Annals of the Institute of Statistical Mathematics manuscript No.
(will be inserted by the editor)

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

Ray Bai

Received: June 2023 / Revised: August 2025

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, Poisson, negative binomial, and Gaussian regression with group sparsity. We obtain the minimax ℓ_2 convergence rate for both a maximum *a posteriori* (MAP) estimator *and* the full posterior distribution under our prior. Our theoretical results thus justify the use of the posterior mode as a point estimator. The posterior distribution also 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 expectation-maximization (EM) and Markov chain Monte Carlo (MCMC) algorithms. We illustrate our method through simulations and a real data application on predicting human immunodeficiency virus (HIV) drug resistance from protein sequences.

Keywords generalized linear model · group sparsity · spike-and-slab · spike-and-slab group lasso · posterior contraction · rate of convergence · variable selection

1 Introduction

1.1 Motivation

Generalized linear models (GLMs) (McCullagh and Nelder 1989) 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

Ray Bai
Department of Statistics, George Mason University, Fairfax, VA
E-mail: rbai2@gmu.edu

vector of p covariates for the ith observation. GLMs assume that the response variable y_i belongs to the exponential family with conditional density,

$$f_i(y_i \mid \mathbf{x}_i) = \exp\left\{y_i\theta_i - b(\theta_i) + c(y_i)\right\}. \tag{1.1}$$

where $\theta_i = \theta_i(\mathbf{x}_i) \in \Theta \subset \mathbb{R}$ is the natural parameter depending on $\mathbf{x}_i$, $b(\cdot)$ 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 (McCullagh and Nelder 1989). The conditional mean $\mathbb{E}(y_i \mid \mathbf{x}_i) = b'(\theta_i)$ is related to a linear combination of the covariates $\mathbf{x}_i$ through a differentiable link function h so that $(h \circ b') : \Theta \mapsto \mathbb{R}$ is a strictly increasing function, i.e.

$$(h \circ b')(\theta_i) = \sum_{j=1}^{p} x_{ij}\beta_j, \quad i = 1, \ldots, n, \tag{1.2}$$

where $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_p)^\top \in \mathbb{R}^p$ 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 Poisson regression. However, h can also be chosen to be a *non*-canonical link function. Probit regression and negative binomial 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 (Huang et al. 2012). At the individual gene level, mutations in an amino acid sequence can also be represented with group structure (Rhee et al. 2006). 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 (Lee and Cao 2021; Wen et al. 2019). In semiparametric GLMs, continuous functions of the covariates are often estimated by groups of nonlinear basis functions (Lian 2012). Finally, categorical covariates with multiple levels can be represented as groups of dummy variables for each non-baseline category (Breheny and Huang 2015; Meier et al. 2008).

In these scenarios, it is desirable to take advantage of the known 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 conditional mean $\mathbb{E}(y_i \mid \mathbf{x}_i) = b'(\theta_i)$ 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) (Meier et al. 2008; Lian 2012; Blazère et al. 2014; Breheny and Huang 2015). In the Bayesian literature, spike-and-slab priors (Tang et al. 2017a; Lee and Cao 2021) and automatic relevance determination priors (Wen et al. 2019) 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 computational algorithms 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 (2007), Jeong and Ghosal (2021), and Tang and Martin (2024) have also studied posterior 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 (2007), Jeong and Ghosal (2021), and Tang and Martin (2024). 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 (2007), Jeong and Ghosal (2021), and Tang and Martin (2024) all study *point-mass* (discontinuous) spike-and-slab priors. 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. In contrast, we employ an absolutely continuous prior which needs to be handled differently from these other papers.

Finally, we characterize the convergence rate of *both* a maximum *a posteriori* (MAP) estimator *and* the full posterior distribution. In contrast, Jiang (2007), Jeong and Ghosal (2021), and Tang and Martin (2024) studied *only* the full posterior distribution but *not* any specific Bayesian point estimators. One may wonder why it is necessary to study MAP estimators 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 (2004) 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 (2018) 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. (2015) 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 work, we focus on a local MAP estimator rather than the posterior mean or median, because of its practical appeal. Unlike the posterior mean or median, local MAP estimators under our spike-and-slab prior result in *exact* sparsity and are computationally faster to calculate. Given these appealing features, we aim to elucidate the theoretical optimality of MAP estimation for our model.

In short, our main contributions can be summarized as follows:

1. We study a heavy-tailed, *continuous* spike-and-slab prior for *grouped* GLMs (1.3) with both canonical *and* non-canonical link functions. This class of models is much broader than the grouped linear regression model studied by other authors (Yuan and Lin 2006; Bai et al. 2022; Nardi and Rinaldo 2008), since it encompasses any member of the exponential family (1.1) and allows for many possible link functions besides the identity link.
2. We show that both a local MAP estimator *and* the full posterior distribution under our spike-and-slab prior converge at the minimax-optimal ℓ_2 error 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. We are not aware of any previous theoretical studies of Bayesian *point estimates* under spike-and-slab priors in high-dimensional 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.
3. For efficient point estimation under our model, we introduce an expectation-maximization (EM) algorithm. For fully Bayesian posterior inference, we provide Markov chain Monte Carlo (MCMC) algorithms.

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 convergence rate of a local 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. Most of the proofs are deferred to the Supplementary Material.

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 $\mathbf{x}$, $\mathrm{diag}\{\mathbf{x}\}$ 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 $\mathbf{1}_n$ means an n-dimensional vector of all ones, while $\mathbf{0}_n$ denotes an n-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

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

while the observed information matrix is

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

where

$$\boldsymbol{\Sigma}(\boldsymbol{\beta}) = \boldsymbol{\Omega}(\boldsymbol{\beta}) - \mathrm{diag}\{\xi''(\mathbf{X}\boldsymbol{\beta})\}\mathrm{diag}\{\mathbf{Y} - b'(\xi(\mathbf{X}\boldsymbol{\beta}))\}, \tag{1.7}$$

and

$$\boldsymbol{\Omega}(\boldsymbol{\beta}) = \mathrm{diag}\{(h^{-1})'(\mathbf{X}\boldsymbol{\beta})\}\mathrm{diag}\{\xi'(\mathbf{X}\beta)\}. \tag{1.8}$$

Note that $\mathbf{X}^\top \boldsymbol{\Omega}(\boldsymbol{\beta})\mathbf{X}$ is also the *Fisher* information matrix, or the expected value of the Fisher information matrix (1.6). Furthermore, if the canonical link function is used, then $\boldsymbol{\Sigma}(\boldsymbol{\beta}) = \boldsymbol{\Omega}(\boldsymbol{\beta})$, and (1.7) can be greatly simplified to $\boldsymbol{\Sigma}(\boldsymbol{\beta}) = \text{diag}\{b''(\mathbf{X}\boldsymbol{\beta})\}$.

