A}_n : H_n(\boldsymbol{\beta}, \boldsymbol{\beta}_0) > \epsilon \mid \mathbf{Y}\right) \mathbb{1}_{\mathcal{E}_n}(1 - \bar{\varphi}_n) + \mathbb{E}_0 \bar{\varphi}_n \\
&\quad \leq \left\{ \sup_{\boldsymbol{\beta} \in \mathcal{A}_n^\star : H_n(\boldsymbol{\beta}, \boldsymbol{\beta}_0) > \epsilon} \mathbb{E}_{\boldsymbol{\beta}}(1 - \bar{\varphi}_n) + \Pi(\mathcal{A}_n \setminus \mathcal{A}_n^\star) \right\} e^{C_3 n\epsilon_n^2} + \mathbb{E}_0 \bar{\varphi}_n.
\end{aligned}$$

All of the terms in the display except $\Pi(\mathcal{A}_n \setminus \mathcal{A}_n^\star) e^{C_3 n\epsilon_n^2}$ go to zero by choosing $\epsilon = C_4 \epsilon_n$ for a sufficiently large $C_4 > C_3$. To complete the proof then, we must show that

$$\Pi(\mathcal{A}_n \setminus \mathcal{A}_n^\star) e^{C_3 n\epsilon_n^2} \to 0 \text{ as } n \to \infty. \tag{B.19}$$

To establish (B.19), note that

$$\begin{aligned}
\Pi(\mathcal{A}_n \setminus \mathcal{A}_n^\star) &= \Pi(\boldsymbol{\beta} \in \mathbb{R}^p : |\boldsymbol{\nu}(\boldsymbol{\beta})| \leq K_1 s_0, \|\boldsymbol{\beta} - \boldsymbol{\beta}_0\|_2 > G^M) \\
&\leq \Pi(\|\boldsymbol{\beta} - \boldsymbol{\beta}_0\|_2 > G^M) \\
&\leq \sum_{g \in S_0} \Pi\left(\|\boldsymbol{\beta}_g\|_2 > G^M - \|\boldsymbol{\beta}_{0g}\|_2\right) + \sum_{g \in S_0^c} \Pi\left(\|\boldsymbol{\beta}_g\|_2 > G^M\right).
\end{aligned} \tag{B.20}$$

We examine each of the terms in (B.20) separately. Notice that $\pi(\boldsymbol{\beta}_g) \leq \Psi(\boldsymbol{\beta}_g \mid \lambda_1)$ for all $\boldsymbol{\beta}_g$. Moreover, for $\Psi(\boldsymbol{\beta}_g \mid \lambda_1)$, $\|\boldsymbol{\beta}_g\|_2$ follows a gamma distribution with shape parameter m_g and scale parameter λ_1 Letting X_g denote a gamma random variable with shape and scale parameters m_g and λ_1, and recalling that $M \geq 1$, we have

$$\begin{aligned}
\sum_{g \in S_0} \Pi\left(\|\boldsymbol{\beta}_g\|_2 > G^M - \|\boldsymbol{\beta}_{0g}\|_2\right) &\leq \sum_{g \in \mathcal{S}_0} P(X_g > G^M - \|\boldsymbol{\beta}_{0g}\|_2) \\
&\leq \sum_{g \in \mathcal{S}_0} P(X_g > G^M - C_7 \sqrt{G} \log G) \\
&\leq \sum_{g \in S_0} \exp\left[-\lambda_1 (G^M - C_7 \sqrt{G} \log G) + m_{\max}\right] \\
&= \exp\left[-\lambda_1 G^M + \lambda_1 C_7 \sqrt{G} \log G + m_{\max} + \log s_0\right],
\end{aligned} \tag{B.21}$$

where in the second line, we used the fact that by Assumption (B4), $\|\boldsymbol{\beta}_{0g}\|_2 \leq \sqrt{G} \|\boldsymbol{\beta}_0\|_\infty \leq C_7 \sqrt{G} \log G$, for some constant $C_7 > 0$. In the third line of the

display, we used same gamma density tail bound that we used to prove (B.11). Using a similar argument as (B.21), we also have

$$\begin{aligned}\sum_{g\in S_0^c}\Pi\left(\|\boldsymbol{\beta}_g\|_2 > G^M\right) &\le \sum_{g\in S_0^c} P(X_g > G^M)\\ &\le \sum_{g\in S_0^c}\exp\left[-\lambda_1 G^M + m_{\max}\right]\\ &= \exp\left[-\lambda_1 G^M + m_{\max} + \log(G-s_0)\right]. \end{aligned} \tag{B.22}$$

Combining (B.20)-(B.22), we have that for sufficiently large n,

$$\begin{aligned}\Pi(\mathcal{A}_n \setminus \mathcal{A}_n^\star) &\le 2\exp\left[-\lambda_1 G^M + \lambda_1 C_7\sqrt{G}\log G + \log G + m_{\max}\right]\\ &\lesssim \exp(-C_4 s_0 \log G), \end{aligned} \tag{B.23}$$

where we can choose $C_4 > C_3$ so that $-\lambda_1 G^M$ is the dominating term in the first line of the display, and further, $G^M \gg C_4 s_0 n \log G$ when n is large. Thus, from (B.23), we can see that (B.19) holds, and we have established the posterior contraction rate with respect to the average squared Hellinger metric (B.13).

