Bayesian Group Regularization in Generalized Linear Models with a Continuous Spike-and-Slab Prior [*]

Ray Bai[†]

June 30, 2023

Abstract

We study Bayesian group-regularized estimation in high-dimensional generalized linear models (GLMs) under a continuous spike-and-slab prior. Our framework covers both canonical and non-canonical link functions and subsumes logistic regression, Poisson regression, Gaussian regression, and negative binomial regression with group sparsity. Under milder assumptions than those previously assumed for the group lasso, we obtain the convergence rate for both the maximum *a posteriori* (MAP) estimator *and* the full posterior distribution. Our theoretical results thus justify the use of the posterior mode as a point estimator. Furthermore, the posterior distribution contracts at the same rate as the MAP estimator, an attractive feature of our approach which is not the case for the group lasso. For computation, we propose an expectation-maximization (EM) algorithm for rapidly obtaining MAP estimates under our model. We illustrate our method through simulations and a real data application on predicting human immunodeficiency virus (HIV) drug resistance from protein sequences.

1 Introduction

1.1 Motivation

Generalized linear models (GLMs) [21] are widely used in practice and provide a unified way to model both continuous and discrete responses given a set of covariates. Suppose that we observe n independent observations $\{(\mathbf{x}_i, y_i)\}_{i=1}^n$, where $\mathbf{x}_i = (x_{i1}, \ldots, x_{ip})^\top \in \mathbb{R}^p$ denotes a vector of p covariates for the ith observation. GLMs assume that the response variable y_i belongs to the exponential family with density,

$$f_i(y_i) = \exp\{y_i\theta_i - b(\theta_i) + c(y_i)\}\,. \tag{1.1}$$

where $\theta_i \in \Theta \subset \mathbb{R}$ is the natural parameter, $b(\cdot)$ and $c(\cdot)$ are known functions, and the cumulant function b is assumed to be twice differentiable with $b''(\theta) > 0$ for all $\theta \in \Theta$. The family (1.1) includes the Gaussian, Bernoulli, binomial, Poisson, negative binomial, and gamma distributions [21]. The mean response $\mathbb{E}(y_i) = b'(\theta_i)$ is related to a linear combination of the covariates $\mathbf{x}_i$ through a link function h so that $(h \circ b') : \Theta \mapsto \mathbb{R}$ is a strictly increasing function, i.e.

$$(h \circ b')(\theta_i) = \mathbf{x}_i^\top \boldsymbol{\beta}, \ \ i = 1, \ldots, n, \tag{1.2}$$

where $\boldsymbol{\beta} \in \mathbb{R}^p$ is the vector of regression coefficients to be estimated. In practice, h is often chosen to be the canonical link function $h = (b')^{-1}$. This is the case for Gaussian linear regression, logistic regression, and

[*] Keywords and phrases: generalized linear model, group sparsity, spike-and-slab, spike-and-slab group lasso, posterior contraction, rate of convergence, variable selection

[†] Department of Statistics, University of South Carolina, USA. Email: RBAI@mailbox.sc.edu

Poisson regression. However, h can also be chosen to be a *non*-canonical link function. Probit regression, negative binomial regression with a log link, and gamma regression with a log link are examples of GLMs with non-canonical link functions.

Nowadays, it is common to collect datasets where the total number of covariates p is large. In this "large p" setting, the covariates also often exhibit a grouping structure. For example, in genomic data, genes within the same biological pathway function together as a group to affect a clinical phenotype such as disease status or survival time [15]. At the individual gene level, mutations in an amino acid sequence can also be represented with group structure [24]. In our motivating data application in Section 7, each position of a protease gene sequence is represented by a group of binary variables, where each binary variable indicates the presence or absence of a specific amino acid. In MRI imaging, voxels within the same brain region naturally form a group [19, 34]. In semiparametric GLMs, continuous functions of the covariates are often estimated by groups of nonlinear basis functions [20] (see Experiment 3 in Sections 6.2 and 6.3). Finally, categorical covariates with multiple levels are typically represented as groups of dummy variables for each non-baseline category [7, 22] (see Experiment 4 in Sections 6.2 and 6.3).

In these scenarios, it is desirable to take advantage of the natural grouping structure in order to improve estimation and prediction. Suppose that we have G groups. Then we can express each ith covariate vector $\mathbf{x}_i \in \mathbb{R}^p$ as $\mathbf{x}_i = (\mathbf{x}_{i1}^\top, \ldots, \mathbf{x}_{iG}^\top)^\top$, where $\mathbf{x}_{ig} \in \mathbb{R}^{m_g}$ is a group of covariates of size m_g and $\sum_{g=1}^G m_g = p$. In this case, we relate the covariates to the mean response $\mathbb{E}(y_i) = b'(\theta_i)$ through the linear relationship,

$$(h \circ b')(\theta_i) = \sum_{g=1}^{G} \mathbf{x}_{ig}^\top \boldsymbol{\beta}_g, \quad i = 1, \ldots, n, \tag{1.3}$$

where $\boldsymbol{\beta}_g \in \mathbb{R}^{m_g}$ is the gth vector of regression coefficients corresponding to the gth group. It is clear that (1.2) is a special case of the grouped regression model (1.3) where $G = p$, and $m_g = 1$ for all $g \in \{1, \ldots, G\}$ (i.e. each regression coefficient in (1.2) is its own group of size one). Therefore, (1.3) provides a very natural generalization of the traditional GLM structure (1.2). To distinguish these two structures, we henceforth refer to the model (1.3) as a *grouped GLM* and the model (1.2) as an *unstructured GLM*.

When the number of groups G is moderate or large in (1.3), some form of regularization is often desired. In the frequentist literature, penalized group estimators have been extended to GLMs with group structure (1.3) [22, 20, 5, 7]. In the Bayesian literature, spike-and-slab priors [30, 19] and automatic relevance determination priors [34] have been used as sparsity-inducing priors in grouped GLMs. These methods all shrink a large number of the groups in (1.3) towards zero, so that only a few of the groups of covariates are significantly associated with the mean response.

In this paper, we adopt the Bayesian approach and employ a group spike-and-slab prior (to be introduced in Section 2) for estimating the $\boldsymbol{\beta}_g$'s in (1.3). We study our method theoretically and introduce a computationally efficient method for implementing it. Our theory and algorithms apply to any member of the exponential family (1.1) and encompass both canonical and non-canonical link functions. Further, even when there is no known group structure, our results can still be applied to the traditional unstructured GLM (1.2). Thus, our theoretical and computational framework is quite broad.

1.2 Related Work and Our Contributions

The literature on theory for *Bayesian* high-dimensional GLMs is quite sparse. Jiang [17], Jeong and Ghosal [16], and Tang and Martin [29] have also studied contraction rates for Bayesian GLMs in high dimensions. However, our work departs from these other papers in several important ways. First, we study GLMs under *group* sparsity (1.3). This setting is more general than the unstructured setting (1.2) considered by Jiang

[17], Jeong and Ghosal [16], and Tang and Martin [29]. However, since our model subsumes the traditional GLM (1.2) (where all G groups in (1.3) have size one), we obtain results for *both* models (1.2) and (1.3).

Secondly, our study is conducted under a *continuous* spike-and-slab prior, to be introduced in Section 2. In contrast, Jiang [17], Jeong and Ghosal [16], and Tang and Martin [29] all consider a *point-mass* prior. In these other papers, a model complexity prior is used to first select a random subset of $s < n$ predictors. Conditionally on the chosen set, the s coordinates are then endowed with a multivariate prior, while the other regression coefficients are modeled with a Dirac delta density at zero. On the other hand, we use an absolutely continuous prior which puts *zero* probability on exactly sparse vectors. Thus, our prior needs to be handled differently from these other papers.

Finally, we characterize the convergence rate of *both* the maximum *a posteriori* (MAP) estimator *and* the full posterior distribution. In contrast, Jiang [17], Jeong and Ghosal [16], and Tang and Martin [29] studied *only* the full posterior distribution but *not* any specific Bayesian point estimators. One may wonder why it is necessary to study the MAP estimator and the posterior distribution separately. First, practitioners typically report a point estimate (e.g. the posterior mean, median, or mode) when they employ Bayesian methods. Thus, it is important to study the properties of these point estimators. Secondly, many researchers have shown that different Bayesian point estimates may have different asymptotic properties or behave very differently from the full posterior. For example, in the sparse normal means model, Johnstone and Silverman [18] showed that the posterior median under a point mass spike-and-slab prior attains the minimax risk, whereas the posterior mean converges at a slower, suboptimal rate. Under a different empirical Bayes spike-and-slab prior, Castillo and Mismer [8] showed that the posterior mean and median both obtain the optimal rate, but the full posterior converges at a suboptimal rate. In high-dimensional linear regression under a Laplace prior, Castillo et al. [9] showed that the posterior mode converges at the near minimax rate but the full posterior distribution converges much more slowly than the mode. These examples reinforce the argument that Bayesian point estimators need to be analyzed separately from the full posterior.

In this paper, we prove that the MAP estimator and the full posterior distribution under a continuous, heavy-tailed spike-and-slab prior both converge at the same rate in GLMs with group sparsity. Our results thus justify the use of the posterior mode as a point estimator in high-dimensional Bayesian GLMs. At the same time, the posterior contraction rate also implies that the full posterior distribution provides valid inference in the sense that posterior credible sets have radius of an optimal size. As a byproduct of our analysis, we also obtain the first theoretical results for the univariate (non-grouped) spike-and-slab prior introduced by Ročková and George [27] in the context of high-dimensional GLMs. For computation, we adopt the penalized likelihood perspective and propose an EM algorithm to obtain MAP estimates.

The rest of this paper is structured as follows. In Section 2, we describe our prior specification and discuss Bayesian estimation of $\boldsymbol{\beta}$. In Section 3, we study the MAP estimator under our approach. In Section 4, we characterize the convergence rate for the *entire* posterior distribution and show that it can overcome the slow posterior contraction rate of the group lasso. Section 5 discusses how to implement our method. We conduct simulation studies in Section 6 and an analysis of an HIV drug resistance dataset in Section 7. Section 8 concludes the paper. Most of the proofs are deferred to the Appendix.

1.3 Notation and Preliminaries

For two sequences of positive real numbers a_n and b_n, we write $a_n = o(b_n)$ or $a_n \prec b_n$ if $\lim_{n\to\infty} a_n/b_n = 0$, $a_n = O(b_n)$ or $a_n \lesssim b_n$ if $|a_n/b_n| \leq M$ for some positive real number M independent of n, and $a_n \asymp b_n$ if $b_n \lesssim a_n \lesssim b_n$. For a set $\mathcal{S}$, we denote its cardinality by $|\mathcal{S}|$, and for a subset $\mathcal{T} \subset \mathcal{S}$, $\mathcal{T}^c$ means $\mathcal{T}^c = \mathcal{S} \setminus \mathcal{T}$.

The ℓ_∞, ℓ_2 and ℓ_1 norms of a vector $\mathbf{v}$ are denoted by $\|\mathbf{v}\|_\infty$, $\|\mathbf{v}\|_2$ and $\|\mathbf{v}\|_1$ respectively. For an $m \times n$ matrix $\mathbf{A}$ with entries a_{ij}, we denote $\|\mathbf{A}\|_{2,\infty} = \max_{1\leq i\leq n}(\sum_{j=1}^m a_{ij}^2)^{1/2}$ as the maximum row length of $\mathbf{A}$ and $\|\mathbf{A}\|_{\max} = \max_{i,j} |a_{ij}|$ as the maximum entry in absolute value. For a symmetric matrix $\mathbf{C}$, we denote its minimum and maximum eigenvalues by $\lambda_{\min}(\mathbf{C})$ and $\lambda_{\max}(\mathbf{C})$ respectively. For a vector $\mathbf{x}$, diag$(\mathbf{x})$ denotes the diagonal matrix determined by entries of $\mathbf{x}$. If f is a univariate function, then $f(\mathbf{x})$ means that f is applied elementwise to the entries in the vector $\mathbf{x}$. The notation $\mathbf{1}_m$ means an m-dimensional vector of all ones, while $\mathbf{0}_m$ denotes an m-dimensional zero vector.

To succinctly express a GLM with group sparsity (1.3) under the exponential family (1.1), we denote $\mathbf{Y} = (y_1, \ldots, y_n)^\top \in \mathbb{R}^n$, $\mathbf{X} = (\mathbf{X}_1, \ldots, \mathbf{X}_G) \in \mathbb{R}^{n \times p}$, and $\boldsymbol{\beta} = (\boldsymbol{\beta}_1^\top, \ldots, \boldsymbol{\beta}_G^\top)^\top \in \mathbb{R}^p$, where $p = \sum_{g=1}^G m_g$. We define the function ξ as $\xi = (h \circ b')^{-1}$. Then the log-likelihood for (1.3) can be written (up to normalizing constant) as

$$\ell_n(\boldsymbol{\beta}) = \mathbf{Y}^\top \xi(\mathbf{X}\boldsymbol{\beta}) - \mathbf{1}_n^\top b(\xi(\mathbf{X}\boldsymbol{\beta})). \tag{1.4}$$

The gradient of $\ell_n(\boldsymbol{\beta})$ is then

$$\nabla \ell_n(\boldsymbol{\beta}) = \mathbf{X}^\top \text{diag}(\xi'(\mathbf{X}\boldsymbol{\beta}))(\mathbf{Y} - b'(\mathbf{X}\boldsymbol{\beta})). \tag{1.5}$$

while the Hessian is

$$\nabla^2 \ell_n(\boldsymbol{\beta}) = -\mathbf{X}^\top \boldsymbol{\Sigma}(\boldsymbol{\beta})\mathbf{X}, \tag{1.6}$$

where $\boldsymbol{\Sigma}(\boldsymbol{\beta})$ is an $n \times n$ diagonal matrix that depends on $\boldsymbol{\beta}$ through

$$\boldsymbol{\Sigma}(\boldsymbol{\beta}) = \text{diag}\{(h^{-1})'(\mathbf{X}\boldsymbol{\beta})\xi'(\mathbf{X}\boldsymbol{\beta})\}. \tag{1.7}$$

2 Prior specification and Bayesian estimation

2.1 Spike-and-slab group lasso

Given a high-dimensional GLM with group structure (1.3), a Bayesian approach to estimation and variable selection is to put a prior on the parameter $\boldsymbol{\beta}$. For an $m_g \times 1$ random vector $\boldsymbol{\beta}_g$, we first define the multivariate density function,

$$\boldsymbol{\Psi}(\boldsymbol{\beta}_g \mid \lambda) = \frac{\lambda^{m_g} e^{-\lambda \|\boldsymbol{\beta}_g\|_2}}{2^{m_g} \pi^{m_g - 1} \Gamma((m_g + 1)/2)}. \tag{2.1}$$

It is important to note that (2.1) is a multivariate *Laplace* distribution, *not* a multivariate Gaussian. The exponent term of (2.1) contains the ℓ_2 norm $\|\boldsymbol{\beta}_g\|_2$ rather than the squared ℓ_2 norm $\|\boldsymbol{\beta}_g\|_2^2$. As a result, the density (2.1) has tails that are *heavier* than normal. It is easy to see that if $m_g = 1$, then (2.1) reduces to a univariate *Laplace* density with scale λ^{-1}. The hyperparameter λ controls how concentrated $\boldsymbol{\beta}_g$ is around the zero vector $\mathbf{0}_{m_g}$, with larger values of λ leading to a Laplace density that is more peaked around $\mathbf{0}_{m_g}$.