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 (Fang et al. 1990), *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 parameter 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 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. (2022),

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

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. (2022) 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. (2022) 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.). Using the prior (2.2), we also develop asymptotic theory for *point estimation* in high-dimensional Bayesian GLMs, which to our knowledge, has not been studied before.

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 (2018) 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. (2017b) 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

$$\log \pi(\boldsymbol{\beta} \mid \mathbf{Y}) = \ell_n(\boldsymbol{\beta}) + \log \pi(\boldsymbol{\beta}). \tag{2.5}$$

Hence, a very natural point estimator for $\boldsymbol{\beta}$ is a local MAP estimator $\widehat{\boldsymbol{\beta}}$, i.e.

$$\widehat{\boldsymbol{\beta}} \text{ such that } \nabla \log \pi(\widehat{\boldsymbol{\beta}} \mid \mathbf{Y}) = \mathbf{0}_p. \tag{2.6}$$

Local MAP estimators (or local posterior modes) are useful point estimates to consider because standard optimization algorithms can be used to rapidly obtain an estimate $\widehat{\boldsymbol{\beta}}$. These optimization algorithms are often faster and much more scalable than MCMC algorithms. In general, local MAP estimators may not be unique, and the log-likelihood function $\ell_n(\boldsymbol{\beta})$ may be unbounded. However, if the prior $\pi(\boldsymbol{\beta})$ in (2.5) is proper, does not depend on the observed data, and sufficiently restricts the parameter space for $\boldsymbol{\beta}$ (e.g. by shrinking most of the coefficients in $\boldsymbol{\beta}$ towards zero), then the posterior density $\pi(\boldsymbol{\beta} \mid \mathbf{Y})$ will be proper and bounded, i.e. $\sup_{\boldsymbol{\beta}} \pi(\boldsymbol{\beta} \mid \mathbf{Y}) < \infty$. This enables us to find a local MAP estimator (2.6).

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

$$\log \pi(\boldsymbol{\beta} \mid \mathbf{Y}) = \ell_n(\boldsymbol{\beta}) - \lambda \sum_{g=1}^{G} \|\boldsymbol{\beta}_g\|_2, \tag{2.7}$$

which is the objective function for the group lasso estimator of Yuan and Lin (2006). Thus, the group lasso estimator corresponds to the MAP estimator under independent multivariate Laplace priors (2.1) on the $\boldsymbol{\beta}_g$'s, and the MAP estimator for each group $\boldsymbol{\beta}_g$ 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 local SSGL MAP estimators (2.6) will *also* be exactly sparse, since the mixture components in (2.2) are both multivariate Laplace. However, whereas the group lasso (2.7) 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 local MAP estimators (2.6) under the SSGL prior (2.2) makes the SSGL very appealing for both group selection *and* estimation. In contrast, neither the posterior mean nor the posterior median under the SSGL prior is exactly sparse. Estimating the posterior mean or median also often requires the use of MCMC. In Sections 3.2 and 4.2, we further 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 theoretically study MAP estimation under the SSGL prior (2.2). Our goal is to establish the existence of a local MAP estimator (2.6) with the minimax estimation rate in terms of ℓ_2 risk. To the best of our knowledge, our work is the first one to investigate *point estimation* under a spike-and-slab prior in high-dimensional *Bayesian* GLMs. Other authors (Jiang 2007; Jeong and Ghosal 2021; Tang and Martin 2024) have only studied the convergence rate for the *full* posterior distribution in Bayesian GLMs, which we will also consider in Section 4.

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

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

where h and b are the known link function and cumulant function respectively, while θ_{0i} is the true natural parameter in (1.1). Further, 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 $S_0^c = \{1, \ldots, G\} \setminus S_0$. For a p-dimensional vector $\boldsymbol{\beta}$, let $\boldsymbol{\beta}_{S_0}$ denote the subvector of $\boldsymbol{\beta}$ with the groups in S_0, and let $\boldsymbol{\beta}_{S_0^c}$ be the subvector with groups in S_0^c. Let $\mathbf{X}_{S_0}$ denote the submatrix of the design matrix $\mathbf{X}$ with the

$\sum_{g \in S_0} m_g$ columns of $\mathbf{X}$. Recall that $\boldsymbol{\Sigma}(\boldsymbol{\beta})$ and $\boldsymbol{\Omega}(\boldsymbol{\beta})$ are the matrices defined respectively in (1.7) and (1.8).

Without loss of generality, assume that the first s_0 groups in $\boldsymbol{\beta}_0$ are nonzero, so that $\boldsymbol{\beta}_0 = (\boldsymbol{\beta}_{0S_0}^\top, \mathbf{0}^\top)^\top$. We additionally make the following set of assumptions:

(A1) $G \gg n$ and $\log G = o(n^{1/2})$.
(A2) $s_0 = o(n^{1/2}/\log G)$ and $m_{\max} = O(\log n \wedge (\log G/\log n))$, where $m_{\max} = \max_{1 \le g \le G} m_g$.
(A3) The design matrix $\mathbf{X}$ satisfies the following conditions:
  (i) All the entries x_{ij} of $\mathbf{X}$ satisfy $|x_{ij}| = O(\log G)$.
  (ii) Define the neighborhood $\mathcal{N}_0 = \{\boldsymbol{\delta} \in \mathbb{R}^p : \|\boldsymbol{\delta} - \boldsymbol{\beta}_0\|_2 \le (s_0 \log G/n)^{1/2}\}$. For any $\boldsymbol{\delta} \in \mathcal{N}_0$, $\lambda_{\min}(n^{-1}\mathbf{X}_{S_0}^\top \boldsymbol{\Sigma}(\boldsymbol{\delta})\mathbf{X}_{S_0}) \gtrsim 1$, $\lambda_{\min}(n^{-1}\mathbf{X}_{S_0}^\top \boldsymbol{\Omega}(\boldsymbol{\delta})\mathbf{X}_{S_0}) \gtrsim 1$, and $\lambda_{\max}(n^{-1}\mathbf{X}_{S_0}^\top \boldsymbol{\Omega}(\boldsymbol{\delta})\mathbf{X}_{S_0}) \prec \log G/\log n$.
  (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 $\mathbb{E}(|y_i - b'(\xi(\mathbf{x}_i^\top \boldsymbol{\beta}_0))|^k) \le \frac{k!}{2}\mathbb{E}[y_i^2]L^{k-2}$ for some $L > 0$ and every integer $k \ge 2$. 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 $\boldsymbol{\beta} \in \mathbb{R}^p$ and $i = 1, \ldots, n$.