Step 2: Contraction rate for the regression coefficients. We first define the uniform compatibility number ϕ_1 as

$$\phi_1(s;\boldsymbol{\Omega}) = \inf_{\boldsymbol{\beta}:1\le|\boldsymbol{\nu}(\boldsymbol{\beta})|\le s}\frac{\|\boldsymbol{\Omega}^{1/2}\mathbf{X}\boldsymbol{\beta}\|_2|\boldsymbol{\nu}(\boldsymbol{\beta})|^{1/2}}{n^{1/2}\|\boldsymbol{\beta}\|_1},$$

and the smallest scaled singular value ϕ_2 as

$$\phi_2(s;\boldsymbol{\Omega}) = \inf_{\boldsymbol{\beta}:1\le|\boldsymbol{\nu}(\boldsymbol{\beta})|\le s}\frac{\|\boldsymbol{\Omega}^{1/2}\mathbf{X}\boldsymbol{\beta}\|_2}{n^{1/2}\|\boldsymbol{\beta}\|_2}.$$

According to Theorem 3 of Jeong and Ghosal (2021), as long as

$$\frac{s_0^2\log G\|\mathbf{X}\|_{\max}}{\phi_1^2\left((K_2+1)s_0;\boldsymbol{\Omega}(\boldsymbol{\beta}_0)\right)} = o(n), \tag{B.24}$$

then for some $C_8 > 0$,

$$\{\boldsymbol{\beta} : H_n(\boldsymbol{\beta},\boldsymbol{\beta}_0) \le C_4\epsilon_n\} \subset \left\{\boldsymbol{\beta} : \|\boldsymbol{\beta}-\boldsymbol{\beta}_0\|_2 \le \frac{C_8\epsilon_n}{\phi_2\left((K_2+1)s_0;\boldsymbol{\Omega}(\boldsymbol{\beta}_0)\right)}\right\}, \tag{B.25}$$

where $\epsilon_n = (s_0\log G/n)^{1/2}$. Notice that for any set S such that $s := |S| \le (K_2+1)s_0$, we have

$$\begin{aligned}\phi_2(s;\boldsymbol{\Omega}(\boldsymbol{\beta}_0)) &= \inf_{\boldsymbol{\beta}:1\le|\boldsymbol{\nu}(\boldsymbol{\beta})|\le s}\frac{\|\boldsymbol{\Omega}^{1/2}(\boldsymbol{\beta}_0)\mathbf{X}\boldsymbol{\beta}\|_2}{\|\boldsymbol{\beta}\|_2 n^{1/2}}\\ &\ge \frac{\left[\lambda_{\min}\left(\mathbf{X}_S^\top\boldsymbol{\Omega}(\boldsymbol{\beta}_0)\mathbf{X}_S\right)\right]^{1/2}\|\boldsymbol{\beta}\|_2}{\|\boldsymbol{\beta}\|_2 n^{1/2}}\\ &= \left[\lambda_{\min}\left(\frac{1}{n}\mathbf{X}_S^\top\boldsymbol{\Omega}(\boldsymbol{\beta}_0)\mathbf{X}_S\right)\right]^{1/2} \gtrsim 1, \end{aligned} \tag{B.26}$$

by Assumption (B3)(ii). We also have $\|\mathbf{X}\|_{\max} \lesssim \log G$ by Assumption (B3)(i), and $\phi_1(s;\mathbf{\Omega}) \geq \phi_2(s;\mathbf{\Omega})$. It follows from (B.26) that

$$\frac{s_0^2 \log G \|\mathbf{X}\|_{\max}}{\phi_1^2\left((K_2+1)s_0;\mathbf{\Omega}(\boldsymbol{\beta}_0)\right)} \lesssim \frac{s_0^2 (\log G)^2}{\phi_2^2\left((K_2+1)s_0;\mathbf{\Omega}(\boldsymbol{\beta}_0)\right)} \lesssim s_0^2 (\log G)^2.$$

Thus, by Assumption (A2) that $s_0 = o(n^{1/2}/\log G)$, it must be that (B.24) holds. Thus, for some $K_3 > 0$, we have by (B.25) and Theorem 2 that

$$\mathbb{E}_0 \Pi\left(\boldsymbol{\beta} : \|\boldsymbol{\beta} - \boldsymbol{\beta}_0\|_2 \leq K_3 \epsilon_n \mid \mathbf{Y}\right) \geq \Pi(\boldsymbol{\beta} : H_n(\boldsymbol{\beta}, \boldsymbol{\beta}_0) \leq C_4 \epsilon_n) \to 1 \text{ as } n \to \infty,$$

where the smallest scaled singular value $\phi_2((K_2+1)s_0;\mathbf{\Omega}(\boldsymbol{\beta}_0))$ is a positive constant (by (B.26)) that is absorbed into the constant K_3. We are done. □

B.3 Proof of Proposition 1

Throughout, we use P_0 and $\mathbb{E}_0$ to respectively denote the probability and expectation operators under the law induced by the model $\mathbf{Y} \sim \mathcal{N}_n(\boldsymbol{\beta}_0, \mathbf{I}_n)$.

The posterior contraction proof in Proposition 1 follows a similar strategy as the proof of Theorem 7 in Castillo et al. (2015). However, it should be noted that Castillo et al. (2015) deal exclusively with the case where $\boldsymbol{\beta}_0 = (\beta_{01}, \ldots, \beta_{0n})^\top$ is the zero vector, i.e. $\boldsymbol{\beta}_0 = \mathbf{0}_n$, and the *individual* regression coefficients $\beta_{0i}, i = 1, \ldots, n$, are endowed with *univariate* Laplace$(0, \lambda^{-1})$ priors. In contrast, we consider the *grouped* scenario where $\boldsymbol{\beta}_0 = (\boldsymbol{\beta}_{01}^\top, \ldots, \boldsymbol{\beta}_{0G}^\top)^\top$ with $\boldsymbol{\beta}_{01} \neq \mathbf{0}_{m_1}$, and the groups $\boldsymbol{\beta}_{0g}, g = 1, \ldots, G$, are endowed with independent *multivariate* Laplace priors (9).