To induce group sparsity in $\boldsymbol{\beta}$ under (1.3), we endow $\boldsymbol{\beta}$ with the spike-and-slab group lasso (SSGL) prior of Bai et al. [2],

$$\pi(\boldsymbol{\beta}) = \prod_{g=1}^{G} \left[(1-\theta)\boldsymbol{\Psi}(\boldsymbol{\beta}_g \mid \lambda_0) + \theta \boldsymbol{\Psi}(\boldsymbol{\beta}_g \mid \lambda_1) \right], \tag{2.2}$$

where $\theta \in (0,1)$ is a mixing proportion. In (2.2), λ_0 is set to be a large value so that $\boldsymbol{\Psi}(\boldsymbol{\beta}_g \mid \lambda_0)$, i.e. the "spike," is highly concentrated around the zero vector $\mathbf{0}_{m_g}$. Meanwhile, $\lambda_1 \ll \lambda_0$ is set to be small so that $\boldsymbol{\Psi}(\boldsymbol{\beta}_g \mid \lambda_1)$, i.e. the "slab," is a diffuse and relatively flat density. In (2.2), the slab density models the nonzero groups, while the spike density models the zero groups. The SSGL prior was originally introduced by Bai et al. [2] in the Gaussian linear regression model, $\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}, \boldsymbol{\varepsilon} \sim \mathcal{N}_n(\mathbf{0}, \sigma^2 \mathbf{I}_n)$. This paper extends the work of Bai et al. [2] to the much more general GLM setting where the response variables $\mathbf{Y}$ can also be discrete or non-Gaussian (e.g. binary, binomial, Poisson, negative binomial, gamma, etc.).

When $m_1 = \ldots = m_G = 1$, the SSGL prior (2.2) reduces to a two-component mixture of *univariate* Laplace densities,

$$\pi(\boldsymbol{\beta}) = \prod_{j=1}^{p} \left[(1-\theta)\psi(\beta_j \mid \lambda_0) + \theta \psi(\beta_j \mid \lambda_1) \right]. \tag{2.3}$$

where $\psi(\beta_j \mid \lambda) = (\lambda/2)\exp(-\lambda|\beta_j|)$ denotes the density of a univariate Laplace distribution. The prior (2.3) is the spike-and-slab lasso (SSL) originally introduced by Ročková and George [27] in non-grouped linear regression. In order to conduct Bayesian inference for unstructured GLMs (1.2), we can place the SSL prior (2.3) on the individual regression coefficients in (1.2). Tang et al. [31] extended the SSL (2.3) to high-dimensional GLMs. However, the theoretical properties for the SSL in GLMs have thus far not been investigated. As a byproduct of our theoretical analysis of the SSGL (2.2), we *also* obtain the rates of convergence for the SSL (2.3) in GLMs in Sections 3 and 4.

2.2 Bayesian Estimation

After endowing the groups of regression coefficients $\boldsymbol{\beta}$ in (1.3) with an appropriate prior distribution $\pi(\boldsymbol{\beta})$, we obtain the posterior distribution for $\boldsymbol{\beta}$,

$$\pi(\boldsymbol{\beta} \mid \mathbf{Y}) = \frac{\exp(\ell_n(\boldsymbol{\beta}))\pi(\boldsymbol{\beta})}{\int \exp(\ell_n(\boldsymbol{\beta}))\pi(\boldsymbol{\beta}) d\boldsymbol{\beta}}, \tag{2.4}$$

where $\ell_n(\boldsymbol{\beta})$ is the log-likelihood (1.4). The posterior (2.4) is typically intractable, but Markov chain Monte Carlo (MCMC) can be used to draw samples from the approximate posterior. From (2.4), we also see that the log-posterior (up to normalizing constant) is

$$\log \pi(\boldsymbol{\beta} \mid \mathbf{Y}) = \ell_n(\boldsymbol{\beta}) + \log \pi(\boldsymbol{\beta}).$$

Hence, a very natural point estimator for $\boldsymbol{\beta}$ is the MAP estimator,

$$\widehat{\boldsymbol{\beta}} = \underset{\boldsymbol{\beta} \in \mathbb{R}^p}{\arg\max} \left\{ \ell_n(\boldsymbol{\beta}) + \log \pi(\boldsymbol{\beta}) \right\}. \tag{2.5}$$

In particular, if $\pi(\boldsymbol{\beta}) = \prod_{g=1}^{G} \boldsymbol{\Psi}(\boldsymbol{\beta}_g \mid \lambda)$, where $\boldsymbol{\Psi}(\cdot \mid \lambda)$ is the multivariate Laplace prior (2.1), then (2.5) becomes

$$\widehat{\boldsymbol{\beta}} = \underset{\boldsymbol{\beta} \in \mathbb{R}^p}{\arg\max} \left\{ \ell_n(\boldsymbol{\beta}) - \lambda \sum_{g=1}^{G} \|\boldsymbol{\beta}_g\|_2 \right\}, \tag{2.6}$$

which is the objective function for the group lasso of Yuan and Lin [35]. Thus, the group lasso corresponds to the MAP estimator (2.5) under independent multivariate Laplace priors (2.1) on the $\boldsymbol{\beta}_g$'s, and the MAP estimator for each $\boldsymbol{\beta}_g$ is either exactly $\mathbf{0}_{m_g}$ or nonzero. If instead, we use the SSGL prior (2.2) for $\pi(\boldsymbol{\beta})$ in (2.5), then the SSGL MAP estimator will *also* be exactly sparse, since the mixture components in (2.2) are both multivariate Laplace. However, whereas the group lasso (2.6) applies the same amount of shrinkage λ to every group, the SSGL (2.2) allows for *adaptive* shrinkage. This is because the slab density $\boldsymbol{\Psi}(\cdot \mid \lambda_1)$ of the SSGL (2.2) prevents groups with larger coefficients from being downward biased.

The combination of exact group sparsity and adaptive shrinkage of the MAP estimator (2.5) under the SSGL prior (2.2) makes the SSGL very appealing for both group selection and estimation. In Sections 3.2 and 4.2, we demonstrate the theoretical advantages of the SSGL prior (2.2) over the group lasso prior (2.1). SSGL is also empirically shown to significantly outperform the group lasso in Sections 6 and 7.

3 Characterization of the MAP estimator

3.1 Convergence rate of the MAP estimator

We first consider the MAP estimator (2.5) under the SSGL prior (2.2). To the best of our knowledge, our work is the first one to investigate a Bayesian point estimator (other than the group lasso (2.6)) in high-dimensional Bayesian GLMs. Other authors [17, 16, 29] have only studied the convergence rate for the full posterior distribution in Bayesian GLMs, which we will also visit in Section 4.

Suppose that the true regression coefficients vector is $\boldsymbol{\beta}_0 = (\boldsymbol{\beta}_{01}^\top, \ldots, \boldsymbol{\beta}_{0G}^\top)^\top \in \mathbb{R}^p$, where $\boldsymbol{\beta}_{0g} \in \mathbb{R}^{m_g}$. Then the true model is

$$(h \circ b')(\theta_{0i}) = \sum_{g=1}^{G} \mathbf{x}_{ig}\boldsymbol{\beta}_{0g}, \quad i = 1, \ldots, n, \tag{3.1}$$

where h and b are the known link function and cumulant function respectively, while θ_{0i} is the true natural parameter in (1.1).

Below, we let $S_0 \subset \{1, \ldots, G\}$ be the set of indices of the true nonzero groups in $\boldsymbol{\beta}_0$, with cardinality $s_0 = |S_0|$. Then $\mathbf{X}_{S_0}$ denotes the submatrix of the design matrix $\mathbf{X}$ with the $\sum_{g \in S_0} m_g$ columns of $\mathbf{X}$, and the complement of S_0 is $S_0^c = \{1, \ldots, G\} \setminus S_0$. Recall that $\boldsymbol{\Sigma}(\boldsymbol{\beta})$ is defined as in (1.7), and by (1.6), the negative Hessian of the log-likelihood is $\mathbf{X}^\top\boldsymbol{\Sigma}(\boldsymbol{\beta})\mathbf{X}$. Finally, we denote $m_{\max} = \max_{1 \leq g \leq G} m_g$. We make the following assumptions:

(A1) $G \gg n$ and $\log G = o(n)$.

(A2) $s_0 = o((n/\log G)^{1/2})$ and $m_{\max} = O(\log n \wedge (\log G/\log n))$.

(A3) The design matrix $\mathbf{X}$ satisfies the following conditions:

- (i) All the entries x_{ij} of $\mathbf{X}$ satisfy $|x_{ij}| \leq D$ for some constant $D > 0$.
- (ii) Define the neighborhood $\mathcal{N}_0 = \{\boldsymbol{\delta} \in \mathbb{R}^p : \|\boldsymbol{\delta} - \boldsymbol{\beta}_0\|_2 \leq (s_0 \log G/n)^{1/2}\}$. Then for any $\boldsymbol{\delta} \in \mathcal{N}_0$, there exists a positive constant $\underline{\tau} > 0$ such that $\lambda_{\min}(n^{-1}\mathbf{X}_{S_0}^\top\boldsymbol{\Sigma}(\boldsymbol{\delta})\mathbf{X}_{S_0}) \geq \underline{\tau}$. Furthermore, $\lambda_{\max}\left(n^{-1}\mathbf{X}_{S_0}^\top\boldsymbol{\Sigma}(\boldsymbol{\delta})\mathbf{X}_{S_0}\right) \lesssim \log G/\log n$.
- (iii) For any group $g \in S_0^c$ and $\boldsymbol{\delta} \in \mathcal{N}_0$, $\|\mathbf{X}_{S_0}^\top\boldsymbol{\Sigma}(\boldsymbol{\delta})\mathbf{X}_g\|_{2,\infty} = O(n)$.

(A4) The observations $\{(\mathbf{x}_i, y_i)\}_{i=1}^n$ satisfy the following conditions:

- (i) The responses $\{y_i\}_{i=1}^n$ satisfy $\mathbb{E}(|y_i - b'(\xi(\mathbf{x}_i^\top\boldsymbol{\beta}_0))|^k) \leq \frac{k!}{2}\mathbb{E}[y_i^2]L^{k-2}$ for some $L > 0$ and every integer $k \geq 2$. In addition, $\mathrm{Var}(y_i \mid \mathbf{x}_i) \lesssim G$ for all $i = 1, \ldots, n$.
- (ii) For the function $\xi = (h \circ b')^{-1}$ in (1.4), $\xi'(\mathbf{x}_i^\top\boldsymbol{\beta}) < \infty$ for all $i = 1, \ldots, n$.

Condition (A1) allows the number of groups G to diverge at a nearly exponential rate with n. Condition (A2) allows the number of nonzero groups s_0 and the group sizes m_g's to diverge but at rates slower than n. Part (ii) of Condition (A3) is a restricted eigenvalue condition on the negative Hessian. This assumption is routinely employed in the high-dimensional GLM literature [12, 29]. By (A2), the number of columns in $\mathbf{X}_{S_0}$ is always strictly less than n, and thus, it is sensible for $\lambda_{\min}(n^{-1}\mathbf{X}_{S_0}^\top\boldsymbol{\Sigma}(\boldsymbol{\delta})\mathbf{X}_{S_0})$ to be bounded away from zero. These conditions ensure the identifiability of $\boldsymbol{\beta}_0$. Part (iii) of Condition (A3) is an irrepresentability condition [36, 12], which limits the correlations between the active covariates $\mathbf{X}_{S_0}$ and the inactive ones $\mathbf{X}_{S_0^c}$. Since $\mathbf{X}_{S_0}^\top\boldsymbol{\Sigma}(\boldsymbol{\delta})\mathbf{X}_g$ has m_g columns and $m_g \ll n$ for all $g \in \{1, \ldots, G\}$, the condition that $\|\mathbf{X}_{S_0}^\top\boldsymbol{\Sigma}(\boldsymbol{\delta})\mathbf{X}_g\|_{2,\infty} = O(n)$ is also fairly mild. Fan and Lv [12] make a very similar assumption to ours.

Part (i) of Condition (A4) is an assumption on the central moments of the response variables and implies that the tails decay exponentially. This assumption is satisfied for the Gaussian, Poisson, Bernoulli, gamma, and Laplace distributions, among others [3]. In addition, the assumption that $\text{Var}(y_i \mid \mathbf{x}_i) \lesssim G$ is a very weak assumption, in light of condition (A1) that allows $G = O(e^{n^{\xi}})$, for some $\xi \in (0,1)$. Part (ii) of Condition (A4) is also satisfied for many GLMs, even if $\mathbf{x}_i^\top \boldsymbol{\beta}$ is unbounded. For example, the canonical link function $h = (b')^{-1}$ is usually used in practice, e.g. in logistic, Poisson, and Gaussian regression. In this case, $\xi(u) = (h \circ b')^{-1}(u) = u$ and $\xi'(u) = 1$ for all $u \in \mathbb{R}$. In negative binomial regression with the log link $h(u) = \log u$ and a given number of failures r, we have $b(u) = -r\log(1-e^u)$, $\xi(u) = -\log(re^{-u}+1)$. Thus, $\xi'(u) = r/(r+e^u) \leq 1$ for all $u \in \mathbb{R}$. However, even if $\xi'(u)$ is unbounded in $\mathbb{R}$, we can still satisfy $\xi'(\mathbf{x}_i^\top \boldsymbol{\beta}_0) < \infty$ if we make a stronger assumption that $\|\mathbf{X}\boldsymbol{\beta}_0\|_\infty < \infty$.

Theorem 1 (convergence rate of the MAP estimator under SSGL)**.** *Suppose that we have a grouped GLM* (3.1)*, and we endow* $\boldsymbol{\beta}_0$ *with the SSGL prior* (2.2) *where the hyperparameters* $(\lambda_0, \lambda_1, \theta)$ *satisfy* $\lambda_0 = (1-\theta)/\theta \asymp G^c$*, where* $c > 2$*, and* $\lambda_1 \asymp 1/n$*. Further, assume that conditions (A1)-(A4) hold. Then there exists a MAP estimator* $\widehat{\boldsymbol{\beta}}$ (2.5) *so that as* $n \to \infty$*,*

$$\|\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0\|_2 = O_p\left(\sqrt{\frac{s_0 \log G}{n}}\right). \tag{3.2}$$

Remark 1. *We have treated the mixing weight* θ *as a deterministic quantity which depends on* G*. However, Theorem 1 still holds if we instead put a prior* $\pi(\theta)$ *on* θ*, as long as* $\pi(\theta)$ *satisfies* $P((1-\theta)/\theta \geq G^c) \geq 1 - e^{-Ms_0 \log G}$ *where* $M > 0$ *and* $c > 2$*. Then, since with probability tending to one,* $\lambda_0 = (1-\theta)/\theta \geq G^c$*, we can condition our analysis on the high probability event* $\mathcal{A} = \{(1-\theta)/\theta \geq G^c\}$ *and our theory still holds. This will be satisfied, for example, when* $\theta \sim \text{Beta}(1, G^c)$ *[2].*

We also have the following corollary which gives the convergence rate of the MAP estimator for $\boldsymbol{\beta}$ under the SSL prior (2.3) of Ročková and George [27] on the regression coefficients in unstructured GLMs (1.2).