Assumption (A1) allows the number of groups G to diverge at a nearly exponential rate with n and is analogous to the growth rate for univariate GLMs in Fan and Lv (2011). Assumption (A2) allows the number of nonzero groups s_0 and the group sizes m_g's to diverge but at rates slower than n. Note that Assumption (A2) implies a sparsity condition on $\boldsymbol{\beta}_0$, where the true $\boldsymbol{\beta}_0$ only contains a small number of nonzero groups. Assumption (A3)(i) holds automatically if the entries of $\mathbf{X}$ are uniformly bounded. If the entries of $\mathbf{X}$ are sub-Gaussian with scale factor σ, then this assumption also holds with probability greater than $1 - 2\exp(-Dn/2\sigma^2)$ for some constant $D > 0$.

Assumption (A3)(ii) gives restricted eigenvalue conditions for the submatrix $\mathbf{X}_{S_0}$ of $\mathbf{X}$. Note that the maximum eigenvalue for $n^{-1}\mathbf{X}_{S_0}^\top \boldsymbol{\Omega}(\boldsymbol{\delta})\mathbf{X}_{S_0}$ can diverge as $n \to \infty$. Restricted eigenvalue conditions are routinely employed in the high-dimensional GLM literature (Fan and Lv 2011; Tang and Martin 2024). These conditions ensure the identifiability of $\boldsymbol{\beta}_0$. It should also be noted that since our theory covers both canonical *and* non-canonical link functions $h(\cdot)$, we require eigenvalue conditions for both $n^{-1}\mathbf{X}_{S_0}^\top \boldsymbol{\Sigma}(\boldsymbol{\delta})\mathbf{X}_{S_0}$ and $n^{-1}\mathbf{X}_{S_0}^\top \boldsymbol{\Omega}(\boldsymbol{\delta})\mathbf{X}_{S_0}$. If a non-canonical link function is used, then in general, $\boldsymbol{\Sigma}(\boldsymbol{\beta}) \neq \boldsymbol{\Omega}(\boldsymbol{\beta})$. However, if the canonical link function is used, then $\boldsymbol{\Sigma}(\boldsymbol{\beta}) = \boldsymbol{\Omega}(\boldsymbol{\beta})$, and the first two eigenvalue conditions in (A3)(ii) are the same.

Assumption (A3)(iii) is an irrepresentability-type condition (Zhao and Yu 2006; Fan and Lv 2011), 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. In fact, Assumption (A3)(iii) is much weaker than the strong irrepresentability condition of Zhao and Yu (2006) for the LASSO, which requires that $\|\mathbf{X}_{S_0^c}^\top \mathbf{X}_{S_0}(\mathbf{X}_{S_0}^\top \mathbf{X}_{S_0})^{-1}\|_\infty \leq C < \infty$. In contrast, we can allow $\|\mathbf{X}_{S_0}^\top \boldsymbol{\Sigma}(\boldsymbol{\delta})\mathbf{X}_g\|_{2,\infty}$ to diverge on the order of n. A nearly identical assumption as Assumption (A3)(iii) was made by Fan and Lv (2011). When $G \gg n$, it does not seem as though this type of irrepresentability condition can be removed. In low-dimensional (i.e. fixed G) settings, Assumption (A3)(iii) is not needed (Nardi and Rinaldo 2008). However, if G grows much faster than n, then conditions like (A3)(iii) may be required (Huang et al. 2008; Fan and Lv 2011). If we consider weaker notions of convergence such as convergence in ℓ_∞ norm or if we focus on the problem of support recovery (i.e. the ability to asymptotically identify the correct subset of nonzero groups), then we may also be able to remove the irrepresentability condition (Loh and Wainwright 2017). However, we conjecture that this type of condition is necessary when considering convergence in ℓ_2 norm in the high-dimensional $G \gg n$ regime.

Assumption (A4)(i) 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 (Baraud 2010). In addition, the assumption that $\text{Var}(y_i \mid \mathbf{x}_i) \lesssim G$ is a very weak assumption, in light of Assumption (A1) that allows $G = O(e^{n^\xi})$, for some $\xi \in (0, 1/2)$. Assumption (A4)(ii) 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 $\|\mathbf{X}\boldsymbol{\beta}_0\|_\infty < \infty$.

The minimax-optimal ℓ_2 risk for estimating $\boldsymbol{\beta}_0$ in the grouped regression model (3.1) is $\{[s_0 \log(G/s_0) + \sum_{g \in S_0} m_g]/n\}^{1/2}$ (Huang and Zhang 2010). Under Assumptions (A1)-(A2), the minimax ℓ_2 convergence rate is of the same order as $(s_0 \log G/n)^{1/2}$. Our first theorem below certifies that there exists a local MAP estimator (2.6) under the SSGL prior (2.2) which achieves this minimax-optimal rate with respect to the ℓ_2 risk. This justifies the use of MAP estimation for Bayesian grouped regression with the SSGL prior.

Theorem 1 (convergence rate of a local 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 Assumptions (A1)-(A4) hold. Then there exists a strict local MAP estimator* $\widehat{\boldsymbol{\beta}} = (\widehat{\boldsymbol{\beta}}_{S_0}^\top, \widehat{\boldsymbol{\beta}}_{S_0^c}^\top)^\top$ *of the log-posterior* (2.5) *such that, as* $n \to \infty$*,* $\widehat{\boldsymbol{\beta}}_{S_0^c} = \mathbf{0}$ *and*

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

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 $P((1-\theta)/\theta \geq G^c) \geq 1 - e^{-Ds_0 \log G}$ where $D > 0$ 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 $\theta \sim \text{Beta}(1, G^c)$ (Bai et al. 2022).