Proof of Proposition 1 We first prove (25). Let $\mathbf{Y} = (\mathbf{Y}_1^\top, \ldots, \mathbf{Y}_G^\top)^\top$, where the indices of $\mathbf{Y}_g$ correspond to those of $\boldsymbol{\beta}_{0g}$ in $\boldsymbol{\beta}_0 = (\boldsymbol{\beta}_{01}^\top, \ldots, \boldsymbol{\beta}_{0G}^\top)^\top$ for all $g = 1, \ldots, G$. When $\mathbf{Y} \sim \mathcal{N}_n(\boldsymbol{\beta}_0, \mathbf{I}_n)$, the group lasso estimator has a closed form,

$$\widehat{\boldsymbol{\beta}}_g = \left(1 - \frac{\lambda}{\|\mathbf{Y}_g\|_2}\right)_+ \mathbf{Y}_g, \quad \text{for all} \quad g = 1, \ldots, G,$$

where $x_+ = \min\{x, 0\}$. Therefore, we must have $\|\widehat{\boldsymbol{\beta}}_g\|_2 < \|\mathbf{Y}_g\|_2$ for all $g \in \{1, \ldots, G\}$. We can decompose the expected squared ℓ_2 risk as

$$\mathbb{E}_0\|\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0\|_2^2 = \mathbb{E}_0\|\widehat{\boldsymbol{\beta}}_1 - \boldsymbol{\beta}_{01}\|_2^2 + \sum_{g=2}^{G} \mathbb{E}_0\|\widehat{\boldsymbol{\beta}}_g\|_2^2, \tag{B.27}$$

We first upper-bound the first term in (B.27). Since $\mathbf{Y}_1 \sim \mathcal{N}_m(\boldsymbol{\beta}_{01}, \mathbf{I}_m)$, $\|\widehat{\boldsymbol{\beta}}_1\|_2 < \|\mathbf{Y}_1\|_2$, and $\widehat{\boldsymbol{\beta}}_1 = \mathbf{0}_m$ when $\|\mathbf{Y}_1\|_2 \leq \lambda$, we have

$$\begin{aligned}
\mathbb{E}_0\|\widehat{\boldsymbol{\beta}}_1 - \boldsymbol{\beta}_{01}\|_2^2 &\leq 2\mathbb{E}_0\|\mathbf{Y}_1 - \boldsymbol{\beta}_{01}\|_2^2 + 2\mathbb{E}_0\|\widehat{\boldsymbol{\beta}}_1 - \mathbf{Y}_1\|_2^2 \\
&\leq 2m + 2\mathbb{E}_0\{\|\mathbf{Y}_1\|_2^2 \mathbb{I}(\|\mathbf{Y}_1\|_2 \leq \lambda)\} + 2\mathbb{E}_0\{\lambda^2 \mathbb{I}(\|\mathbf{Y}_1\|_2 > \lambda)\} \\
&\leq 2m + 2\lambda^2 + 2\lambda^2 P_0(\|\mathbf{Y}_1\|_2 > \lambda) \\
&\leq 2m + 4\lambda^2 \asymp 2\log n,
\end{aligned} \tag{B.28}$$

since m is constant and $\lambda \asymp \sqrt{2\log n}$. Next, we upper-bound the second term in (B.27). Again, using the fact that $\|\widehat{\boldsymbol{\beta}}_g\|_2 < \|\mathbf{Y}_g\|_2$ and $\widehat{\boldsymbol{\beta}}_g = \mathbf{0}_m$ when $\|\mathbf{Y}_g\|_2 \leq \lambda$, each of the summands can be upper-bounded as

$$\mathbb{E}_0\|\widehat{\boldsymbol{\beta}}_g\|_2^2 \leq \mathbb{E}_0\{\|\mathbf{Y}_g\|_2^2 \mathbb{I}(\|\mathbf{Y}_g\|_2^2 > \lambda^2)\}.$$

Since $\mathbf{Y}_g \sim \mathcal{N}_m(\mathbf{0}_m, \mathbf{I}_m)$ for all $g \neq 1$, $\|\mathbf{Y}_g\|_2^2 \sim \chi_m^2$. Letting $\Gamma(a,x) = \int_x^\infty t^{a-1}e^{-t}dt$ denote the incomplete gamma function, we proceed to upper-bound the individual summands as

$$\begin{aligned}
\mathbb{E}_0\|\widehat{\boldsymbol{\beta}}_g\|_2^2 &\leq \int_{\lambda^2}^\infty \left[\Gamma\left(\frac{m}{2}\right)\right]^{-1} \left(\frac{u}{2}\right)^{m/2} e^{-u/2}du \\
&= 2\left[\Gamma(\frac{m}{2})\right]^{-1} \Gamma\left(\frac{m}{2}+1, \lambda^2\right) \\
&\leq 2\left[\Gamma\left(\frac{m}{2}\right)\right]^{-1} \left(\frac{m}{2}+1\right) e^{-\lambda^2}(\lambda^2)^{m/2} \quad \text{for large } n \\
&\asymp n^{-2}(\log n)^{m/2}, \qquad \text{(B.29)}
\end{aligned}$$

where we used the bound $\Gamma(a,x) \leq ae^{-x}x^{a-1}$ when $a \geq 1$ and $x > a$ (Pinelis, 2020) in the third line, and the fact that m is fixed and $\lambda \asymp \sqrt{2\log n}$ in the last line. Combining (B.27)-(B.29) gives

$$\mathbb{E}_0\|\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0\|_2^2 \lesssim 2\log n + (G-1)n^{-2}(\log n)^{m/2} = 2\log n + o(1) \quad \text{as } n \to \infty,$$

where we used the fact that $G = n/m$ and m is a constant. This completes the proof of (25).