Corollary 1 (convergence rate of the MAP estimator under SSL)**.** *Suppose that we have an unstructured GLM* (1.2)*, and we endow* $\boldsymbol{\beta}_0$ *with the SSL prior* (2.3) *where the hyperparameters* $(\lambda_0, \lambda_1, \theta)$ *satisfy* $\lambda_0 = (1-\theta)/\theta \asymp p^c$*, where* $p > 2$*, and* $\lambda_1 \asymp 1/n$*. Further, assume that* $p \gg n$*,* $\log p = o(n)$*,* $s_0 = o((n/\log p)^{1/2})$*, and Assumptions (A3)-(A4) hold with* G *replaced by* p*. Then there exists a MAP estimator* $\widehat{\boldsymbol{\beta}}$ *so that as* $n \to \infty$*,*

$$\|\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0\|_2 = O_p\left(\sqrt{\frac{s_0 \log p}{n}}\right). \tag{3.3}$$

Proof. The model (1.2) is a special case of the grouped model (1.3) where $m_1 = \ldots = m_G = 1$. Since we can also treat the SSL prior (2.3) as a special case of the SSGL prior (2.2) with $m_1 = \cdots = m_G = 1$, the result follows from Theorem 1. □

Theorems 1 and Corollary 1 justify the use of the MAP estimator (2.5) as a point estimate under the SSGL (2.2) and SSL (2.3) priors in high-dimensional GLMs. In Section 4, we will turn our attention to the asymptotic behavior of the *full* posterior distribution.

3.2 Comparison of our work to the group lasso

The group lasso estimator (2.6) of Yuan and Lin [35] has been studied theoretically in GLMs by Blazère et al. [5]. Similar to the SSGL, Blazère et al. [5] obtained the convergence rate of $O((s_0 \log G/n)^{1/2})$ for the group lasso. However, Blazère et al. [5] require the assumption that $\sum_{g=1}^{G} \sqrt{m_g}\|\boldsymbol{\beta}_{0g}\|_2 < \infty$ (condition (H.3) in Blazère et al. [5]) in order to achieve this rate. The condition that $\sum_{g=1}^{G} \sqrt{m_g}\|\boldsymbol{\beta}_{0g}\|_2 < \infty$ seems highly restrictive, especially if the number of groups G diverges to infinity as in Assumption (A1). The only way for this assumption to be satisfied in practice is if *all* of the following conditions hold: (i) the group sizes m_g do *not* diverge with n, (ii) the number of nonzero groups s_0 is fixed and does *not* diverge with n, and (iii) $\|\boldsymbol{\beta}_0\|_\infty < \infty$.

In contrast, we do not make such a strong assumption. Theorem 1 allows both the group sizes m_g *and* the number of nonzero groups s_0 to diverge (Assumption (A2)), while the maximum signal strength $\|\boldsymbol{\beta}_0\|_\infty$ can also grow to infinity. All of these assumptions clearly cause Condition (H.3) of Blazère et al. [5] to be violated. In short, we have derived the convergence rate for the SSGL MAP estimator in GLMs under much weaker conditions than those previously used for the group lasso estimator (2.6). In Section 4.2, we further demonstrate the advantage of SSGL (2.2) over the group lasso from a fully Bayesian perspective.

4 Characterization of the full posterior

4.1 Posterior contraction rate

As mentioned in Section 1.2, it is *not* necessarily the case that the posterior contracts around the true $\boldsymbol{\beta}_0$ in (3.1) at the same rate as point estimates from the posterior [9, 8]. In this section, we analyze the *full* SSGL posterior (2.4) and show that the posterior $\pi(\boldsymbol{\beta} \mid \mathbf{Y})$ inherits the nice theoretical properties of the SSGL MAP estimator (2.5).

In order to derive theory for the SSGL posterior in high-dimensional GLMs, we require a different set of conditions on the design matrix and on the maximum signal strength of $\boldsymbol{\beta}_0$. In particular, we replace conditions (A3)-(A4) with the following conditions. Recall that $S_0 \subset \{1, \ldots, G\}$ is the set of true nonzero groups which has cardinality $|S_0| = s_0$, and $\boldsymbol{\Sigma}(\boldsymbol{\beta})$ is defined as in (1.7).

(B3) The design matrix $\mathbf{X}$ satisfies the following conditions:

(i) All the entries x_{ij} of $\mathbf{X}$ satisfy $|x_{ij}| \leq D$ for some constant $D > 0$.

(ii) For a set of indices $S \subset \{1, \ldots, G\}$, let $\mathbf{X}_S$ denote the submatrix of $\mathbf{X}$ whose columns contain the groups $g \in S$. For any S where $|S| \leq M s_0$, $M > 0$, we have $\lambda_{\min}\left(n^{-1}\mathbf{X}_S \boldsymbol{\Sigma}(\boldsymbol{\beta}_0)\mathbf{X}_S\right) \geq \underline{\delta}$, where $\underline{\delta} > 0$. Meanwhile, for any $g \in \{1, \ldots, G\}$, $\lambda_{\max}(n^{-1}\mathbf{X}_g \boldsymbol{\Sigma}(\boldsymbol{\beta})\mathbf{X}_g) \lesssim \log G$.

(B4) The maximum signal strength satisfies $\|\boldsymbol{\beta}_0\|_\infty = O(\log G)$.

Part (ii) of Condition (B3) imposes eigenvalue conditions on the design matrix. However, these conditions are stronger than those in part (ii) of condition (A3). The minimum restricted eigenvalue condition $\lambda_{\min}(n^{-1}\mathbf{X}_S \boldsymbol{\Sigma}(\boldsymbol{\beta}_0)\mathbf{X}_S) > 0$ is required to hold for *all* submatrices $\mathbf{X}_S$ where $|S| \leq M s_0$. The maximum eigenvalue condition $\lambda_{\max}(n^{-1}\mathbf{X}_g \boldsymbol{\Sigma}(\boldsymbol{\beta})\mathbf{X}_g) \lesssim \log G$ also needs to hold for all individual submatrices $\mathbf{X}_g$ (not necessarily the entire $n \times p$ design matrix $\mathbf{X}$). In contrast, part (ii) of (A3) only imposes eigenvalue conditions for a *single* submatrix $\mathbf{X}_{S_0}$. Intuitively, this is because in Theorem 1, we only need to be able to find one local mode in a small neighborhood around $\boldsymbol{\beta}_0$. Contrastingly, Theorems 2 and 3 require the *entire* posterior $\pi(\boldsymbol{\beta} \mid \mathbf{Y})$ to concentrate all its mass on configurations where no more than a constant multiple of s_0 of the entries in $\boldsymbol{\beta}$ have magnitude much larger than zero. On the other hand, however, the irrepresentability condition in part (iii) of (A3) is *not* needed for the full posterior to contract.

Condition (B4) also replaces the moment conditions on the responses in (A4) with a more direct condition on the maximum signal strength for $\boldsymbol{\beta}_0$. We need this condition because the SSGL prior (2.2) must be able to put sufficient prior mass in a neighborhood of the true $\boldsymbol{\beta}_0$ for the posterior to contract around $\boldsymbol{\beta}_0$. Overall, our theoretical results underscore the importance of studying posterior point estimates separately from the full posterior, because different conditions may be required for convergence of these two objects.

A crucial difference between our theory and that of Jiang [17], Jeong and Ghosal [16], and Tang and Martin [29] is that the SSGL prior (2.2) is an absolutely *continuous* spike-and-slab prior. Although the posterior *mode* under (2.2) is exactly sparse, the SSGL prior itself puts zero probability on exactly sparse vectors. Therefore, in order to analyze the *full* posterior (2.4), we must resort to a notion of "approximate" sparsity known as the *generalized dimensionality* [4, 27, 2]. Following Bai et al. [2], we use a small quantity $\omega_g > 0$ to define the generalized inclusion indicator $\nu_{\omega_g}(\boldsymbol{\beta}_g)$ and generalized dimensionality $|\boldsymbol{\nu}(\boldsymbol{\beta})|$ respectively as

$$\nu_{\omega_g}(\boldsymbol{\beta}_g) = I(\|\boldsymbol{\beta}_g\|_2 > \omega_g) \text{ and } |\boldsymbol{\nu}(\boldsymbol{\beta})| = \sum_{g=1}^{G} \nu_{\omega_g}(\boldsymbol{\beta}_g). \tag{4.1}$$

As long as ω_g in (4.1) tends to zero as $n \to \infty$ for all $g \in \{1, \ldots, G\}$, then $|\boldsymbol{\nu}(\boldsymbol{\beta})|$ will provide a good approximation to the number of nonzero groups, $\#\{g : \boldsymbol{\beta}_g \neq \mathbf{0}_{m_g}\}$. For the threshold ω_g in (4.1), we use

$$\omega_g = \frac{1}{\lambda_0 - \lambda_1} \log \left[\frac{1-\theta}{\theta} \frac{\lambda_0^{m_g}}{\lambda_1^{m_g}} \right], \tag{4.2}$$

where $(\lambda_0, \lambda_1, \theta)$ are the hyperparameters in the SSGL prior (2.2). As described in Bai et al. [2], any vectors $\boldsymbol{\beta}_g$ that satisfy $\|\boldsymbol{\beta}_g\|_2 = \omega_g$ are the intersection points between the spike density $\boldsymbol{\Psi}(\cdot \mid \lambda_0)$ and the slab density $\boldsymbol{\Psi}(\cdot \mid \lambda_1)$ in (2.2). Therefore, if $\|\boldsymbol{\beta}_g\|_2 > \omega_g$, then $\boldsymbol{\beta}_g$ is much more likely to belong to the slab rather than the spike.

We first show in Theorem 2 that the posterior $\pi(\boldsymbol{\beta} \mid \mathbf{Y})$ under the SSGL prior (2.2) asymptotically puts all of its mass on vectors where the generalized dimensionality (4.1) is no larger than a constant multiple of the true model size s_0. That is, the SSGL posterior concentrates on sparse sets in high-dimensional GLMs. The posterior contraction rate is then given in Theorem 3.

Theorem 2 (posterior concentration on approximately sparse sets)**.** *Assume the same setup as Theorem 1, and suppose that conditions (A1)-(A2) and (B3)-(B4) hold. Then for some $K_1 > 0$,*

$$\mathbb{E}_0 \Pi \left(\boldsymbol{\beta} : |\boldsymbol{\nu}(\boldsymbol{\beta})| > K_1 s_0 \mid \mathbf{Y} \right). \to 0 \tag{4.3}$$

Theorem 3 (posterior contraction rate under SSGL)**.** *Assume the same setup as Theorem 1, and suppose that conditions (A1)-(A2) and (B3)-(B4) hold. Then for some $K_2 > 0$, as $n \to \infty$,*

$$\mathbb{E}_0 \Pi \left(\boldsymbol{\beta} : \|\boldsymbol{\beta} - \boldsymbol{\beta}_0\|_2 > K_2 \sqrt{\frac{s_0 \log G}{n}} \;\middle|\; \mathbf{Y} \right) \to 0. \tag{4.4}$$

Remark 2. *Similarly as with the posterior mode in Theorem 1, we can also obtain the same posterior contraction rate for the full posterior when we endow θ in (2.2) with a prior $\pi(\theta)$. As long as $\pi(\theta)$ satisfies $P((1-\theta)/\theta > G^c) > 1 - e^{-Ms_0 \log G/n}$, where $M > 0$ and $c > 2$, then $\lambda_0 = (1-\theta)/\theta \geq G^c$ with probability tending to one, and the theory still holds. This will be the case if $\theta \sim$ Beta$(1, G^c)$. See Bai et al. [2] for the proof in the case of Gaussian grouped linear regression, which also holds in the GLM setting.*

Remark 3. *In the cases of Gaussian regression and logistic regression, one may be able to remove the restriction (B4) on the maximum signal strength (see, e.g., [9, 27, 1]). However, since we consider GLMs under the general exponential family (1.1), it seems unlikely that a condition such as (B4) can be totally removed for GLMs in general. To see why, consider Poisson regression, where the cumulant function is $b(\theta_{0i}) = e^{\mathbf{x}_i^\top \boldsymbol{\beta}_0}$ and the diagonal entries of $\boldsymbol{\Sigma}(\boldsymbol{\beta}_0)$ are $\{e^{\mathbf{x}_i^\top \boldsymbol{\beta}_0}\}_{i=1}^n$. In this scenario, it seems difficult to control the approximation error without any restrictions on $\|\boldsymbol{\beta}_0\|_\infty$.*

The following corollary is immediate from Theorem 2. If we have an unstructured GLM (1.2) and we endow the regression coefficients in $\boldsymbol{\beta}_0$ with the SSL prior (2.3) of Ročková and George [27], we obtain the following posterior contraction rate.

Corollary 2 (posterior contraction rate under SSL)**.** *Assume the same setup as Corollary 1. Suppose that* $p \gg n$, $\log p = o(n)$, $s_0 = o((n/\log p)^{1/2})$, *and Assumptions (B3)-(B4) hold with* G *replaced by* p. *Then for some* $K_3 > 0$, *as* $n \to \infty$,

$$\mathbb{E}_0 \Pi \left(\boldsymbol{\beta} : \|\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0\|_2 > K_3 \sqrt{\frac{s_0 \log p}{n}} \;\middle|\; \mathbf{Y} \right) \to 0. \tag{4.5}$$

Theorem 3 and Corollary 2 show that for high-dimensional GLMs, the posterior distributions under the SSGL (2.2) and SSL (2.3) priors converge at the *same* rate as their respective MAP estimators. This is not guaranteed to be the case in general. Our results also suggest that the SSGL and the SSL posteriors $\pi(\boldsymbol{\beta} \mid \mathbf{Y})$ provide valid inference in high-dimensional GLMs. Determining the posterior contraction rate of the full posterior is often the first step towards obtaining frequentist guarantees about uncertainty quantification for Bayesian procedures [25]. In particular, the posterior contraction rate gives an indication as to how large we can expect the posterior credible sets to be [25]. However, beyond just their size, more detailed study is typically required to guarantee that these credible sets are also honest, or have asymptotic coverage probability greater than or equal to the prescribed confidence level $1 - \alpha, \alpha \in (0, 1)$ [25]. The issue of honest coverage of posterior credible sets is beyond the scope of this paper.

4.2 Suboptimality of the group lasso for fully Bayesian inference

In Section 3.2, we demonstrated that the MAP estimator under the SSGL converges under weaker assumptions than those previously assumed for the group lasso (2.6). It turns out that from the fully Bayesian perspective, the SSGL also has an advantage over the group lasso. As discussed in Section 2.2, the group lasso estimator (2.6) is the MAP estimator under the prior $\pi(\boldsymbol{\beta}) = \prod_{g=1}^{G} \boldsymbol{\Psi}(\boldsymbol{\beta}_g \mid \lambda)$, where $\boldsymbol{\Psi}(\cdot \mid \lambda)$ is a *single* multivariate Laplace density (2.1). In contrast to the two-group SSGL (2.2), the group lasso posterior might contract much *slower* than its posterior mode. This renders the group lasso less useful for uncertainty quantification. This is formalized in the next proposition.