Remark 2 By the continuous mapping theorem, $f(\widehat{\boldsymbol{\beta}})$ is also a consistent estimator of $f(\boldsymbol{\beta}_0)$ for any function f that is continuous at $\boldsymbol{\beta}_0$. For example, let $S \subset \{1, \ldots, p\}$, and let $\boldsymbol{\beta}_S$ be the subvector of $\boldsymbol{\beta}$ with indices in S. Taking $f(\boldsymbol{\beta}) = \mathbf{A}_S\boldsymbol{\beta}$, where $\mathbf{A}_S \in \mathbb{R}^{|S| \times p}$ is a matrix with unit vector rows such that $\mathbf{A}_S\boldsymbol{\beta} = \boldsymbol{\beta}_S$, it follows that $\widehat{\boldsymbol{\beta}}_S$ is a consistent estimator of $\boldsymbol{\beta}_S$.

Remark 3 Theorem 1 implies that as $n \to \infty$, $\theta \asymp 1/(G^c+1) \to 0$. Since θ can be interpreted as the prior probability that the group $\boldsymbol{\beta}_g$ belongs to the slab density $\boldsymbol{\Psi}(\cdot \mid \lambda_1)$ rather than the spike density $\boldsymbol{\Psi}(\cdot \mid \lambda_0)$, this implies that the SSGL prior (2.2) will classify a smaller proportion of the groups as nonzero as $n \to \infty$. This matches our assumption that the true proportion of nonzero groups, i.e. s_0/G, also decays to zero as $n \to \infty$. However, the diffuse slab $\boldsymbol{\Psi}(\cdot \mid \lambda_1)$ still enables us to identify the true signals. To see this, notice that the prior probability that $\boldsymbol{\beta}_g$ belongs to the slab can be written as

$$\begin{aligned} p^\star_\theta(\boldsymbol{\beta}_g) &= \frac{\theta\boldsymbol{\Psi}(\boldsymbol{\beta}_g \mid \lambda_1)}{\theta\boldsymbol{\Psi}(\boldsymbol{\beta}_g \mid \lambda_1) + (1-\theta)\boldsymbol{\Psi}(\boldsymbol{\beta}_g \mid \lambda_0)} \\ &= \frac{1}{1 + \left(\frac{1-\theta}{\theta}\right)\left(\frac{\lambda_0}{\lambda_1}\right)^{m_g} \exp\left[-(\lambda_0 - \lambda_1)\|\boldsymbol{\beta}_g\|_2\right]}. \end{aligned} \tag{3.3}$$

We can see from (3.3) that if $\boldsymbol{\beta}_g \neq \mathbf{0}_{m_g}$, then $p^\star_\theta(\boldsymbol{\beta}_g) \approx 1$ as $n \to \infty$. On the other hand, if $\boldsymbol{\beta}_g = \mathbf{0}_{m_g}$, then $p^\star_\theta(\boldsymbol{\beta}_g) \approx 0$ as $n \to \infty$.

We also have the following corollary which gives the convergence rate of a local MAP estimator for $\boldsymbol{\beta}$ under the SSL prior (2.3) of Ročková and George (2018) on the regression coefficients in unstructured GLMs (1.2).

Corollary 1 (convergence rate of a local 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 $p \gg n$, $\log p = o(n^{1/2})$, $s_0 = o(n^{1/2}/\log p)$, and Assumptions (A3)-(A4) hold with G replaced by p. Then there exists a strict local MAP estimator $\widehat{\boldsymbol{\beta}} = (\widehat{\boldsymbol{\beta}}_{S_0}^\top, \widehat{\boldsymbol{\beta}}_{S_0^c}^\top)^\top$ of the log-posterior* (2.5) *such that, as $n \to \infty$, $\widehat{\boldsymbol{\beta}}_{S_0^c} = \mathbf{0}$ and*

$$\|\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0\|_2 = O_p\left(\sqrt{\frac{s_0 \log p}{n}}\right). \tag{3.4}$$

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 using MAP estimation for point estimation 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 (2.7) of Yuan and Lin (2006) has been studied theoretically in GLMs by Blazère et al. (2014). Similar to the SSGL, Blazère et al. (2014) obtained the convergence rate of $O((s_0 \log G/n)^{1/2})$ for the group lasso. However, Blazère et al. (2014) require the assumption that $\sum_{g=1}^{G} \sqrt{m_g}\|\boldsymbol{\beta}_{0g}\|_2 < \infty$ (condition (H.3) in Blazère et al. (2014)) 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 violate Condition (H.3) of Blazère et al. (2014). 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 (2.7). 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

For fully Bayesian inference, the minimax rate is a useful benchmark for studying contraction rates, because the posterior cannot contract faster than the minimax rate (Ghosal et al. 2000). In this section, we analyze the *full* SSGL posterior (2.4) and show that the full posterior $\pi(\boldsymbol{\beta} \mid \mathbf{Y})$ inherits the nice theoretical properties of the SSGL MAP estimator (2.6). As discussed in Section 1.2, it is *not* automatically the case that the posterior contracts around the true parameter $\boldsymbol{\beta}_0$ in (3.1) at the same rate as posterior point estimates (Castillo et al. 2015; Castillo and Mismer 2018). Thus, we require a separate analysis of the full posterior.

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 $\boldsymbol{\Omega}(\boldsymbol{\beta})$ is defined as in (1.8).

(B3) The design matrix $\mathbf{X}$ satisfies the following conditions:
  (i) All the entries x_{ij} of $\mathbf{X}$ satisfy $|x_{ij}| = O(\log G)$.
  (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 $|S| \leq s_0$, we have $\lambda_{\min}\left(n^{-1}\mathbf{X}_S^\top \boldsymbol{\Omega}(\boldsymbol{\beta}_0)\mathbf{X}_S\right) \gtrsim 1$. Meanwhile, for any $g \in \{1, \ldots, G\}$ and any $\boldsymbol{\beta}$ such that $\|\boldsymbol{\beta} - \boldsymbol{\beta}_0\|_2 \leq G^M$ for some $M \geq 1$, $\lambda_{\max}(n^{-1}\mathbf{X}_g^\top \boldsymbol{\Omega}(\boldsymbol{\beta})\mathbf{X}_g) \lesssim \log G$.