Next, we prove (26). We follow the proof strategy of Theorem 7 in Castillo et al. (2015). According to Lemma 7.1 of Castillo and van der Vaart (2012), $\mathbb{E}_0\Pi(\boldsymbol{\beta} : \|\boldsymbol{\beta} - \boldsymbol{\beta}_0\|_2 < s_n \mid \mathbf{Y}) \to 0$ for any s_n for which there exists a sequence r_n such that

$$\frac{\Pi(\boldsymbol{\beta} : \|\boldsymbol{\beta} - \boldsymbol{\beta}_0\|_2 < s_n)}{\Pi(\boldsymbol{\beta} : \|\boldsymbol{\beta} - \boldsymbol{\beta}_0\|_2 < r_n)} e^{2r_n^2} = o(1). \qquad \text{(B.30)}$$

Since the prior is $\boldsymbol{\beta}_g \overset{iid}{\sim} \Psi(\boldsymbol{\beta}_g \mid \lambda)$ for all $g = 1, \ldots, G$, $\|\boldsymbol{\beta}_1\|_2 \sim \text{Gamma}(m, \lambda)$ and $\sum_{g=1}^G \|\boldsymbol{\beta}_g\|_2 \sim \text{Gamma}(n, \lambda)$ where $n = mG$.

We first upper-bound the numerator in (B.30). Since $\|\boldsymbol{\beta}_0\|_2 = \|\boldsymbol{\beta}_{01}\|_2 = L$ and $\|\boldsymbol{\beta} - \boldsymbol{\beta}_0\|_2 \geq \|\boldsymbol{\beta}_1 - \boldsymbol{\beta}_{01}\|_2$, we have that for sufficiently large n such that $s_n > L$,

$$\begin{aligned}
\Pi(\boldsymbol{\beta} : \|\boldsymbol{\beta} - \boldsymbol{\beta}_0\|_2 < s_n) &\leq \Pi(\boldsymbol{\beta}_1 : \|\boldsymbol{\beta}_1 - \boldsymbol{\beta}_0\|_2 < s_n) \\
&\leq \Pi\left(\boldsymbol{\beta}_1 : \left|\|\boldsymbol{\beta}_1\|_2 - \|\boldsymbol{\beta}_{01}\|_2\right| < s_n\right) \\
&= \Pi(\boldsymbol{\beta}_1 : \|\boldsymbol{\beta}_1\|_2 < s_n + L) \\
&= \int_0^{s_n+L} \frac{\lambda^m}{\Gamma(m)} u^{m-1}e^{-\lambda u}du \\
&\leq \frac{[\lambda(s_n+L)]^m}{\Gamma(m+1)}. \qquad \text{(B.31)}
\end{aligned}$$

Next, we lower-bound the denominator in (B.30). For sufficiently large n so that $r_n > L$, we have

$$\begin{aligned}\Pi(\boldsymbol{\beta} : \|\boldsymbol{\beta} - \boldsymbol{\beta}_0\|_2 < r_n) &\geq \Pi(\boldsymbol{\beta} : \|\boldsymbol{\beta}\|_2 + \|\boldsymbol{\beta}_0\|_2 < r_n) \\ &= \Pi(\boldsymbol{\beta} : \|\boldsymbol{\beta}\|_2 < r_n - L) \\ &\geq \Pi\left(\boldsymbol{\beta} : \sum_{g=1}^{G} \|\boldsymbol{\beta}_g\|_2 < r_n - L\right) \\ &= \int_0^{r_n - L} \frac{\lambda^n}{\Gamma(n)} u^{n-1} e^{-\lambda u} du \\ &\geq \frac{\lambda^n}{n\Gamma(n)} e^{-\lambda(r_n - L)} (r_n - L)^n. \end{aligned} \tag{B.32}$$

Combining (B.30)-(B.32) gives

$$\begin{aligned}\frac{\Pi(\boldsymbol{\beta} : \|\boldsymbol{\beta} - \boldsymbol{\beta}_0\|_2 < s_n)}{\Pi(\boldsymbol{\beta} : \|\boldsymbol{\beta} - \boldsymbol{\beta}_0\|_2 < r_n)} &\leq n\lambda^{m-n} \frac{\Gamma(n)}{\Gamma(m+1)} \frac{(s_n + L)^m}{(r_n - L)^n} e^{\lambda(r_n - L)} \\ &\lesssim e^{\lambda\sqrt{n} r_n} \left(\frac{s_n}{r_n}\right)^n \\ &= \exp\left[\lambda\sqrt{n} r_n - n \log\left(\frac{r_n}{s_n}\right)\right]. \end{aligned} \tag{B.33}$$

Setting $r_n = \sqrt{n}(\lambda^{-1} \wedge 1)$ and $s_n = \delta r_n$ for δ small enough in (B.33) satisfies (B.30), and thus, by Lemma 7.1 of Castillo and van der Vaart (2012),

$$\mathbb{E}_0 \Pi\left(\boldsymbol{\beta} : \|\boldsymbol{\beta} - \boldsymbol{\beta}_0\|_2 < \delta\sqrt{n}\left(\frac{1}{\lambda} \wedge 1\right) \;\middle|\; \mathbf{Y}\right) \to 0 \;\text{ as }\; n \to \infty.$$

By setting $\lambda \asymp \sqrt{2 \log n}$ in the above display, we obtain the posterior contraction rate of $\sqrt{n/\log n}$ in (26). □