Proposition 1. *Suppose that we have a grouped GLM* (3.1)*, and we endow* $\boldsymbol{\beta}_0$ *with the group lasso prior* $\pi(\boldsymbol{\beta}) = \prod_{g=1}^{G} \boldsymbol{\Psi}(\boldsymbol{\beta}_g \mid \lambda)$*, where* $\boldsymbol{\Psi}(\cdot \mid \lambda)$ *is as in* (2.1)*. Then there exist scenarios where the full posterior distribution* $\pi(\boldsymbol{\beta} \mid \mathbf{Y})$ *converges slower than the MAP estimator* (2.6)*.*

Proof. Consider $\mathbf{Y} \sim \mathcal{N}_n(\boldsymbol{\beta}_0, \boldsymbol{I}_n)$, which falls under the exponential family (1.1) and is a special case of model (3.1), with $\mathbf{X} = \mathbf{I}_n$. First, suppose that $\boldsymbol{\beta}_0 = \mathbf{0}_n$ and $m_1 = \ldots = m_G = 1$ so that the group lasso prior reduces to $\pi(\boldsymbol{\beta}) = \prod_{g=1}^{G} (\lambda/2) \exp(-\lambda|\beta_g|)$. Setting $\lambda = \sqrt{2 \log n}$ yields the near-minimax risk of $\sqrt{\log n}$ for the MAP estimator $\widehat{\boldsymbol{\beta}}$, i.e. $\|\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0\|_2 = O_p(\sqrt{\log n})$ as $n \to \infty$ [11, 9]. However, by Theorem 7 of Castillo et al. [9], the posterior $\pi(\boldsymbol{\beta} \mid \mathbf{Y})$ then places no probability on the ball $\{\|\boldsymbol{\beta}\|_2 \leq \sqrt{n/\log n}\}$ as $n \to \infty$. That is, for some $K_4 > 0$,

$$\mathbb{E}_{\boldsymbol{\beta}_0 = \mathbf{0}_n} \Pi \left(\boldsymbol{\beta} : \|\boldsymbol{\beta}\|_2 \leq K_4 \sqrt{\frac{n}{\log n}} \;\middle|\; \mathbf{Y} \right) \to 0 \;\text{ as } n \to \infty,$$

and $\sqrt{\log n} \ll \sqrt{n/\log n}$ when n is large. In the general group case where $\boldsymbol{\beta}_0 = (\boldsymbol{\beta}_{01}^\top, \ldots, \boldsymbol{\beta}_{0G})^\top$, suppose that $m_g = O(1)$ and $\boldsymbol{\beta}_{0g} = \mathbf{0}_{m_g}$ for all $g \in \{1, \ldots, G\}$. In this scenario, $\lambda = \sqrt{2 \log n}$ still yields the near-minimax risk of $\sqrt{\log n}$ for $\widehat{\boldsymbol{\beta}}$ under expected ℓ_2 loss. However, the posterior contraction rate under the group lasso prior $\pi(\boldsymbol{\beta}) = \prod_{g=1}^{G} \boldsymbol{\Psi}(\boldsymbol{\beta}_g \mid \lambda)$ is still slower than $\sqrt{n/\log n}$. □

Proposition 1 implies that the full posterior distribution under the group lasso prior might put all of its mass in an ℓ_2 ball with radius that is *substantially larger* than the convergence rate of the posterior mode. Intuitively, this is because there is a conflict between shrinking the coefficients to zero and optimally estimating the nonzero signals. In order to estimate very sparse parameters well, the group lasso needs to set λ to be large. But if λ is too large, then there will be too much bias in the resulting estimator. The addition of the slab density $\boldsymbol{\Psi}(\cdot \mid \lambda_1)$ in the SSGL prior (2.2) alleviates this tension by *preventing* overshrinkage of true signals. Thus, the full SSGL posterior (2.4) shares the same convergence properties as its MAP estimator (2.5).

5 Implementation

5.1 EM algorithm

In order to implement the SSGL model for GLMs, we adopt the penalized likelihood perspective and perform MAP estimation. The MAP estimator (2.5) is appealing, not just because of its nice theoretical properties, but also because it is *exactly* sparse. Thus, the MAP estimator can be used for both estimation and variable selection in GLMs. To obtain the MAP estimator, we extend the EM variable selection (EMVS) approach of Ročková and George [26] to the GLM setting with grouped variables.

For the theoretical results in Sections 3 and 4, we treated the mixing proportion θ in (2.2) as a deterministic quantity (that depends on n and G). Our theoretical results still hold with a prior on θ, as long as the prior ensures that $\mathcal{A} = \{(1-\theta)/\theta \geq G^c, c > 2\}$ is a very high probability event (see Remarks 1 and 2).

For practical implementation, we also recommend endowing θ with a prior $\pi(\theta)$ in order to model the inherent uncertainty in θ and *adaptively* learn the true sparsity level from the data. To this end, we endow θ in (2.2) with a beta prior with shape parameters $a > 0, b > 0$,

$$\theta \sim \mathcal{B}(a, b). \tag{5.1}$$

Our prior specification for $(\boldsymbol{\beta}, \theta)$ is then given by $\pi(\boldsymbol{\beta}, \theta) = \pi(\boldsymbol{\beta} \mid \theta)\pi(\theta)$, where $\pi(\boldsymbol{\beta} \mid \theta)$ is as in (2.2) and $\pi(\theta)$ is as in (5.1). The complete log-posterior is then

$$\log \pi(\boldsymbol{\beta}, \theta \mid \mathbf{Y}) = \ell_n(\boldsymbol{\beta}) + \log \pi(\boldsymbol{\beta} \mid \theta) + \log \pi(\theta). \tag{5.2}$$

We use a variant of the EMVS algorithm [26] to iteratively solve for the MAP estimator $(\widehat{\boldsymbol{\beta}}, \widehat{\theta})$ in the optimization problem,

$$(\widehat{\boldsymbol{\beta}}, \widehat{\theta}) = \underset{\boldsymbol{\beta} \in \mathbb{R}^p, \theta \in (0,1)}{\arg\max} \ \log \pi(\boldsymbol{\beta}, \theta \mid \mathbf{Y}). \tag{5.3}$$

The EMVS approach of Ročková and George [26] introduces latent variables $\boldsymbol{\gamma} = (\gamma_1, \ldots, \gamma_G)$, $\gamma_G \in \{0, 1\}$, where $\gamma_g = 1$ indicates that the gth group of coefficients $\boldsymbol{\beta}_g$ should be included in the model. These indicator variables are treated as missing data in the E-step of our algorithm. To be precise, we reparameterize the SSGL prior (2.2) as a beta-Bernoulli prior,

$$\begin{aligned} \pi(\boldsymbol{\beta} \mid \boldsymbol{\gamma}) &= \prod_{g=1}^{G} \left[(1-\gamma_g)\boldsymbol{\Psi}(\boldsymbol{\beta}_g \mid \lambda_0) + \theta_g \boldsymbol{\Psi}(\boldsymbol{\beta}_g \mid \lambda_1) \right], \\ \pi(\boldsymbol{\gamma} \mid \theta) &= \prod_{g=1}^{G} \theta^{\gamma_g} (1-\theta)^{1-\gamma_g}, \end{aligned} \tag{5.4}$$

where $\boldsymbol{\gamma}$ is a binary vector. As shown in Appendix C, $\mathbb{E}[\gamma_g \mid \mathbf{Y}, \boldsymbol{\beta}, \theta] = p_g^{\star}(\boldsymbol{\beta}_g, \theta)$, where

$$p_g^{\star}(\boldsymbol{\beta}_g, \theta) = \frac{\theta \boldsymbol{\Psi}(\boldsymbol{\beta}_g \mid \lambda_1)}{\theta \boldsymbol{\Psi}(\boldsymbol{\beta}_g \mid \lambda_1) + (1-\theta)\boldsymbol{\Psi}(\boldsymbol{\beta}_g \mid \lambda_0)} \tag{5.5}$$

is the conditional posterior probability that $\boldsymbol{\beta}_g$ is drawn from the slab distribution rather than from the spike. In the E-step, we compute $p_g^{\star(t-1)} := p_g^{\star}(\boldsymbol{\beta}_g^{(t-1)}, \theta^{(t-1)}) = \mathbb{E}[\gamma_g \mid \mathbf{Y}, \boldsymbol{\beta}^{(t-1)}, \theta^{(t-1)}], g = 1, \ldots, G$. In the M-step, we then update θ as

$$\theta^{(t)} = \frac{a - 1 + \sum_{g=1}^{G} p_g^{\star(t-1)}}{a + b + G - 2}, \tag{5.6}$$

and $\boldsymbol{\beta}$ as

$$\boldsymbol{\beta}^{(t)} = \arg\max_{\boldsymbol{\beta}} \left\{ \ell(\boldsymbol{\beta}) - \sum_{g=1}^{G} \lambda_g^{\star(t-1)} \|\boldsymbol{\beta}_g\|_2 \right\}, \tag{5.7}$$

where each $\lambda_g^{\star(t-1)} = \lambda_1 p_g^{\star(t-1)} + \lambda_0(1 - p_g^{\star(t-1)})$ is an *adaptive* weight ensuring that insignificant groups are shrunk aggressively to zero, while significant groups incur minimal shrinkage. The objective (5.7) is simply a group lasso optimization with known group-specific weights $(\lambda_1^{\star(t-1)}, \ldots, \lambda_G^{\star(t-1)})$. In Appendix C, we describe how to efficiently solve (5.7). In summary, our EMVS algorithm proceeds as follows.

1. Initialize $(\boldsymbol{\beta}^{(0)}, \theta^{(0)})$. For example, we can initialize $\boldsymbol{\beta}^{(0)} = \mathbf{0}_p$ and $\theta^{(0)} = 0.5$.

2. For $t = 1, 2, \ldots$, repeat until convergence:

   i. **E-step**: For $g = 1, \ldots, G$, compute $p_g^{\star(t-1)} = p_g^{\star}(\boldsymbol{\beta}_g^{(t-1)}, \theta^{(t-1)})$ as in (5.5).

   ii. **M-step**: Update $\theta^{(t)}$ as in (5.6) and $\boldsymbol{\beta}^{(t)}$ as in (5.7).

To determine convergence of the algorithm, we recommend using the criterion $\|\boldsymbol{\beta}^{(t)} - \boldsymbol{\beta}^{(t-1)}\|_2^2 / \|\boldsymbol{\beta}^{(t-1)}\|_2^2 < \varepsilon$, where ε is a small value (e.g. $\varepsilon = 10^{-6}$). Since the EM algorithm has the ascent property, our algorithm is guaranteed to converge to a local mode. However, the SSGL objective function is nonconvex, and thus, convergence to the global mode is not guaranteed. Other nonconvex group penalties such as the group minimax concave penalty and the group smoothly clipped absolute deviation penalty also cannot guarantee global convergence [7]. However, we have not found local convergence to be a practical issue for the SSGL.

5.2 Choice of hyperparameters

The performance of the SSGL is mainly governed by the three parameters $(\lambda_0, \lambda_1, \theta)$ in the prior (5.4) on $\boldsymbol{\beta}$. We recommend fixing the slab hyperparameter $\lambda_1 = 1$, so that the $\boldsymbol{\beta}_g$'s with large entries incur very minimal shrinkage. To induce sparsity, the mixing proportion θ should also be small with high probability, so that most of the $\boldsymbol{\beta}_g$'s belong to the spike density. To this end, we recommend setting $a = 1, b = G$ for the $\mathcal{B}(a, b)$ prior (5.1) on θ. This ensures that most of the $\boldsymbol{\beta}_g$'s will be shrunk to zero.

The spike hyperparameter λ_0 in (5.4) controls how sparse our final model is, with larger values of λ_0 leading to more sparsity. For unstructured GLMs (1.2) with the SSL prior (2.3), Tang et al. [30] recommended tuning λ_0 from cross-validation (CV). However, we found CV to be very time-consuming, even for moderate-sized p. CV also tended to produce overly dense models, since CV prioritizes out-of-sample predictive accuracy rather than correct variable selection.

As an alternative to CV, we propose choosing λ_0 in (5.4) using the generalized information criterion (GIC) of Fan and Tang [13]. Let $\widehat{\boldsymbol{\beta}}_{\lambda_0}$ denote the MAP estimator (2.5) for $\boldsymbol{\beta}$ with λ_0 as the spike hyperparameter in the SSGL prior. The GIC is defined as

$$\text{GIC}(\lambda_0) = \frac{1}{n} \left\{ D(\widehat{\boldsymbol{\mu}}_{\lambda_0}; \mathbf{Y}) + a_n \times \# \text{ of nonzero elements in } \widehat{\boldsymbol{\beta}}_{\lambda_0} \right\}, \tag{5.8}$$

where a_n is a diverging sequence that penalizes model size, and $D(\widehat{\boldsymbol{\mu}}; \mathbf{Y}) = 2\{\ell_n(\mathbf{Y}; \mathbf{Y}) - \ell(\widehat{\boldsymbol{\mu}}_{\lambda_0}; \mathbf{Y})\}$ is the deviance function. Here, $\ell_n(\boldsymbol{\mu}; \mathbf{Y})$ denotes the log-likelihood (1.4) reexpressed as a function of the expectation $\boldsymbol{\mu} = b'(\xi(\mathbf{X}\boldsymbol{\beta}))$; hence, $\widehat{\boldsymbol{\mu}}_{\lambda_0} = b'(\xi(\mathbf{X}\widehat{\boldsymbol{\beta}}_{\lambda_0}))$, while $\ell_n(\mathbf{Y}; \mathbf{Y})$ is the saturated model with $\boldsymbol{\mu} = \mathbf{Y}$. By Theorems 1 and 2 of Fan and Tang [13], $\widehat{\boldsymbol{\beta}}_{\lambda_0}$ consistently estimates the true model if λ_0 is chosen to minimize the GIC criterion (5.8) and the sequence a_n diverges at an appropriate rate with n. Theorems 1 and 2 of Fan and Tang [13] hold for *any* sparse estimator $\widehat{\boldsymbol{\beta}}$, including the SSGL MAP estimator (2.5), even when p grows exponentially fast with n.

As recommended by Fan and Tang [13], we set $a_n = \{\log(\log n)\} \log p$ in (5.8). We choose the spike parameter λ_0 which minimizes GIC(λ_0) over an equispaced grid $\lambda_0 \in (\lambda_1, \ldots, \lambda_{\max}]$, where $\lambda_{\max}$ is the smallest value of λ_0 so that $\widehat{\boldsymbol{\beta}} = \mathbf{0}_p$. It is not hard to see that $\lambda_{\max} = \max_{1 \leq g \leq G} \|\nabla_g \, \ell_n(\mathbf{0}_p)\|_2$, where ∇_g denotes the subvector of the gradient (1.5) corresponding to the gth group. Typically, $\lambda_{\max}$ has an analytical form. For example, in logistic regression, $\lambda_{\max} = \max_{1 \leq g \leq G} \|0.25\mathbf{X}_g^\top(\mathbf{Y} - 0.5\mathbf{1}_n)\|_2$, and in Poisson regression, $\lambda_{\max} = \max_{1 \leq g \leq G} \|\mathbf{X}_g^\top(\mathbf{Y} - \mathbf{1}_n)\|_2$.

To account for potentially different group sizes m_g, we further rescale λ_0 for each gth group, so that the spike parameter for each $\boldsymbol{\beta}_g$ is $\lambda_{0g} = \lambda_0 \sqrt{m_g}$. As discussed in Huang et al. [15], scaling of the regularization penalty by group size is needed to ensure that groups are not unfairly penalized simply for being smaller or erroneously included simply for being larger.

6 Simulation studies

6.1 Setup and performance metrics

We investigated the performance of the SSGL prior (2.2) in numerical experiments with $G < n$ and $G > n$. We considered grouped logistic regression for binary data and grouped Poisson regression for count data. In particular, Experiments 3 and 4 in Sections 6.2 and 6.3 are meant to mimic two real applications where our methodology is especially useful: a) semiparametric additive models with continuous covariates, and b) genetic association studies involving single nucleotide polymorphisms (SNPs) [7]. In Appendix D, we also present some simulation results for grouped negative binomial regression with a log link, which is an example of our method with a *non*-canonical link function.