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

Assumption (B3)(ii) imposes restricted eigenvalue conditions which are stronger than those in Assumption (A3)(ii). The assumption $\lambda_{\min}(n^{-1}\mathbf{X}_S^\top \boldsymbol{\Omega}(\boldsymbol{\beta}_0)\mathbf{X}_S) \gtrsim 1$ is required to hold for *all* submatrices $\mathbf{X}_S$ where $|S| \leq s_0$. The condition $\lambda_{\max}(n^{-1}\mathbf{X}_g^\top \boldsymbol{\Omega}(\boldsymbol{\beta})\mathbf{X}_g) \lesssim \log G$ also needs to hold for *individual* submatrices $\mathbf{X}_g$ (not necessarily the entire $n \times p$ design matrix $\mathbf{X}$). In contrast, (A3)(ii) 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 groups in $\boldsymbol{\beta}$ have magnitude much larger than zero. As a trade-off, however, the irrepresentability condition in Assumption (A3)(iii) is *not* needed for the full posterior to contract at the minimax-optimal rate.

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 $\boldsymbol{\beta}_0$. Taken together, Assumptions (A2) and (B4) imply restrictions on the true parameter space for $\boldsymbol{\beta}_0$, both in terms of the sparsity of $\boldsymbol{\beta}_0$ and the magnitude of its entries. 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 (2007), Jeong and Ghosal (2021), and Tang and Martin (2024) 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 posterior is continuous and thus 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* (Bhattacharya et al. 2015; Ročková and George 2018; Bai et al. 2022). Following Bai et al. (2022), 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

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

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. (2022), 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 the SSGL prior (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. This also implies that if the threshold ω_g in (4.2) tends to zero as $n \to \infty$ for all $g \in \{1, \ldots, G\}$, then $|\boldsymbol{\nu}(\boldsymbol{\beta})|$ can estimate the number of nonzero groups in $\boldsymbol{\beta}$ as $n \to \infty$.

We first establish 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 approximately sparse sets in high-dimensional GLMs. Theorem 3 then verifies that the SSGL posterior contracts at the minimax-optimal rate around the true parameter $\boldsymbol{\beta}_0$. In the subsequent theorems, the notation $\mathbb{E}_0$ denotes the expectation operator with the true parameter $\boldsymbol{\beta}_0$ in (3.1).

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

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

Theorem 3 (posterior contraction rate under SSGL) *Assume the same setup as Theorem 1, and suppose that Assumptions (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 4 Similarly as with the local MAP estimator 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 $P((1-\theta)/\theta > G^c) > 1 - e^{-D s_0 \log G / n}$, where $D > 0$ 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 $\theta \sim \text{Beta}(1, G^c)$. See Bai et al. (2022) for the proof in the case of Gaussian grouped linear regression, which also holds in the GLM setting.

Remark 5 If $f(\boldsymbol{\beta})$ is a continuous function, then the posterior distribution for $f(\boldsymbol{\beta})$ is also consistent at $f(\boldsymbol{\beta}_0)$. In particular, this implies that the marginal posteriors for the individual coefficients or subsets of the coefficients in $\boldsymbol{\beta}$ also concentrate around their respective true values (see Remark 2).

Remark 6 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., Castillo et al. (2015), Ročková and George (2018), and Atchadé (2017)). 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 (2018), we obtain the following posterior contraction rate.

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

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

Theorem 3 and Corollary 2 show that for high-dimensional GLMs, the posterior distributions under the SSGL (2.2) and SSL (2.3) priors *also* converge at the fastest possible (i.e. the minimax) rates. It is not necessarily the case that the full posterior has the same asymptotic behavior as posterior point estimates. For instance, Castillo and Mismer (2018) and Castillo et al. (2015) provide examples where the full posterior actually contracts much slower than the posterior mean, median, or mode.

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 (Rousseau 2016). In particular, the posterior contraction rate gives an indication as to how large we can expect the posterior credible sets to be (Rousseau 2016). 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 $1 - \alpha, \alpha \in (0, 1)$ (Rousseau 2016). 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 there exists a local MAP estimator (2.6) for SSGL which converges at the minimax-optimal rate under weaker assumptions than those previously assumed for the group lasso (2.7). 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 (2.7) 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), however, the group lasso posterior might contract much *slower* than its posterior mode. This renders the group lasso less useful for uncertainty quantification. An example of this phenomenon is provided in the following proposition. Recall that $\boldsymbol{\beta}_0 = (\boldsymbol{\beta}_{01}^{\top}, \ldots, \boldsymbol{\beta}_{0G}^{\top})^{\top}$, where $\boldsymbol{\beta}_{0g}$ is the gth group of size m_g.

Proposition 1 *Suppose that* $\mathbf{Y} \sim \mathcal{N}_n(\boldsymbol{\beta}_0, \mathbf{I}_n)$ *and all the group sizes are constant and equal, i.e.* $m_1 = \cdots = m_G = m$ *where* $1 \leq m < \infty$. *Further, assume that* $\boldsymbol{\beta}_{01} \neq \mathbf{0}_m$ *with* $\|\boldsymbol{\beta}_{01}\|_2 = L < \infty$, *while* $\boldsymbol{\beta}_{0g} = \mathbf{0}_m$ *for all* $g \in \{2, \ldots, G\}$. *Suppose that we endow* $\boldsymbol{\beta}_0$ *with the prior* $\pi(\boldsymbol{\beta}) = \prod_{g=1}^{G} \boldsymbol{\Psi}(\boldsymbol{\beta}_g \mid \lambda)$, *where* $\boldsymbol{\Psi}(\cdot \mid \lambda)$ *is the group lasso prior in* (2.1). *Then if* $\lambda \asymp \sqrt{2 \log n}$, *the MAP estimator* $\widehat{\boldsymbol{\beta}}$ *satisfies*

$$\mathbb{E}_0 \|\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0\|_2 \lesssim \sqrt{2 \log n} \text{ as } n \to \infty, \tag{4.6}$$

but for some $K_4 > 0$,

$$\mathbb{E}_0 \Pi \left(\boldsymbol{\beta} : \|\boldsymbol{\beta} - \boldsymbol{\beta}_0\|_2 \leq K_4 \sqrt{\frac{n}{\log n}} \;\middle|\; \mathbf{Y} \right) \to 0 \text{ as } n \to \infty. \tag{4.7}$$