C Additional details for the EM algorithm in Section 5.1

C.1 Derivation of the EM algorithm

Recall that the log-likelihood function $\ell_n(\boldsymbol{\beta})$ is given in (4). With the reparameterization of the hierarchical SSGL prior in (30) and the $\mathcal{B}(a, b)$ prior (27) on θ, we can write the log-posterior $\log \pi(\boldsymbol{\beta}, \theta \mid \mathbf{Y})$ in (28) as

$$\begin{aligned}\log \pi(\boldsymbol{\beta}, \theta \mid \mathbf{Y}) = \ell_n(\boldsymbol{\beta}) &+ \sum_{g=1}^{G} \log\left((1 - \gamma_g)\lambda_0^{m_g} e^{-\lambda_0 \|\boldsymbol{\beta}_g\|_2} + \gamma_g \lambda_1^{m_g} e^{-\lambda_1 \|\boldsymbol{\beta}_g\|_2}\right) \\ &+ \left(a - 1 + \sum_{g=1}^{G} \gamma_g\right) \log\theta + \left(b - 1 + p - \sum_{g=1}^{G} \gamma_g\right) \log(1 - \theta). \end{aligned} \tag{C.1}$$

From (C.1), it is straightforward to verify that $\mathbb{E}[\gamma_g \mid \mathbf{Y}, \boldsymbol{\beta}, \theta] = p_g^\star(\boldsymbol{\beta}_g, \theta)$, where $p_g^\star(\boldsymbol{\beta}_g, \theta)$ is as in (31).

In the E-step of our EM algorithm, we treat $\boldsymbol{\gamma}$ as missing data and take expectation with respect to γ_g for all $g = 1, \ldots, G$, holding the parameters $(\boldsymbol{\beta}, \theta)$ fixed at their previous values. That is, we compute $p_g^{\star(t-1)} := p^\star(\boldsymbol{\beta}_g^{(t-1)}, \theta^{(t-1)}) = \mathbb{E}[\gamma_g \mid \mathbf{Y}, \boldsymbol{\beta}^{(t-1)}, \theta^{(t-1)}]$ for all $g = 1, \ldots, G$, given the previous estimates $(\boldsymbol{\beta}^{(t-1)}, \theta^{(t-1)})$.

For the M-step, we then maximize the following function:

$$\mathbb{E}[\log \pi(\boldsymbol{\beta}, \theta \mid \mathbf{Y}) \mid \boldsymbol{\beta}^{(t-1)}, \theta^{(t-1)}] = \ell_n(\boldsymbol{\beta}) - \sum_{g=1}^{G} \lambda_g^{\star(t-1)} \|\boldsymbol{\beta}_g\|_2 + \left(a - 1 + \sum_{g=1}^{G} p_g^{\star(t-1)}\right) \log\theta + \left(b - 1 + G - \sum_{g=1}^{G} p_g^{\star(t-1)}\right) \log(1-\theta), \tag{C.2}$$

where $\lambda_g^{\star(t-1)} = \lambda_1 p_g^{\star(t-1)} + \lambda_0(1 - p_g^{\star(t-1)})$. From (C.2), it is easy to derive the update for θ in (32) by taking the derivative of (C.2) with respect to θ. The update for $\boldsymbol{\beta}$ in (33) is obtained by isolating the terms on the right-hand side of (C.2) that only depend on $\boldsymbol{\beta}$. We discuss the M-step for updating $\boldsymbol{\beta}$ in detail in the next section.

C.2 M-step for updating the regression coefficients

In the M-step of our EM algorithm, we need to solve the optimization problem (33), or equivalently,

$$\boldsymbol{\beta}^{(t)} = \arg\min_{\beta} \left\{ -\ell_n(\boldsymbol{\beta}) + \sum_{g=1}^{G} \lambda_g^{\star(t-1)} \|\boldsymbol{\beta}_g\|_2 \right\}, \tag{C.3}$$

where $-\ell_n(\boldsymbol{\beta})$ is the negative of the log-likelihood in (4). While (C.3) may appear to be intractable, we can apply the standard iteratively reweighted least squares (IRLS) algorithm (McCullagh and Nelder, 1989) for GLMs to efficiently solve (C.3).

The IRLS algorithm in GLMs is based on a quadratic approximation of the negative log-likelihood. Denote the mean response as $\mu_i = b'(\theta_i)$, and let $V(\mu_i) = b''((b')^{-1}(\mu_i))$ be the variance function, where b is the cumulant function in (1). With the link function h in (3), we define the "working response" vector $\mathbf{Z}$ at the kth iteration of the optimization problem (C.3) as $\mathbf{Z} = \mathbf{X}\boldsymbol{\beta}^{(k)} + \boldsymbol{\zeta}^{(k)}$, where $\zeta_i^{(k)} = h'(\mu_i^{(k)})(y_i - \mu_i^{(k)})$, $i = 1, \ldots, n$. Similarly, we define the weights matrix $\mathbf{W} = \mathrm{diag}(w_1^{(k)}, \ldots, w_n^{(k)})$, where the weights are $w_i^{(k)} = [V(\mu_i^{(k)})(h'(\mu_i^{(k)}))^2]^{-1}$. As shown in McCullagh and Nelder (1989), the negative log-likelihood for any GLM in the exponential dispersion family (1.1) can then be approximated as

$$-\ell_n(\boldsymbol{\beta}) \approx \frac{1}{2}(\mathbf{Z} - \mathbf{X}\boldsymbol{\beta})^\top \mathbf{W}(\mathbf{Z} - \mathbf{X}\boldsymbol{\beta}). \tag{C.4}$$

Thus, substituting the approximation (C.4) for $-\ell_n(\boldsymbol{\beta})$ in (C.3), we see that the M-step (C.3) for $\boldsymbol{\beta}$ can alternatively be written as

$$\boldsymbol{\beta}^{(t)} = \arg\min_{\boldsymbol{\beta}} \left\{ \frac{1}{2}(\mathbf{Z} - \mathbf{X}\boldsymbol{\beta})^\top \mathbf{W}(\mathbf{Z} - \mathbf{X}\boldsymbol{\beta}) + \sum_{g=1}^{G} \lambda_g^{\star(t-1)} \|\boldsymbol{\beta}_g\|_2 \right\}. \tag{C.5}$$

With the quadratic approximation to the negative log-likelihood, (C.5) is now a standard linear regression model with a group lasso penalty (Yuan and Lin, 2006) and group-specific weights $\lambda_g^{\star(t-1)}$ for each gth group. Solving (C.5) can be done with any standard block coordinate descent algorithm for the regular group lasso (Yuan and Lin, 2006) in linear regression. In particular, if the canonical link function $h = (b')^{-1}$ is used, then the Fisher information matrix and the negative Hessian matrix are equal, and we use the majorization-minimization (MM) algorithm in Breheny and Huang (2015) to solve (C.5). On the other hand, if a non-canonical link function is used, then we use the least squares approximation (LSA) approach of Wang and Leng (2007) with a group lasso penalty.