In semiparametric additive models (Experiment 3 in Sections 6.2 and 6.3), we flexibly model the effects of continuous covariates x_j on the mean response as univariate functions $f_j(x_j)$. The f_j's are approximated using linear combinations of K basis functions $g_{jk}(x_j)$, i.e. $f_j(x_j) \approx \sum_{k=1}^K \beta_{jk} g_{jk}(x_j)$. The jth main effect f_j is then estimated as $\widehat{f}_j(x_j) = 0$ if $\boldsymbol{\beta}_j = (\beta_{j1}, \ldots, \beta_{jK})^\top = \mathbf{0}_K$ or as $\widehat{f}_j(x_j) \neq 0$ if $\boldsymbol{\beta}_j \neq \mathbf{0}_K$.

In the simulated genetic association studies (Experiment 4 in Sections 6.2 and 6.3), the responses are either binary or count phenotypes, and the covariates are simulated SNPs. SNPs are categorical variables coded as one of three values {"0", "1", or "2"}, depending on the number of minor alleles present [7]. We thus represent each SNP as a factor with two levels, i.e. a group of two indicator variables. Assuming that "2" is the baseline, we can represent each jth SNP x_j with two dummy variables $\mathbb{I}(x_j = 0)$ and $\mathbb{I}(x_j = 1)$. If $x_j = 2$, then $\mathbb{I}(x_j = 0) = \mathbb{I}(x_j = 1) = 0$.

We have implemented SSGL for GLMs in the R package `SSGL`. In all of our experiments, we chose the hyperparameters in the SSGL prior as described in Section 5.2. We compared the performance of SSGL to the group lasso (gLASSO) [35], the group minimax concave penalty (gMCP) [7], and the group smoothly clipped absolute deviation (gSCAD) [7]. Unlike the SSGL, these other methods only have a single regularization parameter $\lambda > 0$ controlling the sparsity. We implemented gLASSO, gMCP, and gSCAD using the R package `grpreg`. This package does not currently compute the GIC criterion (5.8). However, `grpreg` does support model selection using the extended Bayesian Information Criterion (EBIC), which is a very similar criterion to (5.8) that also consistently estimates the true active set when p grows faster than n [10]. Thus, we used EBIC to choose λ for gLASSO, gMCP, and gSCAD.

We computed the following performance metrics: mean squared error (MSE), mean squared prediction error (MSPE), sensitivity (Sens), specificity (Spec), precision (Prec), and Matthews Correlation Coefficient (MCC), defined as

$$\begin{aligned}
&\text{MSE} = \frac{1}{p}\|\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}\|_2^2, \ \ \text{MSPE} = \frac{1}{n_{\text{test}}}\sum_{i=1}^{n}(y_{i,\text{test}} - \widehat{y}_{i,\text{test}})^2, \\
&\text{Sens} = \frac{\text{TP}}{\text{TP}+\text{FN}}, \ \ \text{Spec} = \frac{\text{TN}}{\text{TN}+\text{FP}}, \ \ \text{Prec} = \frac{\text{TP}}{\text{TP}+\text{FP}}, \\
&\text{MCC} = \frac{\text{TP}\times\text{TN} - \text{FP}\times\text{FN}}{\sqrt{(\text{TP}+\text{FP})(\text{TP}+\text{FN})(\text{TN}+\text{FP})(\text{TN}+\text{FN})}},
\end{aligned}$$

where TP, TN, FP, and FN are the number of true positives, true negatives, false positives, and false negatives respectively. The MSPE was computed using $n_{\text{test}} = 100$ out-of-sample test points $\{(\mathbf{x}_{i,\text{test}}, y_{i,\text{test}})\}_{i=1}^{n_{\text{test}}}$, and $\widehat{y}_{i,\text{test}} = b'(\widehat{\theta}_{i,\text{test}})$, where $\widehat{\theta}_i$ is the predicted natural parameter with $\mathbf{x}_{i,\text{test}}$ in (1.3). MCC takes values between -1 and 1, with higher values indicating better overall variable selection performance. For the logistic regression experiments in Section 6.2, we also recorded the area under the receiver operating characteristic curve (AUC) on the test data.

6.2 Grouped logistic regression

For logistic regression, we have $h(u) = \log\{u/(1-u)\}$ and $b(u) = \log(1+e^u)$ in (1.3), so that the left-hand side of (1.3) is $\log(\theta_i/(1-\theta_i))$, and the responses are independently drawn from $y_i \mid \mathbf{x}_i \sim \text{Bernoulli}(1/(1+\exp(-\theta_i))), i = 1, \ldots, n$. We considered the following four experiments:

Experiment 1 ($G < n$). We set $n = 100$ and $G = 40$. We simulated the groups to have high within-group correlation. Namely, the rows of each $\mathbf{X}_g$ in (1.3) were generated independently from a multivariate Gaussian with mean $\mathbf{0}_{m_g}$ and covariance matrix $\sigma^2\boldsymbol{\Omega}_g$, where $\sigma^2 = 1$ and $\boldsymbol{\Omega}_g$ had all off-diagonal entries equal to 0.8 and diagonal entries equal to one. The group sizes m_g were randomly chosen from $\{3, 4, 5\}$, and $s_0 = 5$ of the vectors $\boldsymbol{\beta}_g$ were randomly chosen to be nonzero with entries randomly chosen from $\{-2.5, -2, -1.5, 1.5, 2, 2.5\}$. Then we modeled

$$\log\left(\frac{\theta}{1-\theta}\right) = \mathbf{x}^\top\boldsymbol{\beta}.$$

Experiment 2 ($G > n$). We repeated Experiment 1 with $n = 100$, but we increased the number of groups to $G = 200$.

Experiment 3 (semiparametric regression). We set $n = 100$ and $G = 80$ and generated the entries of the $n \times G$ design matrix $\mathbf{X}$ from independent $\text{Uniform}(-1, 1)$ random variables. Then we modeled

$$\log\left(\frac{\theta}{1-\theta}\right) = 5\sin(3x_1) - 5x_5e^{0.5x_5^2},$$

i.e. only the covariates x_1 and x_5 had a non-null and nonlinear effect on the mean response, and $f_j(x_j) = 0$ for all $j \notin \{1, 5\}$. We represented each covariate as a six-term B-spline basis expansion.

Experiment 4 (genetic association study with $G \gg n$). We set $n = 100$ and $G = 800$. We first generated an $n \times G$ latent matrix $\mathbf{X}$, where each ith row $\mathbf{x}_i$ was drawn from a multivariate Gaussian with mean $\mathbf{0}_G$ and covariance matrix $\mathbf{\Gamma}$, where the (j,k)th entry of $\mathbf{\Gamma}$ was $\Gamma_{jk} = 0.5^{|j-k|}$. Then each entry in $\mathbf{X}$ was trichotomized as "0," "1," or "2" according to whether it was smaller than $\Phi^{-1}(1/3)$, between $\Phi^{-1}(1/3)$ and $\Phi^{-1}(2/3)$, or greater than $\Phi^{-1}(2/3)$. Here, $\Phi^{-1}(\cdot)$ denotes the inverse cumulative distribution function (cdf) of a standard normal. Thus, the entries in our final design matrix $\mathbf{X}$ were categorical SNP variables with three levels ("0," "1", or "2"). Letting "2" denote the baseline category, we then modeled

$$\begin{aligned}\log\left(\frac{\theta}{1-\theta}\right) = {} & 2.5\mathbb{I}(x_1 = 0) - 2.5\mathbb{I}(x_1 = 1) + 1.4\mathbb{I}(x_{15} = 0) + 2.2\mathbb{I}(x_{15} = 1) \\ & - 1.6\mathbb{I}(x_{25} = 0) - 1.8\mathbb{I}(x_{25} = 1).\end{aligned}$$

i.e. only the SNPs x_1, x_{15}, and x_{25} had a significant association with the phenotype.

We repeated each of the four simulations 200 times. Table 1 reports our experimental results averaged across the 200 replications. Note that in Experiment 3, there is not a "true" $\boldsymbol{\beta}$; rather, each jth ground truth function $f_j(\mathbf{x}_j)$ is estimated by $\widehat{f}_j(\mathbf{x}_j) = \widetilde{\mathbf{X}}_j \widehat{\boldsymbol{\beta}}_j$, where the (i,k)th entry of $\widetilde{\mathbf{X}}_j$ is the kth B-spline basis term $g_{jk}(x_{ij})$. Thus, for Experiment 3, we do not report the MSE.

In all experiments, SSGL had the highest MCC, indicating the best overall ability to distinguish nonzero and zero groups. In Experiments 1, 2, and 4, SSGL estimated the regression coefficients $\boldsymbol{\beta}$ the best, with the lowest MSE. The AUC was also highest for SSGL in Experiments 1, 2, and 4, indicating that SSGL had the best ability to distinguish positive and negative classes.

The advantages of SSGL were especially pronounced in the $G \gg n$ setting (Experiment 4), where the average MSE and MSPE were substantially lower and the AUC was much higher than those of the competing methods. In Experiment 4, gLASSO, gMCP, and gSCAD all faced difficulty picking up the true signals (sometimes estimating a null model), which led to much lower sensitivity and greater estimation and prediction error. On the other hand, the SSGL also estimated more false positives, leading to lower precision.

In Experiment 3, gLASSO outperformed the other methods in terms of prediction, with a lower MSPE and higher AUC. However, gLASSO tended to have many false positives in this experiment, and thus, its MCC was much lower than the competing methods. In this experiment, the group selection performance of gLASSO was much worse than the others, despite having the best predictive performance.

6.3 Grouped Poisson regression

For Poisson regression, we have $h(u) = \log u$ and $b(u) = e^u$ in (1.3), so that the left-hand side of (1.3) is $\log(\theta_i)$, and the response variables are independently drawn from $y_i \mid \mathbf{x}_i \sim \text{Poisson}(\exp(\theta_i)), i = 1, \ldots, n$. Our experiments mimicked those of Section 6.2, except in order to ensure realistic count values, we could not allow the rate parameter $\exp(\theta_i)$ to be too large. Thus, we decreased the magnitude of the entries in the design matrix $\mathbf{X}$ and/or the signal sizes. Our four simulations were as follows:

Table 1: Simulation results for grouped logistic regression under the SSGL, gLASSO, gMCP, and gSCAD models, averaged across 200 replicates. The empirical standard error is reported in parentheses below the average.

Experiment 1

	MSE	MSPE	AUC	Sens	Spec	Prec	MCC
SSGL	**0.471**	**0.168**	**0.916**	**0.593**	**0.992**	**0.927**	**0.706**
	(0.058)	(0.026)	(0.043)	(0.182)	(0.015)	(0.132)	(0.146)
gLASSO	0.507	0.191	0.856	0.522	0.986	0.887	0.629
	(0.052)	(0.017)	(0.053)	(0.216)	(0.022)	(0.163)	(0.153)
gMCP	0.496	0.173	0.846	0.506	0.990	0.919	0.636
	(0.054)	(0.025)	(0.052)	(0.204)	(0.019)	(0.153)	(0.158)
gSCAD	0.507	0.191	0.856	0.521	0.986	0.888	0.629
	(0.052)	(0.017)	(0.052)	(0.215)	(0.022)	(0.162)	(0.152)

Experiment 2

	MSE	MSPE	AUC	Sens	Spec	Prec	MCC
SSGL	**0.096**	0.172	**0.921**	**0.556**	0.992	0.695	**0.603**
	(0.012)	(0.022)	(0.036)	(0.183)	(0.007)	(0.233)	(0.179)
gLASSO	0.101	0.186	0.879	0.496	0.990	0.693	0.546
	(0.011)	(0.015)	(0.039)	(0.216)	(0.011)	(0.268)	(0.163)
gMCP	0.099	**0.168**	0.862	0.467	**0.994**	**0.762**	0.564
	(0.010)	(0.023)	(0.039)	(0.210)	(0.008)	(0.251)	(0.170)
gSCAD	0.101	0.186	0.879	0.496	0.990	0.693	0.546
	(0.010)	(0.015)	(0.039)	(0.216)	(0.011)	(0.268)	(0.163)

Experiment 3

	MSPE	AUC	Sens	Spec	Prec	MCC
SSGL	0.106	0.972	0.980	**1**	**1**	**0.991**
	(0.028)	(0.019)	(0.098)	(0)	(0)	(0.051)
gLASSO	**0.053**	**0.998**	**1**	0.796	0.134	0.320
	(0.028)	(0.005)	(0)	(0.077)	(0.085)	(0.099)
gMCP	0.106	0.966	0.985	0.999	0.998	0.987
	(0.018)	(0.018)	(0.086)	(0.001)	(0.024)	(0.060)
gSCAD	0.103	0.970	0.995	0.996	0.917	0.949
	(0.019)	(0.013)	(0.050)	(0.008)	(0.162)	(0.099)

Experiment 4

	MSE	MSPE	AUC	Sens	Spec	Prec	MCC
SSGL	**0.009**	**0.139**	**0.943**	**0.927**	0.993	0.368	**0.587**
	(0.002)	(0.016)	(0.024)	(0.150)	(0.004)	(0.171)	(0.240)
gLASSO	0.015	0.229	0.756	0.453	**0.999**	0.810	**0.587**
	(0.001)	(0.011)	(0.087)	(0.246)	(0.001)	(0.326)	(0.240)
gMCP	0.015	0.223	0.752	0.442	**0.999**	**0.811**	0.581
	(0.001)	(0.015)	(0.085)	(0.243)	(0.001)	(0.329)	(0.243)
gSCAD	0.015	0.229	0.756	0.453	**0.999**	0.810	0.571
	(0.001)	(0.011)	(0.087)	(0.246)	(0.001)	(0.326)	(0.148)

Experiment 1 ($\boldsymbol{G < n}$). With $n = 100$ and $G = 40$, we simulated $\mathbf{X}$ and $\boldsymbol{\beta}$ the same way as we did in Experiment 1 of Section 6.2, except we set $\sigma^2 = 0.3$, and the entries in the $s_0 = 5$ randomly chosen nonzero vectors were randomly chosen from $\{-1.5, -1, 1, 1.5\}$. Then we modeled

$$\log(\theta) = \mathbf{x}^\top \boldsymbol{\beta}.$$

Experiment 2 ($\boldsymbol{G > n}$). We repeated Experiment 1 with $n = 100$, but we increased the number of groups to $G = 200$.

Experiment 3 (semiparametric regression). We set $n = 100$ and $G = 80$ and generated the entries of the $n \times G$ design matrix $\mathbf{X}$ from independent Uniform$(-1, 1)$ random 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.

Experiment 4 (genetic association study with $\boldsymbol{G \gg n}$). With $n = 100$ and $G = 800$, we simulated the SNP categorical variables (“0”, “1”, or “2”) in $\mathbf{X}$ the same way that we did in Experiment 4 of Section 6.2. Then we modeled

$$\begin{aligned} \log{(\theta)} = \ & 2\mathbb{I}(x_1 = 0) - 2\mathbb{I}(x_1 = 1) + 1.6\mathbb{I}(x_{15} = 0) + 1.6\mathbb{I}(x_{15} = 1) \\ & - 1.4\mathbb{I}(x_{25} = 0) - 1.8\mathbb{I}(x_{25} = 1). \end{aligned}$$

Each experiment was repeated 200 times. Table 2 reports the results averaged across the 200 replications. As explained why in Section 6.2, we did not report the MSE in Experiment 3.