Proposition 1 implies that the full posterior distribution under the group lasso prior may asymptotically put all of its mass in an ℓ_2 ball with radius that is *substantially larger* than the convergence rate of the posterior mode. Namely, under the grouped GLM (1.3) where $\mathbf{X} = \mathbf{I}_n$ and the exponential family (1.1) distribution is the normal distribution, the group lasso MAP estimator $\widehat{\boldsymbol{\beta}}$ has an expected ℓ_2 risk of the order $\sqrt{2 \log n}$ when the shrinkage parameter λ is chosen of the order $\sqrt{2 \log n}$. This is the minimax rate for estimating $\boldsymbol{\beta}_0$ (Donoho et al. 1992). However, with this same choice of λ, the posterior $\pi(\boldsymbol{\beta} \mid \mathbf{Y})$ places zero probability on the ball $\{\boldsymbol{\beta} : \|\boldsymbol{\beta} - \boldsymbol{\beta}_0\|_2 \leq \sqrt{n / \log n}\}$ as $n \to \infty$, and $\sqrt{2 \log n} \ll \sqrt{n / \log n}$ when n is large. Intuitively, the discrepancy in the convergence rate between the MAP estimator and the full posterior 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 is 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, unlike the group lasso, the full SSGL posterior (2.4) also contracts at the minimax ℓ_2 convergence rate.

5 Implementation of SSGL for GLMs

5.1 EM algorithm for MAP estimation

We first adopt the penalized likelihood perspective and perform MAP estimation for GLMs under the SSGL model. The MAP estimator (2.6) 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 (2014) 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 4).

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 (Ročková and George 2014) 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 (2014) 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,

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

where $\boldsymbol{\gamma}$ is a binary vector. As shown in Section C of the Supplementary Material, $\mathbb{E}[\gamma_g \mid \mathbf{Y}, \boldsymbol{\beta}, \theta] = p_g^\star(\boldsymbol{\beta}_g, \theta)$, 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 $p_g^{\star(t-1)} := p_g^\star(\boldsymbol{\beta}_g^{(t-1)}, \theta^{(t-1)}) = \mathbb{E}[\gamma_g \mid \mathbf{Y}, \boldsymbol{\beta}^{(t-1)}, \theta^{(t-1)}], g = 1, \ldots, G$. With the hyperprior (5.1) on θ, we then update θ in the M-step as

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

In the M-step, we update $\boldsymbol{\beta}$ 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 Section C of the Supplementary Material, 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.

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. (2017b) recommended tuning λ_0 from cross-validation (CV). For SSGL, we follow Tang et al. (2017b) and similarly tune λ_0 using K-fold CV, with a default of $K = 10$. To accelerate the computational efficiency, we fit the model on each of the K training sets in parallel, allowing for speed-ups on roughly the order of K. In practice, we tune λ_0 from an equispaced grid $\{\lambda_{0,1}, \lambda_{0,2}, \ldots, \lambda_{0,\max}\}$, where $0 < \lambda_{0,1} < \lambda_{0,2} < \cdots < \lambda_{0,\max}$ and $\lambda_{0,\max}$ is the smallest value of λ_0 that results in a MAP estimator of $\widehat{\boldsymbol{\beta}} = \mathbf{0}_p$. It is not hard to see that $\lambda_{0,\max} = \max_{1\leq g\leq G} \|\nabla_g \, \ell_n(\mathbf{0}_p)\|_2$, where ∇_g denotes the subvector of the gradient (1.5) corresponding to the gth group. Typically, $\lambda_{0,\max}$ has an analytical form. For example, in logistic regression, $\lambda_{0,\max} = \max_{1\leq g\leq G} \|0.25\mathbf{X}_g^\top(\mathbf{Y} - 0.5\mathbf{1}_n)\|_2$.

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. (2012), 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.

Remark 7 An inspection of the proofs of Theorems 1-3 reveals that the theoretical results in Theorem 1-3 still hold when we replace the single spike parameter λ_0 with $\lambda_{0g} = \lambda_0\sqrt{m_g}$ for each gth group, as long as we still assume that $\lambda_0 \asymp G^c, c > 2$, as in Theorems 1-3, and $m_{\max} = O(\log n \wedge (\log G/\log n))$, as in Assumption (A2).

5.3 Gibbs sampling for fully Bayesian inference

The EM algorithm in Section 5.1 returns only a single local MAP estimator (2.6). However, *fully* Bayesian inference of the grouped GLM model (1.3) is often desirable, since this allows us to quantify uncertainty for the regression coefficients. For a number of Bayesian GLMs with discrete responses, we can use data augmentation with latent variables (Polson et al. 2013; Albert and Chib 1993) to facilitate Gibbs sampling. In this section, we focus on grouped logistic regression and grouped Poisson regression using Pólya-gamma data augmentation (Polson et al. 2013).

A random variable ω with density function $p(\omega)$ is said to follow a Pólya-gamma (PG) distribution $\mathcal{PG}(a, b)$ with parameters $a > 0$ and $b \in \mathbb{R}$ if

$$\omega = \frac{1}{2\pi^2}\sum_{k=1}^{\infty}\frac{g_k}{\left(k-\frac{1}{2}\right)^2 + \left(\frac{b}{2\pi}\right)^2}, \quad \text{and} \quad g_k \sim \text{Gamma}(a, 1).$$

Polson et al. (2013) established the relation,

$$\frac{(e^\psi)^a}{(1+e^\psi)^b} = 2^{-b}e^{\kappa\psi}\int_0^\infty e^{-\omega\psi^2/2}p(\omega)d\omega, \tag{5.8}$$

where $\kappa = a - 0.5b$ and $\omega \sim \mathcal{PG}(b, 0)$. In logistic regression, the likelihood contribution from the ith observation is

$$\mathcal{L}_i(\boldsymbol{\beta}) = \frac{(\exp(\mathbf{x}_i^\top \boldsymbol{\beta}))^{y_i}}{1 + \exp(\mathbf{x}_i^\top \boldsymbol{\beta})}, \quad y_i \in \{0, 1\}. \tag{5.9}$$

Let $\boldsymbol{\omega} = (\omega_1, \ldots, \omega_n)^\top$. Based on (5.8)-(5.9), the conditional distribution for $\boldsymbol{\beta}$ with prior $\pi(\boldsymbol{\beta})$ in logistic regression is

$$\pi(\boldsymbol{\beta} \mid \boldsymbol{\omega}, \mathbf{Y}) = \pi(\boldsymbol{\beta}) \prod_{i=1}^{n} \mathcal{L}_i(\boldsymbol{\beta}) \propto \pi(\boldsymbol{\beta}) \exp\left\{ -\frac{1}{2} (\mathbf{Z} - \mathbf{X}\boldsymbol{\beta})^\top \boldsymbol{\Omega} (\mathbf{Z} - \mathbf{X}\boldsymbol{\beta}) \right\}, \tag{5.10}$$

where $\mathbf{Z} = (\kappa_1/\omega_1, \ldots, \kappa_n/\omega_n)^\top$, $\boldsymbol{\Omega} = \text{diag}(\omega_1, \ldots, \omega_n)$, and for $i = 1, \ldots, n$, $\kappa_i = y_i - 0.5$ and $\omega_i \sim \mathcal{PG}(1, 0)$.