The SSGL MAP algorithm is on average slower than the numerical algorithms used for group lasso, group SCAD, or group MCP. This is because *each* iteration of the EM algorithm requires numerically solve the optimization problem (C.3) in the M-step. Therefore, if the EM algorithm takes B total iterations to converge, we expect the average runtime of the SSGL method to be on the order of B times slower than that for group lasso, group SCAD, and group MCP. However, we observed that our EM algorithm tended to converge within a small number of iterations (typically fewer than 10), at least in all of the simulation settings we considered. Thus, the computation time for SSGL for fixed λ_0 was still reasonably quick. For example, in Simulations 4 and 8 of the main manuscript where $p = 1600$, the SSGL runtime for any fixed λ_0 was under 30 seconds in all replications.

To further accelerate the computational efficiency for SSGL, we could employ alternative optimization strategies that obviate the need to introduce latent indicator variables. For example, we could potentially approximate the SSGL prior with a neuronized prior (Shin and Liu, 2022), which is a product of a Gaussian weight variable and a scale variable transformed from Gaussian via an activation function. This approximation would allow for a fast coordinate ascent algorithm which avoids the need to marginalize out latent indicator variables (Shin and Liu, 2022). We leave the investigation of more computationally efficient algorithms for SSGL for future work.

D Complete Gibbs sampling algorithms for Section 5.3

Recall that $\boldsymbol{\beta} = (\boldsymbol{\beta}_1^\top, \ldots, \boldsymbol{\beta}_G^\top)^\top$, where $\boldsymbol{\beta}_g \in \mathbb{R}^{m_g}, g = 1, \ldots, G$, and $\boldsymbol{\tau} = (\tau_1, \ldots, \tau_G)^\top$ is defined in (5.14). Let $\mathcal{PG}(a, b)$ denote the Pólya-gamma distribution (Polson et al., 2013), and let $\mathcal{GIG}(a, b, c)$ denote the generalized inverse

Gaussian distribution with density function $f(x; a, b, c) \propto x^{a-1}e^{-(b/x+cx)/2}$. We define $\boldsymbol{\kappa} = (\kappa_1, \ldots, \kappa_n)^\top$, where for $i = 1, \ldots, n$,

$$\kappa_i = \begin{cases} y_i - 0.5 & \text{for logistic regression,} \\ 0.5(y_i - M) & \text{for Poisson regression.} \end{cases} \tag{D.1}$$

We also define $\boldsymbol{\omega} = (\omega_1, \ldots, \omega_n)^\top$, where for $i = 1, \ldots, n$,

$$\begin{cases} \omega_i \sim \mathcal{PG}(1, \mathbf{x}_i^\top \boldsymbol{\beta}) & \text{for logistic regression,} \\ \omega_i \sim \mathcal{PG}(M, \mathbf{x}_i^\top \boldsymbol{\beta} - \log M) & \text{for Poisson regression.} \end{cases} \tag{D.2}$$

Finally, define the vector $\mathbf{Z} = (z_1, \ldots, z_n)^\top$, where for $i = 1, \ldots, n$,

$$z_i = \begin{cases} \kappa_i/\omega_i & \text{for logistic regression,} \\ \kappa_i/\omega_i + \log M & \text{for Poisson regression,} \end{cases} \tag{D.3}$$

and $\boldsymbol{\kappa}$ is defined as in (D.1).

The Gibbs sampler for grouped logistic or Poisson regression with the SSGL prior (5.4) is given in Algorithm 1.

When $p = \sum_{g=1}^{G} m_g$ is larger than n, we can use the fast sampling algorithm of Bhattacharya et al. (2016) to sample $\boldsymbol{\beta}^{(t)}$ in Step 4 of Algorithm 1 in $\mathcal{O}(n^2 p)$ time instead of $\mathcal{O}(p^3)$ time. This sampling algorithm is given in Algorithm 2.

E Simulation studies for negative binomial regression with a log link

Since the mean and variance are equal for the Poisson distribution, Poisson regression may be inappropriate for modeling count data that exhibit overdispersion. In this case, it may be more appropriate to use negative binomial regression. That is, we assume that $y_i \sim NB(\alpha, \mu_i), i = 1, \ldots, n$, where $\alpha > 0$ is a size parameter, and the probability density function (pdf) for each y_i is

$$f(y_i \mid \alpha, \mu_i) = \frac{\Gamma(y_i + \alpha)}{y_i! \Gamma(\alpha)} \left(\frac{\mu_i}{\mu_i + \alpha}\right)^{y_i} \left(\frac{\alpha}{\mu_i + \alpha}\right)^{\alpha}, \quad y_i = 0, 1, 2, \ldots$$

Since $\text{Var}(y_i) = \mu_i + \mu_i^2/\alpha$, the negative binomial distribution is better equipped to handle overdispersed count data than the Poisson distribution.