For grouped Poisson regression, SSGL had uniformly the best group selection performance, with the highest sensitivity, specificity, precision, and MCC in all experiments. In Experiments 1, 2, and 4, SSGL also had the lowest MSE, indicating the best estimation performance of the regression coefficients $\boldsymbol{\beta}$. Finally, SSGL had either the lowest or the second lowest MSPE in the four experimental settings, showing that it performed well in terms of prediction on new test data.

Once again, the SSGL demonstrated its greatest advantage over the competing methods in the $G \gg n$ setting (Experiment 4), where the MSE and MSPE were substantially lower and the precision and MCC were substantially higher for SSGL than for gLASSO, gMCP, or gSCAD. This suggests that SSGL is especially well-suited for estimation and group selection in datasets where G is much larger than n.

Table 2: Simulation results for grouped Poisson regression under the SSGL, gLASSO, gMCP, and gSCAD models, averaged across 200 replicates. The empirical standard error is reported in parentheses below the average.

Experiment 1

	MSE	MSPE	Sens	Spec	Prec	MCC
SSGL	**0.025**	**2.007**	**0.949**	**0.993**	**0.957**	**0.945**
	(0.039)	(1.446)	(0.130)	(0.020)	(0.112)	(0.129)
gLASSO	0.0828	8.440	0.912	0.838	0.489	0.590
	(0.036)	(6.338)	(0.198)	(0.087)	(0.179)	(0.140)
gMCP	0.035	2.033	0.925	0.983	0.901	0.897
	(0.054)	(1.481)	(0.155)	(0.029)	(0.158)	(0.163)
gSCAD	0.035	2.018	0.927	0.985	0.909	0.903
	(0.053)	(1.400)	(0.150)	(0.027)	(0.155)	(0.160)

Experiment 2

	MSE	MSPE	Sens	Spec	Prec	MCC
SSGL	**0.0155**	4.698	**0.524**	**0.996**	**0.745**	**0.609**
	(0.010)	(2.524)	(0.257)	(0.006)	(0.301)	(0.262)
gLASSO	0.0165	26.43	0.337	0.989	0.481	0.357
	(0.003)	(19.20)	(0.286)	(0.016)	(0.390)	(0.243)
gMCP	0.0162	**4.538**	0.510	0.995	0.740	0.596
	(0.010)	(2.349)	(0.257)	(0.007)	(0.296)	(0.257)
gSCAD	0.0167	6.079	0.486	0.995	0.716	0.566
	(0.010)	(4.490)	(0.279)	(0.006)	(0.311)	(0.268)

Experiment 3

	MSPE	Sens	Spec	Prec	MCC
SSGL	1.825	**0.993**	**0.999**	**0.986**	**0.987**
	(0.738)	(0.061)	(0.003)	(0.070)	(0.054)
gLASSO	5.264	0.775	**0.999**	0.764	0.767
	(2.557)	(0.413)	(0.003)	(0.409)	(0.407)
gMCP	**1.822**	0.775	**0.999**	0.764	0.767
	(0.654)	(0.413)	(0.003)	(0.409)	(0.407)
gSCAD	5.264	0.775	**0.999**	0.764	0.767
	(2.557)	(0.413)	(0.003)	(0.409)	(0.407)

Experiment 4

	MSE	MSPE	Sens	Spec	Prec	MCC
SSGL	**0.0001**	**2.412**	**1**	**1**	**0.989**	**0.994**
	(0.0001)	(0.932)	(0)	(0)	(0.052)	(0.028)
gLASSO	0.005	9.096	0.965	0.982	0.175	0.402
	(0.002)	(5.681)	(0.127)	(0.005)	(0.048)	(0.061)
gMCP	0.003	2.423	0.867	0.997	0.721	0.777
	(0.006)	(1.163)	(0.287)	(0.005)	(0.365)	(0.325)
gSCAD	0.002	2.499	0.897	0.998	0.794	0.835
	(0.005)	(1.129)	(0.260)	(0.004)	(0.332)	(0.297)

7 Application to HIV drug resistance data

One of the challenges with drug treatments for HIV is the virus' ability to rapidly mutate and gain resistance to these drugs. The Stanford HIV Drug Resistance Database maintains isolates of HIV that were extracted from infected individuals and sequenced. In a study conducted by Rhee et al. [24], these isolates were used to predict resistance to 16 antiretroviral drugs used in HIV therapy. The outcome in this study was a measure of drug susceptibility, where a higher value indicated greater resistance to the drug.

For our real data application, we focus on the drug Nelfinavir, a protease inhibitor (PI), since the data from the study by Rhee et al. [24] is publicly available.[1] Protease genes are made up of sequences of amino acids. A mutation occurs whenever a position in the sequence contains a different amino acid than the usual amino acid found at that position. Our dataset consists of $n = 842$ isolates and $G = 82$ groups, with a total of $p = 361$ covariates. Each of the G groups represents a specific position in the protease amino acid sequence, and within each gth group, the covariates are 1/0 indicator variables indicating the presence or absence of a specific amino acid mutation at the gth position. For example, if Valine is found at position 13 instead of the usual amino acid at position 13, then the covariate value for Valine in group $g = 13$ would be a "1" instead of a "0."

In Rhee et al. [24], a susceptibility index greater than 20 was considered to be "highly resistant" for PIs. Accordingly, we dichotomized the outcome into two categories according to whether the susceptibility value was greater than 20 or not. This led to 300 positive cases ("highly resistant") and 542 negative cases ("not highly resistant"). We then fit grouped logistic regression models to the data with the dichotomized responses.

In our study, we are mainly interested in prediction of drug resistance to Nelfinavir in HIV-infected individuals [24]. Nevertheless, group regularization can help to prevent overfitting and thus improve model generalization and classification accuracy. We examined the performance of the SSGL model on this dataset and compared it with gLASSO, gMCP, and gSCAD. The hyperparameters and tuning parameters were all chosen the same way as they were in Section 6.

To perform group selection, we fit the four grouped logistic regression models to the full dataset. Next, we assessed the models' predictive power. To do so, we randomly divided the dataset into 70% training and 30% test data (i.e. 590 training observations and 252 test observations). We then fit the models to the training data and evaluated the MSPE and AUC on the held-out test set. We repeated this process 200 times, so that we had 200 different test sets on which to evaluate the methods.

Our results are shown in Table 3. SSGL selected 24 positions. In contrast, gLASSO, gMCP, and gSCAD all selected more parsimonious models. In particular, gLASSO selected only two positions in the amino acid sequence (positions 25 and 30). These positions were also selected by SSGL, gMCP, and gSCAD. However, the gLASSO's out-of-sample MSPE was the highest and its AUC was the lowest.

Despite the fact that the SSGL selected the most sequence positions, SSGL still had the lowest out-of-sample MSPE and the highest out-of-sample AUC. This indicates that the SSGL model did not suffer from overfitting and possessed the best ability to correctly classify whether HIV patients were highly resistant to Nelfinavir or not. On this particular dataset, the SSGL appears to achieve the best tradeoff between model parsimony (24 active positions out of 82) and predictive accuracy.

Our analysis suggests that there may be practical benefits to the SSGL (2.2) having a slab density $\boldsymbol{\Psi}(\cdot \mid \lambda_1)$, in addition to a spike density $\boldsymbol{\Psi}(\cdot \mid \lambda_0)$. In contrast, gLASSO, gMCP, and gSCAD only have a single regularization parameter $\lambda > 0$ controlling the sparsity level. Because of this, these other methods may overshrink many of the signals in the data when λ is large (as suggested by the first column in Table 3). In contrast, the SSGL's slab density is specifically designed to prevent overshrinkage of true nonzero groups, thus allowing SSGL to detect more signals in the data.

[1] https://myweb.uiowa.edu/pbreheny/data/Rhee2006.html. Accessed June 15, 2023.

Table 3: Results for SSGL, gLASSO, gSCAD, and gMCP on the HIV drug resistance dataset. The MSPE and AUC were averaged across 200 test sets, and the empirical standard errors are shown in parentheses.

	Number of positions selected	MSPE	AUC
SSGL	24	**0.098** (0.012)	**0.937** (0.013)
gLASSO	2	0.162 (0.008)	0.805 (0.023)
gMCP	4	0.110 (0.016)	0.914 (0.033)
gSCAD	8	0.144 (0.020)	0.835 (0.050)

8 Discussion

In this paper, we have extended the SSGL prior of Bai et al. [2] from Gaussian linear regression models to GLMs (1.3) under the general exponential family (1.1). This enables the SSGL (2.2) to be applied to binary regression, Poisson regression, negative binomial regression, and gamma regression, among others.

In contrast to high-dimensional Bayesian linear regression models, theoretical developments for high-dimensional Bayesian GLMs have been much slower. Our work builds upon and goes beyond those of Jiang [17], Jeong and Ghosal [16], and Tang and Martin [29] in several ways. First, we conducted our analysis under a *continuous* heavy-tailed spike-and-slab prior (2.2), rather than a point mass spike-and-slab prior. Second, we considered the more general grouped GLM setting (1.3). By treating unstructured GLMs (1.2) as a special case of (1.3), we also obtained the first theoretical results for the SSL prior (2.3) of Ročková and George [27] in high-dimensional GLMs. Finally, we analyzed *both* the MAP estimator (2.5) *and* the full posterior distribution (2.4), confirming that the rate of convergence is the same for both. This is in stark contrast to the original group lasso (2.6). We showed via extensive simulation studies and an analysis of an HIV drug resistance dataset that the SSGL performs very well in grouped GLMs when $G < n$ and $G > n$.

Although we have addressed Bayesian estimation and variable selection in high-dimensional GLMs, we have not addressed the question of uncertainty quantification. While Theorem 3 implies that the full SSGL posterior provides valid inference in terms of the size of credible sets, computationally efficient posterior sampling in high dimensions is a more challenging task than MAP estimation. Apart from the size of the posterior credible sets, theoretical questions about the asymptotic coverage of Bayesian credible sets and/or a Bernstein-von Mises theorem for the limiting shape of the posterior are also important questions that have not been addressed in the literature. We plan to develop scalable algorithms and theory for uncertainty quantification in high-dimensional Bayesian GLMs in future research.

Acknowledgments

The author is grateful to Dr. Seonghyun Jeong for helpful discussions which led to derivations of the theoretical results in this paper. This work was partially supported by National Science Foundation grant DMS-2015528.

References

[1] Atchadé, Y. A. (2017). On the contraction properties of some high-dimensional quasi-posterior distributions. *The Annals of Statistics*, 45(5):2248 – 2273.

[2] 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.

[3] Baraud, Y. (2010). A Bernstein-type inequality for suprema of random processes with applications to model selection in non-Gaussian regression. *Bernoulli*, 16(4):1064 – 1085.

[4] Bhattacharya, A., Pati, D., Pillai, N. S., and Dunson, D. B. (2015). Dirichlet–Laplace priors for optimal shrinkage. *Journal of the American Statistical Association*, 110(512):1479–1490.

[5] Blazère, M., Loubes, J.-M., and Gamboa, F. (2014). Oracle inequalities for a group lasso procedure applied to generalized linear models in high dimension. *IEEE Transactions on Information Theory*, 60(4):2303–2318.

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

[7] 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.

[8] Castillo, I. and Mismer, R. (2018). Empirical Bayes analysis of spike and slab posterior distributions. *Electronic Journal of Statistics*, 12(2):3953 – 4001.

[9] 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.

[10] Chen, J. and Chen, Z. (2008). Extended Bayesian information criteria for model selection with large model spaces. *Biometrika*, 95(3):759–771.

[11] Donoho, D. L. and Johnstone, I. M. (1994). Ideal spatial adaptation by wavelet shrinkage. *Biometrika*, 81(3):425–455.

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

[13] Fan, Y. and Tang, C. Y. (2013). Tuning parameter selection in high dimensional penalized likelihood. *Journal of the Royal Statistical Society. Series B (Statistical Methodology)*, 75(3):531–552.

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

[15] Huang, J., Breheny, P., and Ma, S. (2012). A selective review of group selection in high-dimensional models. *Statistical Science*, 27(4):481 – 499.

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

[17] Jiang, W. (2007). Bayesian variable selection for high dimensional generalized linear models: Convergence rates of the fitted densities. *The Annals of Statistics*, 35(4):1487–1511.

[18] Johnstone, I. M. and Silverman, B. W. (2004). Needles and straw in haystacks: Empirical Bayes estimates of possibly sparse sequences. *The Annals of Statistics*, 32(4):1594 – 1649.

[19] Lee, K. and Cao, X. (2021). Bayesian group selection in logistic regression with application to MRI data analysis. *Biometrics*, 77(2):391–400.

[20] Lian, H. (2012). Variable selection for high-dimensional generalized varying-coefficient models. *Statistica Sinica*, 22(4):1563–1588.

[21] McCullagh, P. and Nelder, J. (1989). *Generalized Linear Models, Second Edition*. Chapman & Hall.

[22] Meier, L., Geer, S. V. D., and Bühlmann, P. (2008). The group lasso for logistic regression. *Journal of the Royal Statistical Society Series B*, 70(1):53–71.

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

[24] Rhee, S.-Y., Taylor, J., Wadhera, G., Ben-Hur, A., Brutlag, D. L., and Shafer, R. W. (2006). Genotypic predictors of human immunodeficiency virus type 1 drug resistance. *Proceedings of the National Academy of Sciences*, 103(46):17355–17360.

[25] Rousseau, J. (2016). On the frequentist properties of Bayesian nonparametric methods. *Annual Review of Statistics and Its Application*, 3(1):211–231.

[26] Ročková, V. and George, E. I. (2014). EMVS: The EM approach to Bayesian variable selection. *Journal of the American Statistical Association*, 109(506):828–846.

[27] Ročková, V. and George, E. I. (2018). The spike-and-slab lasso. *Journal of the American Statistical Association*, 113(521):431–444.

[28] Shen, Y., Solís-Lemus, C., and Deshpande, S. K. (2022). Sparse Gaussian chain graphs with the spike-and-slab lasso: Algorithms and asymptotics. *arXiv preprint arXiv:2207.07020*.

[29] Tang, Y. and Martin, R. (2023). Empirical Bayes inference in sparse high-dimensional generalized linear models. *arXiv preprint arXiv:2303.07854*.

[30] Tang, Z., Shen, Y., Li, Y., Zhang, X., Wen, J., Qian, C., Zhuang, W., Shi, X., and Yi, N. (2017a). Group spike-and-slab lasso generalized linear models for disease prediction and associated genes detection by incorporating pathway information. *Bioinformatics*, 34(6):901–910.

[31] Tang, Z., Shen, Y., Zhang, X., and Yi, N. (2017b). The spike-and-slab lasso generalized linear models for prediction and associated genes detection. *Genetics*, 205(1):77–88.

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

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

[34] Wen, Z., Yu, T., Yu, Z., and Li, Y. (2019). Grouped sparse Bayesian learning for voxel selection in multivoxel pattern analysis of fMRI data. *NeuroImage*, 184:417–430.

[35] 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.

[36] Zhao, P. and Yu, B. (2006). On model selection consistency of lasso. *Journal of Machine Learning Research*, 7(90):2541–2563.