Meanwhile, for Poisson regression, Li et al. (2018) showed that the likelihood contribution from the ith observation can be approximated as

$$\widetilde{\mathcal{L}}_i(\boldsymbol{\beta}) \approx \frac{\left\{\exp(\mathbf{x}_i^\top \boldsymbol{\beta} - \log M)\right\}^{0.5 y_i}}{\left\{1 + \exp(\mathbf{x}_i^\top \boldsymbol{\beta} - \log M)\right\}^{M}}, \tag{5.11}$$

for a sufficiently large $M > 0$. In practice, we can choose $M = 1 + \max_i y_i$. Based on (5.8) and (5.11), the approximate conditional distribution for $\boldsymbol{\beta}$ with prior $\pi(\boldsymbol{\beta})$ in Poisson regression is

$$\pi(\boldsymbol{\beta} \mid \boldsymbol{\omega}, \mathbf{Y}) \approx \pi(\boldsymbol{\beta}) \prod_{i=1}^{n} \widetilde{\mathcal{L}}_i(\boldsymbol{\beta}) \dot{\propto} \pi(\boldsymbol{\beta}) \exp\left\{ -\frac{1}{2} (\mathbf{Z} - \mathbf{X}\boldsymbol{\beta})^\top \boldsymbol{\Omega} (\mathbf{Z} - \mathbf{X}\boldsymbol{\beta}) \right\}, \tag{5.12}$$

where $\mathbf{Z} = (\kappa_1/\omega_1 + \log M, \ldots, \kappa_n/\omega_n + \log M)^\top$, $\boldsymbol{\Omega} = \text{diag}(\omega_1, \ldots, \omega_n)$, and for $i = 1, \ldots, n$, $\kappa_i = 0.5(y_i - M)$ and $\omega_i \sim \mathcal{PG}(M, 0)$.

The exponential terms in (5.10) and (5.12) are Gaussian kernels. Therefore, if the prior $\pi(\boldsymbol{\beta})$ is a Gaussian or a Gaussian scale-mixture, then we can easily draw samples from $\pi(\boldsymbol{\beta} \mid \boldsymbol{\omega}, \mathbf{Y})$ in our Gibbs sampling algorithm.

Remark 8 The conditional distribution (5.12) is not exact. However, (5.12) greatly simplifies posterior sampling for $\boldsymbol{\beta}$ in Poisson regression, since the updates for $\boldsymbol{\beta}$ can be obtained in closed form. Sampling from the *exact* conditional distribution of $\boldsymbol{\beta}$ in Poisson regression typically requires Metropolis-Hastings, which is not practical when $\boldsymbol{\beta}$ is high-dimensional. This is because we cannot decouple the individual components of $\boldsymbol{\beta}$ in the likelihood function, and therefore, we need to update the entire vector $\boldsymbol{\beta}$ in each MCMC iteration. When $\boldsymbol{\beta}$ is a large vector, there is a very low probability of accepting a proposal draw for $\boldsymbol{\beta}$, regardless of the proposal density used. Therefore, we sample $\boldsymbol{\beta}$ directly from the approximate conditional distribution (5.12). In simulation studies, we found that this approximation was very adequate.

Suppose that $\boldsymbol{\beta}_g \sim \boldsymbol{\Psi}(\boldsymbol{\beta}_g \mid \lambda)$, where $\boldsymbol{\Psi}(\boldsymbol{\beta}_g \mid \lambda)$ is the multivariate Laplace density (2.1). Then $\boldsymbol{\beta}_g$ is the marginal prior of the scale mixture,

$$\boldsymbol{\beta}_g \mid \tau_g \sim \mathcal{N}(\mathbf{0}, \tau_g \boldsymbol{I}_{m_g}), \quad \tau_g \sim \text{Gamma}\left(\frac{m_g+1}{2}, \frac{\lambda^2}{2}\right). \tag{5.13}$$

Let $\boldsymbol{\tau} = (\tau_1, \ldots, \tau_G)^\top$ and $\boldsymbol{\gamma} = (\gamma_1, \ldots, \gamma_G)^\top$, and let Bdiag denote a block-diagonal matrix. Based on (5.13) and the reparameterization (5.4) of the SSGL prior, we can rewrite the SSGL prior (2.2) for $\boldsymbol{\beta}$ as a Gaussian scale-mixture hierarchical model,

$$\begin{aligned}
&\boldsymbol{\beta} \mid \boldsymbol{\tau} \sim \mathcal{N}_p(\mathbf{0}_p, \boldsymbol{D}_{\boldsymbol{\tau}}), && \text{where } \boldsymbol{D}_{\boldsymbol{\tau}} = \text{Bdiag}(\tau_1 \boldsymbol{I}_{m_1}, \ldots, \tau_G \boldsymbol{I}_{m_G}),\\
&\tau_g \mid \gamma_g \sim \text{Gamma}\left(\tfrac{m_g+1}{2}, \tfrac{(\lambda_g^\star)^2}{2}\right), && \text{where } \lambda_g^\star = \gamma_g \lambda_1 + (1-\gamma_g)\lambda_0,\\
&\gamma_g \mid \theta \sim \text{Bernoulli}(\theta), && g = 1, \ldots, G.
\end{aligned} \tag{5.14}$$

Due to the hierarchical representation (5.14) of the SSGL prior, we can exploit conditional conjugacy to obtain the conditional distributions for $\boldsymbol{\beta}$ in (5.10) and (5.12) in closed form. With a Beta hyperprior (5.1) for θ in (5.14), the conditional distributions for θ, $\boldsymbol{\tau}$, $\boldsymbol{\gamma}$, and $\boldsymbol{\omega}$ are also all available in closed form.