For negative binomial distributed responses, the natural parameter is $\theta_i = \log(\mu_i/(\mu_i + \alpha))$. With the link function $h(u) = \log(u)$, we have $b(u) = -\alpha \log(1 - e^u)$ in (1). It is readily seen that $\mathbb{E}(y_i) = b'(\theta_i) = \mu_i$. For the function $\xi = (h \circ b')^{-1}$, we have $\xi(u) = -\log(\alpha e^{-u} + 1)$. Since $\xi(u) \neq u$, negative binomial regression with the log link is an example of a GLM with a *non*-canonical link function.

We now present simulation results in grouped negative binomial regression for the SSGL model (10). We compared our results to group lasso (GL), adaptive group lasso (AdGL), group SCAD (GSCAD), and group MCP (gMCP). We used

Algorithm 1 MCMC algorithm for grouped logistic and grouped Poisson regression with the SSGL prior

Input: Initial values $\boldsymbol{\beta}^{(0)}$, $\theta^{(0)}$, $\boldsymbol{\tau}^{(0)}$, T (number of MCMC samples to run), B (number of samples to discard as burn-in)

Output: MCMC samples for regression coefficients $\boldsymbol{\beta}$

for $t = 1, \ldots, T$ **do**

1. **for** $i = 1, \ldots, n$ **do**
   i. Sample $\boldsymbol{\omega}^{(t)}$ as in (D.2) and set $\boldsymbol{\Omega} = \text{diag}\{\boldsymbol{\omega}^{(t)}\}$
   ii. Update $\mathbf{Z}^{(t)}$ as in (D.3)

2. **for** $g = 1, \ldots, G$ **do**
   i. Sample $\gamma_g^{(t)} \sim \text{Bernoulli}\left(\frac{\pi_1}{\pi_1 + \pi_0}\right)$, where
   $\pi_1 = \theta^{(t-1)} \lambda_1^{m_g+1} \exp\left(-\lambda_1^2 \tau_g^{(t-1)}/2\right)$
   and $\pi_0 = \left(1 - \theta^{(t-1)}\right) \lambda_0^{m_g+1} \exp\left(-\lambda_0^2 \tau_g^{(t-1)}/2\right)$
   ii. Sample $\tau_g^{(t)} \sim \mathcal{GIG}\left(\frac{1}{2}, \|\boldsymbol{\beta}_g^{(t-1)}\|_2^2, (\lambda_g^\star)^2\right)$, where
   $\lambda_g^\star = \gamma_g^{(t)} \lambda_1 + (1 - \gamma_g^{(t)}) \lambda_0$

3. Sample $\theta^{(t)} \sim \text{Beta}\left(a + \sum_{g=1}^G \gamma_g^{(t)},\ b + G - \sum_{g=1}^G \tau_g^{(t)}\right)$

4. Sample $\boldsymbol{\beta}^{(t)} \sim \mathcal{N}\left(\boldsymbol{\mu}_{\boldsymbol{\beta}}, \boldsymbol{\Sigma}_{\boldsymbol{\beta}}\right)$, where $\boldsymbol{\mu}_{\boldsymbol{\beta}} = \boldsymbol{\Sigma}_{\boldsymbol{\beta}} \mathbf{X}^\top \boldsymbol{\Omega} \mathbf{Z}$,
   $\boldsymbol{\Sigma}_{\boldsymbol{\beta}} = (\mathbf{X}^\top \boldsymbol{\Omega} \mathbf{X} + \mathbf{D}_{\boldsymbol{\tau}}^{-1})^{-1}$, and $\mathbf{D}_{\boldsymbol{\tau}} = \text{Bdiag}\left(\tau_1^{(t)} \mathbf{I}_{m_1}, \ldots, \tau_G^{(t)} \mathbf{I}_{m_G}\right)$

return $\boldsymbol{\beta}^{(B+1)}, \ldots, \boldsymbol{\beta}^{(T)}$

the same methods for choosing the hyperparameters, regularization parameters, and weights as those described in Section 6.1 of the main article. For our simulation study, we fixed $\alpha = 1$ and generated the responses independently as $y_i \mid \mathbf{x}_i \sim NB(\alpha, \exp(\theta_i)), i = 1, \ldots, n$. We conducted the following simulation studies:

> **Experiment 1**. We set $n = 500$ and $G = 30$. The design matrix $\mathbf{X}$ and the regression coefficients $\boldsymbol{\beta}$ were generated the same way as in Experiment 1 in Section 6.3 of the main manuscript. Then we modeled
>
> $$\log(\theta) = \mathbf{x}^\top \boldsymbol{\beta}.$$
>
> **Experiment 2**. We set $n = 500$ and $G = 30$. We generated the entries of the $n \times G$ design matrix $\mathbf{X}$ from independent Uniform$(-1, 1)$ random

Algorithm 2 Algorithm for sampling $\boldsymbol{\beta}$ in Step 4 of Algorithm 1 when $p > n$

Input: Most recent updates of $\boldsymbol{\Omega}$, $\mathbf{D}_{\boldsymbol{\tau}}$, and $\mathbf{Z}$

Output: An exact sample of $\boldsymbol{\beta}$ in Step 4 of Algorithm 1

1. Set $\widetilde{\mathbf{X}} = \boldsymbol{\Omega}^{1/2}\mathbf{X}$
2. Sample $\mathbf{m} \sim \mathcal{N}(\mathbf{0}, \boldsymbol{D}_{\boldsymbol{\tau}})$ and $\boldsymbol{\delta} \sim \mathcal{N}(\mathbf{0}, \mathbf{I}_n)$ independently
3. Set $\mathbf{v} = \widetilde{\mathbf{X}}\mathbf{m} + \boldsymbol{\delta}$ and $\mathbf{v}^\star = \boldsymbol{\Omega}^{1/2}\mathbf{Z} - \mathbf{v}$
4. Set $\mathbf{A} = \widetilde{\mathbf{X}}\mathbf{D}_{\boldsymbol{\tau}}\widetilde{\mathbf{X}}^\top + \mathbf{I}_n$
5. Set $\mathbf{w} = \mathbf{A}^{-1}\boldsymbol{v}^\star$
6. Set $\boldsymbol{\beta} = \mathbf{m} + \boldsymbol{D}_{\boldsymbol{\tau}}\widetilde{\mathbf{X}}^\top\mathbf{w}$

return $\boldsymbol{\beta}$

variables. Then we modeled

$$\log(\theta) = 1.5\sin(3x_1) - x_5 e^{0.5x_5^2}.$$

We represented each covariate as a six-term B-spline basis expansion.