A Proofs for Section 3

Before proving Theorem 1, it will be necessary to define some terminology. Letting $\pi(\boldsymbol{\beta})$ denote the SSGL prior (2.2), 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 (2.5), it can then be seen from (1.5) and (A.3) that any local maximizer of (2.5) must satisfy

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

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 (2.5),

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

Proof of Theorem 1. Recall that $S_0 \subset \{1, \ldots, p\}$ consists of the indices of the true nonzero groups, i.e. $\boldsymbol{\beta}_{0g} \neq \mathbf{0}_{m_g}$ for all $g \in S_0$, while $\boldsymbol{\beta}_{0g} = \mathbf{0}_{m_g}$ for all $g \in S_0^c$. Recall that the cardinality of S_0 is $|S_0| = s_0$. Let $\boldsymbol{\beta}_{S_0}$ denote the subvector of $\boldsymbol{\beta}$ with all groups in S_0, and $\boldsymbol{\beta}_{S_0^c}$ denote the subvector of $\boldsymbol{\beta}$ with all groups in S_0^c. Let $\epsilon_n = (s_0 \log G / n)^{1/2}$.

Without loss of generality, assume that the first s_0 groups of $\boldsymbol{\beta}$ are nonzero. Then $\boldsymbol{\beta}_0 = (\boldsymbol{\beta}_{0S}^\top, \mathbf{0})^\top$. 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 $\widehat{\boldsymbol{\beta}}$ of (2.5) 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 (2.5) 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,

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

where $\text{pen}(\boldsymbol{\beta}_g)$ 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 $C > 0$,

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

If (A.9) holds, this will imply that with probability at least $1 - \varepsilon$, there exists a local maximum inside the ball $\{\boldsymbol{\beta}_0 + \epsilon_n \boldsymbol{\Delta} : \|\boldsymbol{\Delta}_{S_0}\|_2 \leq C\}$. In order to establish (A.9), we must show that with probability tending to one and sufficiently large $C > 0$,

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

We bound the terms in (A.10) separately. We first focus on bounding $(\text{I}) := \ell_n(\boldsymbol{\beta}_0 + \epsilon_n \boldsymbol{\Delta}) - \ell_n(\boldsymbol{\beta}_0)$. 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

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

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

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

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

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

where the last inequality of the display uses assumption (A2) that $m_{\max} = O(\log n \wedge (\log G/\log n))$ and assumption (A3) that $\lambda_{\max}\left(n^{-1}\mathbf{X}_{S_0}^\top \boldsymbol{\Sigma}(\boldsymbol{\beta}_0)\mathbf{X}_{S_0}\right) \lesssim \log G/\log n$. 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

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

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

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

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

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

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 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,

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

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

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

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

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

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

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

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 $\mathcal{T}=\{\boldsymbol{\beta}\in\mathbb{R}^p:\boldsymbol{\beta}_{S_0^c}=\mathbf{0}\}$.

Step 2 (sparsity). Since $\log\pi(\boldsymbol{\beta})$ 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 [12] 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 $\widehat{\boldsymbol{\beta}}$ from Step 1 satisfies

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

where $\text{pen}(\boldsymbol{\beta}_g)$ is as in (A.1), and

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

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

We first prove (A.17). We have that

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

Thus,

$$\begin{aligned}
&\lambda_{\max}\left\{\nabla^2\ell_n(\widehat{\boldsymbol{\beta}}_g)+\nabla^2\text{pen}(\widehat{\boldsymbol{\beta}}_g)\right\}\\
&\quad\leq -\lambda_{\min}\left(-\mathbf{X}_g^\top\boldsymbol{\Sigma}(\widehat{\boldsymbol{\beta}})\mathbf{X}_g\right)+\frac{(\lambda_0-\lambda_1)^2}{4}\left[\lambda_{\max}\left(\frac{1}{\|\widehat{\boldsymbol{\beta}}_g\|_2}\mathbf{I}_{m_g}\right)-\lambda_{\min}\left(\frac{\widehat{\boldsymbol{\beta}}_g\widehat{\boldsymbol{\beta}}_g^\top}{\|\widehat{\boldsymbol{\beta}}_g\|_2^3}\right)\right]\\
&\quad\leq -n\underline{\tau}+\frac{(\lambda_0-\lambda_1)^2}{4}\left[\frac{1}{\|\widehat{\boldsymbol{\beta}}_g\|_2}-\frac{1}{\|\widehat{\boldsymbol{\beta}}_g\|_2}\right]\\
&\quad= -n\underline{\tau}<0,
\end{aligned}$$

where we used Weyl's inequality, Assumption (A3), the fact that $\lambda_{\max}(-\mathbf{A}) = -\lambda_{\min}(\mathbf{A})$ for a square matrix $\mathbf{A}$, and the fact that $\widehat{\boldsymbol{\beta}}_g \widehat{\boldsymbol{\beta}}_g^\top$ has the same eigenvalue as $\widehat{\boldsymbol{\beta}}_g^\top \widehat{\boldsymbol{\beta}}_g = \|\widehat{\boldsymbol{\beta}}_g\|_2^2$. Thus, we have shown (A.17).

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

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

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 $\widehat{\boldsymbol{\beta}}_{0S_0^c} = \mathbf{0}$, we have

$$h_g(\widehat{\boldsymbol{\beta}}) = h_g(\boldsymbol{\beta}_0) + \nabla_{S_0} h_g(\widetilde{\boldsymbol{\beta}})^\top (\widehat{\boldsymbol{\beta}}_{S_0} - \boldsymbol{\beta}_{0S_0}), \tag{A.19}$$

where $\widetilde{\boldsymbol{\beta}}$ lies on the line segment between $\boldsymbol{\beta}_0$ and $\widehat{\boldsymbol{\beta}}$. By (A.19), we have that

$$\begin{aligned}
& P\left(\exists\, g \in S_0^c, \ \|h_g(\widehat{\boldsymbol{\beta}})\|_2 \geq \lambda_g^\star(\mathbf{0}_{m_g}) \right) \\
& \leq P\left(\exists\, g \in S_0^c, \ \|h_g(\boldsymbol{\beta}_0)\|_2 \geq \frac{\lambda_g^\star(\mathbf{0}_{m_g})}{2} \right) + P\left(\exists\, g \in S_0^c, \ \|\nabla_{S_0} h_g(\widetilde{\boldsymbol{\beta}})^\top (\widehat{\boldsymbol{\beta}}_{S_0} - \boldsymbol{\beta}_{0S_0})\|_2 \geq \frac{\lambda_g^\star(\mathbf{0}_{m_g})}{2} \right) \\
& \leq \sum_{g=s_0+1}^{G} P\left(\|h_g(\boldsymbol{\beta}_0)\|_2 \geq \frac{\lambda_g^\star(\mathbf{0}_{m_g})}{2} \right) + \sum_{g=s_0+1}^{G} P\left(\|\nabla_{S_0} h_g(\widetilde{\boldsymbol{\beta}})^\top (\widehat{\boldsymbol{\beta}}_{S_0} - \boldsymbol{\beta}_{0S_0})\|_2 \geq \frac{\lambda_g^\star(\mathbf{0}_{m_g})}{2} \right) \\
& \triangleq T_1 + T_2.
\end{aligned} \tag{A.20}$$

We will bound each of the terms in (A.20) 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,

$$\begin{aligned}
\lambda_\theta^\star(\mathbf{0}_{m_g}) &= \lambda_1 p_\theta^\star(\mathbf{0}_{m_g}) + \lambda_0 \left[1 - p_\theta^\star(\mathbf{0}_{m_g})\right] \\
&= \frac{\lambda_1 + \lambda_0 (Mn\lambda_0^{m_g+1})}{1 + Mn\lambda_0^{m_g+1}} > \lambda_0 \left(\frac{Mn\lambda_0^{m_g+1}}{1 + Mn\lambda_0^{m_g+1}} \right) > \frac{\lambda_0}{2},
\end{aligned} \tag{A.21}$$

for large n.

We first examine each of the summands in T_1 in (A.20). Since all of the entries of $\mathbf{X}$ are bounded in absolute value by some $D > 0$ (by Assumption (A3)), we have that for any $g \in \{1, \ldots, G\}$, $\|\mathbf{X}_g\|_{2,\infty} \leq D\sqrt{m_g} \leq D\sqrt{m_{\max}}$. Thus, in light of (A.21), we have that for sufficiently large n,

$$\begin{aligned}
P\left(\|h_g(\boldsymbol{\beta}_0)\|_2 \geq \frac{\lambda_\theta^\star(\mathbf{0}_{m_g})}{2} \right) &< P\left(\|\mathbf{X}_g\|_{2,\infty} \|\mathbf{Y} - b'(\mathbf{X}\boldsymbol{\beta}_0)\|_2 \geq \frac{\lambda_0}{2} \right) \\
&\leq P\left(\|\mathbf{Y} - b'(\mathbf{X}\boldsymbol{\beta}_0)\|_2 \geq \frac{\lambda_0}{2D\sqrt{m_{\max}}} \right) \\
&\leq 2\exp\left(-\frac{\lambda_0^2 / 4D^2 m_{\max}}{2nG + L\lambda_0 / D\sqrt{m_{\max}}} \right) \\
&= 2\exp\left(-\frac{\lambda_0^2}{8D^2 m_{\max} nG + 4L\lambda_0 D\sqrt{m_{\max}}} \right)
\end{aligned} \tag{A.22}$$

where we used Assumption (A4) and the Bernstein inequality (specifically, Lemma 2.2.11 of van der Vaart and Wellner [32]). Thus, from (A.22), we have as an upper bound for T_1 in (A.20),

$$T_1 < 2\exp\left(\log G - \frac{\lambda_0^2}{8D^2 m_{\max} nG + 4L\lambda_0 D\sqrt{m_{\max}}}\right) \to 0, \tag{A.23}$$

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.23) dominates the first term as $n \to \infty$.

We next examine each of the summands in T_2 in (A.20). In light of (A.21), we have that for each $g \in S_0^c$ and sufficiently large n,

$$\begin{aligned}
&P\left(\|\nabla_{S_0} h_g(\widetilde{\boldsymbol{\beta}})^\top(\widehat{\boldsymbol{\beta}}_{S_0} - \boldsymbol{\beta}_{0S_0})\|_2 \geq \frac{\lambda_g^\star(\mathbf{0}_{m_g})}{2}\right) \\
&\quad < P\left(\|\nabla_{S_0} h_g(\widetilde{\boldsymbol{\beta}})^\top(\widehat{\boldsymbol{\beta}}_{S_0} - \boldsymbol{\beta}_{0S_0})\|_2 \geq \frac{\lambda_0}{2}\right) \\
&\quad \leq P\left(\|\nabla_{S_0} h_g(\widetilde{\boldsymbol{\beta}})\|_{2,\infty}\|\widehat{\boldsymbol{\beta}}_{S_0} - \boldsymbol{\beta}_{0S_0}\|_2 \geq \frac{\lambda_0}{2}\right) \\
&\quad \leq P\left(\|\mathbf{X}_{S_0}^\top \boldsymbol{\Sigma}(\widetilde{\boldsymbol{\beta}})\mathbf{X}_g\|_{2,\infty}\|\widehat{\boldsymbol{\beta}}_{S_0} - \boldsymbol{\beta}_{0S_0}\|_2 \geq \frac{\lambda_0}{2}\right) \\
&\quad \lesssim \frac{2n}{\lambda_0},
\end{aligned} \tag{A.24}$$

To arrive at (A.24), we used the fact that $\nabla_{S_0} h_g(\widetilde{\boldsymbol{\beta}}) = \mathbf{X}_{S_0}^\top \boldsymbol{\Sigma}(\widetilde{\boldsymbol{\beta}})\mathbf{X}_g$ and Assumption (A3) that for all $g \in \mathcal{S}_0^c$, $\|\mathbf{X}_{S_0}^\top \boldsymbol{\Sigma}(\widetilde{\boldsymbol{\beta}})\mathbf{X}_g\|_{2,\infty} = O(n)$. From (A.24) and 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.20),

$$T_2 \prec \frac{2Gn}{\lambda_0} \asymp \frac{Gn}{G^c} = \frac{n}{G^{c-1}} \to 0 \text{ as } n \to \infty. \tag{A.25}$$

Combining (A.20), (A.23), and (A.25) allows us to conclude that (A.18) holds, 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 [16]. However, a crucial difference is that we study an *absolutely continuous* spike-and-slab prior in this paper, whereas Jeong and Ghosal [16] 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 [16]. In particular, since SSGL (2.2) 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 (1.1), 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 (1.3). 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 B1 (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 [14], this statement will be proven if we can show that

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

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 [16], the KL divergence and KL variation for the exponential family (1.1) 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

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

Since both h^{-1} and ξ are strictly increasing, this implies that if $\max_{1 \leq i \leq n}\{(h^{-1})'(\mathbf{x}_i^\top \boldsymbol{\beta}_0)\xi'(\mathbf{x}_i^\top \boldsymbol{\beta}_0)(\mathbf{x}_i^\top \boldsymbol{\beta} - \mathbf{x}_i^\top \boldsymbol{\beta}_0)^2\} \leq \epsilon_n^2 \to 0$, then $|\mathbf{x}_i^\top \boldsymbol{\beta} - \mathbf{x}_i^\top \boldsymbol{\beta}_0| \to 0$ for all $1 \leq i \leq n$. Therefore, recalling from (1.7) that $\boldsymbol{\Sigma}(\boldsymbol{\beta}_0) = \text{diag}((h^{-1})'(\mathbf{X}\boldsymbol{\beta}_0)\xi'(\mathbf{X}\boldsymbol{\beta}_0))$, 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 $n^{-1}\|\boldsymbol{\Sigma}^{1/2}(\boldsymbol{\beta}_0)\mathbf{X}(\boldsymbol{\beta} - \boldsymbol{\beta}_0)\|_2^2$ on $\mathcal{B}_n$.

Thus, for sufficiently large n, we have for some constant $C_2 > 0$,

$$\begin{aligned} \Pi(\mathcal{B}_n) &\geq \Pi\left(\frac{1}{n}\|\boldsymbol{\Sigma}^{1/2}(\boldsymbol{\beta}_0)\mathbf{X}(\boldsymbol{\beta} - \boldsymbol{\beta}_0)\|_2^2 \leq C_2^2 \epsilon_n^2\right) \\ &\geq \Pi\left(\lambda_{\max}\left(\frac{1}{n}\mathbf{X}_g^\top \boldsymbol{\Sigma}(\boldsymbol{\beta}_0)\mathbf{X}_g\right)\left(\sum_{g=1}^G \|\boldsymbol{\beta}_g - \boldsymbol{\beta}_{0g}\|_2\right)^2 \leq C_2^2 \epsilon_n^2\right) \\ &\geq \Pi\left(\sum_{g=1}^G \|\boldsymbol{\beta}_g - \boldsymbol{\beta}_{0g}\|_2 \leq \frac{C_2\sqrt{s_0}}{\sqrt{n}}\right) \\ &\geq \Pi\left(\sum_{g \in S_0} \|\boldsymbol{\beta}_g - \boldsymbol{\beta}_{0g}\|_2 \leq \frac{C_2\sqrt{s_0}}{2\sqrt{n}}\right) \Pi\left(\sum_{g \in S_0^c} \|\boldsymbol{\beta}_g - \boldsymbol{\beta}_{0g}\|_2 \leq \frac{C_2\sqrt{s_0}}{2\sqrt{n}}\right) \end{aligned}$$

$$\geq \prod_{g \in S_0} \Pi\left(\|\boldsymbol{\beta}_g - \boldsymbol{\beta}_{0g}\|_2^2 \leq \frac{C_2^2}{4n}\right) \prod_{g \in S_0^c} \Pi\left(\|\boldsymbol{\beta}_{S_0^c}\|_2^2 \leq \frac{C_2^2 s_0}{4n(G-s_0)}\right), \tag{B.4}$$

where we used Assumption (B3) 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. [2], we have that

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

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

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

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

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

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

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

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 $\|\boldsymbol{\beta}_{S_0}\|_2 \leq \sqrt{s_0}\|\boldsymbol{\beta}_0\|_\infty$, 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

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

Similar to (D.24) in the proof of Theorem 2 of Bai et al. [2], 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 [16] but work with the generalized dimensionality $|\nu(\boldsymbol{\beta})|$ (4.1). Let $\mathcal{E}_n$ be the event defined in (B.1), where $\epsilon_n = (s_0 \log G/n)^{1/2}$. Define the set $\mathcal{A}_n = \{\boldsymbol{\beta} : |\nu(\boldsymbol{\beta})| \leq K_1 s_0\}$, where $K_1 \geq C_1 \vee 1$, and C_1 is the constant in Lemma B1. Then we have

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

By Lemma B1, $\mathbb{P}_0(\mathcal{E}_n^c) \to 0$ 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 B1. On the other hand, an upper bound for the expected value of the numerator is

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

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. [2]) 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

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

Combining (B.7)-(B.9) and Lemma B1 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 $|\boldsymbol{\nu}(\boldsymbol{\beta})|$ (4.1) equals s_0, where s_0 is the true number of nonzero groups in $\boldsymbol{\beta}_0$.

Lemma B2. *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 $\pi(\boldsymbol{\beta}_g) \geq \Psi_1(\boldsymbol{\beta}_g \mid \lambda_0)$ for all $\boldsymbol{\beta}_g$. Further, for $\boldsymbol{\beta}_g \sim \Psi_1(\boldsymbol{\beta}_g \mid \lambda_0)$, we have that $\|\boldsymbol{\beta}_g\|_2$ follows a gamma distribution with shape parameter m_g and scale parameter λ_0. Thus,

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

Meanwhile,

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

where we used the tail bound for the gamma density on p. 29 of Boucheron et al. [6] and the inequality $1 + x - \sqrt{1+2x} \geq (x-1)/2$ for $x > 0$ (similarly as in Ning et al. [23]) 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,

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

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 [16]. 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

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

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 $\epsilon > 0$,

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

where $\mathcal{E}_n$ is in the event in (B.1). By Lemma B1 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,

$$\mathcal{A}_n^\star = \left\{\boldsymbol{\beta} \in \mathcal{A}_n : \|\boldsymbol{\beta} - \boldsymbol{\beta}_0\|_2 \leq G^{M_2}\right\}, \tag{B.15}$$

for sufficiently large $M_2 \geq 1$. By Lemma 2 of Ghosal and van der Vaart [14], there exists a test function φ_n such that for any $\boldsymbol{\beta}_1$ with $H_n(\boldsymbol{\beta}_0, \boldsymbol{\beta}_1) > \epsilon$,

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

We want to use Lemma 9 of Ghosal and van der Vaart [14]. 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 $\boldsymbol{\beta}_1, \boldsymbol{\beta}_2 \in \mathcal{A}_n^\star$,

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

It then follows that

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

where we used Assumption (B3) that $\lambda_{\max}(n^{-1}\mathbf{X}_g^\top \boldsymbol{\Sigma}(\boldsymbol{\beta}_1)\mathbf{X}_g) \lesssim \log G$ 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

$$N(\epsilon_n/36, \mathcal{A}_n^\star, H_n) \leq N\left(\frac{C_5 \epsilon_n}{36G}, \mathcal{A}_n^\star, \|\cdot\|_2\right) = N\left(\frac{C_5 \epsilon_n}{36G}, \left\{\boldsymbol{\beta} \in \mathcal{A}_n, \|\boldsymbol{\beta} - \boldsymbol{\beta}_0\|_2 \leq G^{M_2}\right\}, \|\cdot\|_2\right). \tag{B.16}$$

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 (4.2) is an increasing function of m_g. Therefore,

$$\left\{\boldsymbol{\beta} \in \mathcal{A}_n, \|\boldsymbol{\beta} - \boldsymbol{\beta}_0\|_2 \leq G^{M_2}\right\} \subset \left\{\boldsymbol{\beta} \in \mathbb{R}^p : \|\boldsymbol{\beta}_{S_0} - \boldsymbol{\beta}_{0S_0}\|_2 \leq G^{M_2}, \text{ and } \|\boldsymbol{\beta}_g\|_2 \leq \widetilde{\omega} \text{ for all } g \in S_0^c\right\}$$

$$\subset \left\{\|\boldsymbol{\beta}_{S_0} - \boldsymbol{\beta}_{0S_0}\|_2 \leq G^{M_2}\right\} \times [-\widetilde{\omega}, \widetilde{\omega}]^{G-K_1 s_0}.$$

By Lemma S4 of Shen et al. [28], 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. [28], 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. [28] and the fact that we can upper bound the covering number by the packing number, we can further upper bound (B.16) by

$$\binom{G}{K_1 s_0} D\left(\frac{C_5 \epsilon_n}{36G\sqrt{2}}, \left\{\|\boldsymbol{\beta}_S - \boldsymbol{\beta}_{0S}\|_2 \leq G^{M_2}\right\}, \|\cdot\|_2\right) \leq \binom{G}{K_1 s_0} \left(\frac{108\sqrt{2}G^{M_2+1}}{C_5 \epsilon_n}\right)^{K_1 s_0}, \tag{B.17}$$

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

$$\log N(\epsilon_n/36, \mathcal{A}_n^\star, H_n) \leq K_1 s_0 \log G + K_1 s_0 \left[(M_2 + 1)\log G + \log(108\sqrt{2n})\right] \lesssim n\epsilon_n^2, \tag{B.18}$$

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 [14], which implies that for every $\epsilon > \epsilon_n$, there exists a test $\bar{\varphi}_n$ such that

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

for some $C_6 > 0$. By Lemma B2, $\Pi(|\boldsymbol{\nu}(\boldsymbol{\beta})| = s_0) \geq e^{-C_3 n\epsilon_n^2}$ 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

$$\begin{aligned}
\Pi(\mathcal{A}_n \setminus \mathcal{A}_n^\star) &= \Pi(\boldsymbol{\beta} \in \mathbb{R}^p : |\boldsymbol{\nu}(\boldsymbol{\beta})| \leq K_1 s_0, \|\boldsymbol{\beta} - \boldsymbol{\beta}_0\|_2 > G^{M_2}) \\
&\leq \Pi(\|\boldsymbol{\beta} - \boldsymbol{\beta}_0\|_2 > G^{M_2}) \\
&\leq \sum_{g \in S_0} \Pi\left(\|\boldsymbol{\beta}_g\|_2 > G^{M_2} - \|\boldsymbol{\beta}_{0g}\|_2\right) + \sum_{g \in S_0^c} \Pi\left(\|\boldsymbol{\beta}_g\|_2 > G^{M_2}\right).
\end{aligned} \tag{B.20}$$

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 $M_2 \geq 1$, we have

$$\begin{aligned}\sum_{g \in S_0} \Pi\left(\|\boldsymbol{\beta}_g\|_2 > G^{M_2} - \|\boldsymbol{\beta}_{0g}\|_2\right) &\leq \sum_{g \in \mathcal{S}_0} P(X_g > G^{M_2} - \|\boldsymbol{\beta}_{0g}\|_2) \\ &\leq \sum_{g \in \mathcal{S}_0} P(X_g > G^{M_2} - M_3\sqrt{G}\log G) \\ &\leq \sum_{g \in \mathcal{S}_0} \exp\left[-\lambda_1(G^{M_2} - M_3\sqrt{G}\log G) + m_{\max}\right] \\ &= \exp\left[-\lambda_1 G^{M_2} + \lambda_1 M_3\sqrt{G}\log G + m_{\max} + \log s_0\right], \qquad \text{(B.21)}\end{aligned}$$

where in the second line, we used the fact that by Assumption (B4), $\|\boldsymbol{\beta}_{0g}\|_2 \leq \sqrt{G}\|\boldsymbol{\beta}_0\|_\infty \leq M_3\sqrt{G}\log G$, for some constant $M_3 > 0$. 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

$$\begin{aligned}\sum_{g \in S_0^c} \Pi\left(\|\boldsymbol{\beta}_g\|_2 > G^{M_2}\right) &\leq \sum_{g \in S_0^c} P(X_g > G^{M_2}) \\ &\leq \sum_{g \in S_0^c} \exp\left[-\lambda_1 G^{M_2} + m_{\max}\right] \\ &= \exp\left[-\lambda_1 G^{M_2} + m_{\max} + \log(G - s_0)\right]. \qquad \text{(B.22)}\end{aligned}$$

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

$$\begin{aligned}\Pi(\mathcal{A}_n \setminus \mathcal{A}_n^\star) &\leq 2\exp\left[-\lambda_1 G^{M_2} + \lambda_1 M_3\sqrt{G}\log G + \log G + m_{\max}\right] \\ &\lesssim \exp(-C_4 s_0 \log G), \qquad \text{(B.23)}\end{aligned}$$

where we can choose $C_4 > C_3$ so that $-\lambda_1 G^{M_2}$ is the dominating term in the first line of the display, and further, $G^{M_2} \gg C_4 s_0 n \log G$ 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

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

and the smallest scaled singular value ϕ_2 as

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

According to Theorem 3 of Jeong and Ghosal [16], as long as

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

then

$$\{\boldsymbol{\beta} : H_n(\boldsymbol{\beta}, \boldsymbol{\beta}_0) \leq C_4 \epsilon_n\} \subset \left\{ \boldsymbol{\beta} : \|\boldsymbol{\beta} - \boldsymbol{\beta}_0\|_2 \leq \frac{C_7 \epsilon_n}{\phi_2\left((K_2+1)s_0; \boldsymbol{\Sigma}(\boldsymbol{\beta}_0)\right)} \right\}, \tag{B.25}$$

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

$$\begin{aligned} \phi_2(s; \boldsymbol{\Sigma}(\boldsymbol{\beta}_0)) &= \inf_{\boldsymbol{\beta}: 1 \leq |\boldsymbol{\nu}(\boldsymbol{\beta})| \leq s} \frac{\|\boldsymbol{\Sigma}^{1/2}(\boldsymbol{\beta}_0)\mathbf{X}\boldsymbol{\beta}\|_2}{\|\boldsymbol{\beta}\|_2 n^{1/2}} \\ &\geq \frac{[\lambda_{\min}\left(\mathbf{X}_S^\top \boldsymbol{\Sigma}(\boldsymbol{\beta}_0)\mathbf{X}_S\right)]^{1/2}\|\boldsymbol{\beta}\|_2}{\|\boldsymbol{\beta}\|_2 n^{1/2}} \\ &= \left[\lambda_{\min}\left(\frac{1}{n}\mathbf{X}_S^\top \boldsymbol{\Sigma}(\boldsymbol{\beta}_0)\mathbf{X}_S\right)\right]^{1/2} > 0, \end{aligned}$$

by Assumption (B3). We also have $\|\mathbf{X}\|_{\max} \leq D < \infty$ by Assumption (B3), and $\phi_1(s; \boldsymbol{\Sigma}) \geq \phi_2(s; \boldsymbol{\Sigma})$. Therefore,

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

Thus, by Assumption (A2) that $s_0^2 = o(n/\log G)$, 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 $\phi_2((K_2+1)s_0; \boldsymbol{\Sigma}(\boldsymbol{\beta}_0)) > 0$ is a constant that is absorbed into the constant K_3. This proves the theorem. □

C Additional details for the EM algorithm of Section 5.1

C.1 Derivation of the EM Algorithm

Recall that the log-likelihood function $\ell_n(\boldsymbol{\beta})$ is given in (1.4). With the reparameterization of the hierarchical SSGL prior in (5.4) and the $\mathcal{B}(a,b)$ prior (5.1) on θ, we can write the log-posterior $\log \pi(\boldsymbol{\beta}, \theta \mid \mathbf{Y})$ in (5.2) 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 (5.5).

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 (5.6) by taking the derivative of (C.2) with respect to θ. The update for $\boldsymbol{\beta}$ in (5.7) 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 (5.7), 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 (1.4). While (C.3) may appear to be intractable, we can apply the standard iteratively reweighted least squares (IRLS) algorithm [21] 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 (1.1). With the link function h in (1.3), 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 $\mathbf{W} = \text{diag}(w_1^{(k)}, \ldots, w_n^{(k)})$, where the weights are $w_i^{(k)} = [V(\mu_i^{(k)})(h'(\mu_i^{(k)}))^2]^{-1}$. As shown in McCullagh and Nelder [21], 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 [35] 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 [35] 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 [7] 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 [33] with a group lasso penalty.

D 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 responses, the natural parameter is $\theta_i = \log\left(\mu_i/(\mu_i + \alpha)\right)$. With the link function $h(u) = \log(u)$, we have $b(u) = -\alpha \log(1 - e^u)$ in (1.1). 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 (2.2). We compared our results to gLASSO, gSCAD, and gMCP. 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 variables. Then we modeled
>
> $$\log(\theta) = 1.5 \sin(3x_1) - x_5 e^{0.5 x_5^2}.$$
>
> We represented each covariate as a six-term B-spline basis expansion.

We repeated each experiment 200 times. Table D1 reports our results averaged across the 200 replications. In Experiment 1, SSGL, gMCP, and gSCAD all performed equally well in terms of group selection. However, SSGL had the lowest MSE and MSPE, indicating that SSGL performed the best in terms of estimation and prediction.

In Experiment 2, SSGL, gMCP, and gSCAD also performed equally well in terms of group selection. In this case, gSCAD slightly outperformed SSGL and gMCP in terms of prediction with the lowest MSPE, but the results were quite comparable for these three methods. The gLASSO performed the worst in our negative binomial experiments.

Table D1: Simulation results for grouped negative binomial regression under the SSGL, gLASSO, gMCP, and gSCAD models, averaged across 200 replicates. The empirical standard error is reported in parentheses below the average.

Experiment 1

	MSE	MSPE	Sens	Spec	Prec	MCC
SSGL	**0.012**	**31.72**	**1**	**0.998**	**0.993**	**0.996**
	(0.005)	(15.18)	(0)	(0.009)	(0.037)	(0.021)
gLASSO	0.062	36.48	0.998	0.978	0.917	0.944
	(0.018)	(12.19)	(0.020)	(0.033)	(0.114)	(0.076)
gMCP	0.014	32.07	**1**	**0.998**	**0.993**	**0.996**
	(0.006)	(16.41)	(0)	(0.001)	(0.033)	(0.021)
gSCAD	0.015	32.07	**1**	**0.998**	**0.993**	**0.996**
	(0.006)	(16.41)	(0)	(0.008)	(0.033)	(0.021)

Experiment 2

	MSPE	Sens	Spec	Prec	MCC
SSGL	15.67	**1**	**1**	**1**	**1**
	(4.909)	(0)	(0)	(0)	(0)
gLASSO	18.56	1	0.999	0.997	0.998
	(5.869)	(0)	(0.004)	(0.033)	(0.020)
gMCP	15.67	**1**	**1**	**1**	**1**
	(4.638)	(0)	(0)	(0)	(0)
gSCAD	**15.64**	1	0.998	0.983	0.990
	(4.678)	(0)	(0.008)	(0.073)	(0.043)