We repeated each experiment 200 times. Table 1 reports our results averaged across the 200 replications. In Experiment 1, SSGL had the lowest MSE and the second lowest MSPE. Meanwhile, all five methods had a TPR of 1, indicating that all methods had equally high power to detect the significant groups in this particular setting. However, SSGL had the highest average TNR and precision, indicating superior overall group selection performance.

In Experiment 2, SSGL also had the lowest MSPE, indicating superior average prediction on out-of-sample data. In this experiment, all five methods once again had a TPR of 1. However, SSGL had the highest average TNR and precision, indicating that SSGL was the best suited for function selection in this particular generalized additive model.

References

Bai, R., Moran, G. E., Antonelli, J. L., Chen, Y., and Boland, M. R. (2022). Spike-and-slab group lassos for grouped regression and sparse generalized additive models. *Journal of the American Statistical Association*, 117(537):184–197.

Bhattacharya, A., Chakraborty, A., and Mallick, B. K. (2016). Fast sampling with Gaussian scale mixture priors in high-dimensional regression. *Biometrika*, 103(4):985–991.

Table 1: Simulation results for grouped negative binomial regression under the SSGL, GL, AdGL, GMCP, and GSCAD models, averaged across 200 replicates. The empirical standard error is reported in parentheses below the average.

Experiment 1

	MSE	MSPE	TPR	TNR	Prec
SSGL	**0.008**	30.04	**1**	**0.999**	**0.999**
	(0.006)	(10.56)	(0)	(0.007)	(0.029)
GL	0.029	**29.43**	**1**	0.387	0.255
	(0.008)	(9.52)	(0)	(0.146)	(0.051)
AdGL	0.018	31.09	**1**	0.809	0.562
	(0.008)	(12.32)	(0)	(0.119)	(0.180)
GMCP	0.012	31.15	**1**	0.892	0.705
	(0.005)	(15.98)	(0)	(0.094)	(0.192)
GSCAD	0.013	31.31	**1**	0.702	0.434
	(0.005)	(14.59)	(0)	(0.133)	(0.130)

Experiment 2

	MSPE	TPR	TNR	Prec
SSGL	**17.43**	**1**	**1**	**1**
	(5.09)	(0)	(0)	(0)
GL	18.35	1	0.383	0.114
	(5.85)	(0)	(0.184)	(0.044)
AdGL	18.16	**1**	0.759	0.296
	(5.47)	(0)	(0.140)	(0.186)
GMCP	17.52	**1**	0.921	0.695
	(5.10)	(0)	(0.123)	(0.317)
GSCAD	17.49	1	0.790	0.363
	(5.33)	(0)	(0.160)	(0.236)

Boucheron, S., Lugosi, G., and Massart, P. (2013). *Concentration Inequalities: A Nonasymptotic Theory of Independence*. Oxford University Press.

Breheny, P. and Huang, J. (2015). Group descent algorithms for nonconvex penalized linear and logistic regression models with grouped predictors. *Statistics and Computing*, 25(2):173–187.

Castillo, I., Schmidt-Hieber, J., and van der Vaart, A. (2015). Bayesian linear regression with sparse priors. *The Annals of Statistics*, 43(5):1986 – 2018.

Castillo, I. and van der Vaart, A. (2012). Needles and straw in a haystack: Posterior concentration for possibly sparse sequences. *The Annals of Statistics*, 40(4):2069 – 2101.

Fan, J. and Lv, J. (2011). Nonconcave penalized likelihood with NP-dimensionality. *IEEE Transactions on Information Theory*, 57(8):5467–5484.

Ghosal, S. and van der Vaart, A. (2007). Convergence rates of posterior distributions for noniid observations. *The Annals of Statistics*, 35(1):192 – 223.

Jeong, S. and Ghosal, S. (2021). Posterior contraction in sparse generalized linear models. *Biometrika*, 108(2):367–379.

McCullagh, P. and Nelder, J. (1989). *Generalized Linear Models, Second Edition*. Chapman & Hall, New York.

Ning, B., Jeong, S., and Ghosal, S. (2020). Bayesian linear regression for multivariate responses under group sparsity. *Bernoulli*, 26(3):2353 – 2382.

Pinelis, I. (2020). Exact lower and upper bounds on the incomplete gamma function. *Mathematical Inequalities & Applications*, 23(4):1261–1278.

Polson, N. G., Scott, J. G., and Windle, J. (2013). Bayesian inference for logistic models using Pólya–gamma latent variables. *Journal of the American Statistical Association*, 108(504):1339–1349.

Shen, Y., Solís-Lemus, C., and Deshpande, S. K. (2024). Estimating sparse direct effects in multivariate regression with the spike-and-slab lasso. *Bayesian Analysis (to appear)*.

Shin, M. and Liu, J. S. (2022). Neuronized priors for Bayesian sparse linear regression. *Journal of the American Statistical Association*, 117(540):1695–1710.

van der Vaart, A. W. and Wellner, J. A. (2006). *Weak Convergence and Empirical Process: With Applications to Statistics*. Springer.

Wang, H. and Leng, C. (2007). Unified lasso estimation by least squares approximation. *Journal of the American Statistical Association*, 102:1039–1048.

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