Our Gibbs sampling algorithm proceeds by sequentially drawing samples from the conditional distributions of $\boldsymbol{\beta}$, $\boldsymbol{\tau}$, $\boldsymbol{\gamma}$, θ, and $\boldsymbol{\omega}$, holding the other variables fixed at their current values. If $p = \sum_{g=1}^{G} m_g$ is greater than n, then we can also use the fast sampling algorithm of Bhattacharya et al. (2016) to sample $\boldsymbol{\beta}$ in $\mathcal{O}(n^2 p)$ time rather than $\mathcal{O}(p^3)$ time. The complete MCMC algorithms for Bayesian grouped logistic regression and grouped Poisson regression with the SSGL prior (2.2) are provided in Section D of the Supplementary Material.

Remark 9 Although we focused on logistic and Poisson regression in this section, data augmentation with Pólya-gamma latent variables also works for other Bayesian GLMs such as negative binomial regression and multinomial regression (Polson et al. 2013). Data augmentation with other types of latent variables can also be used for other GLMs, e.g. Bayesian probit regression models can be implemented using the approach of Albert and Chib (1993).

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) (Breheny and Huang 2015).

In Section E of the Supplementary Material, 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 (Breheny and Huang 2015). 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 other group-regularized estimators of the form,

$$\underset{\boldsymbol{\beta} \in \mathbb{R}^p}{\arg\max}\, \ell_n(\boldsymbol{\beta}) + \sum_{g=1}^{G} \text{pen}_\lambda(\|\boldsymbol{\beta}_g\|_2), \tag{6.1}$$

where $\ell_n(\boldsymbol{\beta})$ is the grouped GLM log-likelihood function defined in (1.4) and $\text{pen}_\lambda(\|\boldsymbol{\beta}_g\|_2)$ is a penalty function on $\|\boldsymbol{\beta}_g\|_2$ which depends on tuning parameter $\lambda > 0$. Our competitors were the group lasso (GL) (Yuan and Lin 2006), the adaptive group lasso (AdGL) (Wang and Leng 2008), the group minimax concave penalty (GMCP) (Breheny and Huang 2015), and the group smoothly clipped absolute deviation (GSCAD) (Breheny and Huang 2015). We implemented GL, AdGL, GMCP, and GSCAD using the R package `grpreg`, where the regularization parameter $\lambda > 0$ was tuned using K-fold CV with $K = 10$. For each of these methods, the penalty was further scaled by $\sqrt{m_g}$ for each gth group.

AdGL adds group-specific weights w_g to the groups $\boldsymbol{\beta}_g, g = 1, \ldots, G$, to counteract the well-known bias of GL. For AdGL, we chose the weights as

$$w_g = \begin{cases} \|\widetilde{\boldsymbol{\beta}}_g\|_2^{-1}, & \text{if } \|\widetilde{\boldsymbol{\beta}}_g\|_2 > 0, \\ \infty, & \text{if } \|\widetilde{\boldsymbol{\beta}}_g\|_2 = 0, \end{cases}$$

where $\widetilde{\boldsymbol{\beta}}_g$ was the GL estimator of $\boldsymbol{\beta}_g$. For GMCP and GSCAD, there is an additional parameter γ which controls the concavity of the penalty (Breheny and Huang 2015). We used the default settings of $\gamma = 3$ for GMCP and $\gamma = 4$ for GSCAD in `grpreg`. While it may be desirable to further tune γ, performing a two-dimensional grid search for (λ, γ) with CV is very computationally

expensive. In the simulation settings that we considered, we also did not find that these methods were very sensitive to the choice of γ, as long as $\gamma > 1$ for GMCP, $\gamma > 2$ for GSCAD, and γ was not overwhelmingly large for either method. Thus, for GMCP and GSCAD, we opted to fix γ and only tuned the penalty parameter λ.

We computed the following performance metrics: mean squared error (MSE), mean squared prediction error (MSPE), true positive rate (TPR), true negative rate (TNR), and precision (Prec), defined as

$$\text{MSE} = \frac{1}{p}\|\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}\|_2^2, \quad \text{MSPE} = \frac{1}{n_{\text{test}}}\sum_{i=1}^{n}(y_{i,\text{test}} - \widehat{y}_{i,\text{test}})^2,$$
$$\text{TPR} = \frac{\text{TP}}{\text{TP}+\text{FN}}, \quad \text{TNR} = \frac{\text{TN}}{\text{TN}+\text{FP}}, \quad \text{Prec} = \frac{\text{TP}}{\text{TP}+\text{FP}},$$

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). 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 Uniform$(-1, 1)$ random variables. Then we modeled

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

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 $\boldsymbol{\Gamma}$, where the (j, k)th entry of $\boldsymbol{\Gamma}$ 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 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, we do not report MSE in Experiment 3.

Table 1 shows that in all four experiments, SSGL had either the lowest or second lowest average MSE and MSPE. SSGL also had the highest average AUC in all experiments. These results demonstrate that SSGL performed very well for both estimation and prediction. In terms of group selection, SSGL had the highest average TPR, indicating that SSGL had the highest power to detect the truly significant groups. However, this higher TPR came at a loss of TNR and precision, where GMCP and GSCAD performed the best on average.

The advantages of SSGL were especially pronounced in the $G \gg n$ setting (Experiment 4), where the average MSE and MSPE were substantially lower for SSGL. In Experiment 4, GL, AdGL, GMCP, and GSCAD all faced greater difficulty picking up the true signals, which led to lower average TPR and higher estimation and prediction error. However, SSGL also estimated more false positives, leading to lower precision.

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

Experiment 1

	MSE	MSPE	AUC	TPR	TNR	Prec
SSGL	0.350	0.128	**0.901**	**0.999**	0.580	0.254
	(0.044)	(0.037)	(0.066)	(0.014)	(0.017)	(0.068)
GL	**0.348**	0.136	0.885	**0.999**	0.523	0.232
	(0.037)	(0.037)	(0.071)	(0.014)	(0.050)	(0.020)
AdGL	0.356	0.136	0.892	**0.999**	0.572	0.251
	(0.039)	(0.040)	(0.071)	(0.014)	(0.037)	(0.015)
GMCP	0.438	0.139	0.883	0.599	**0.979**	**0.894**
	(0.044)	(0.042)	(0.066)	(0.200)	(0.037)	(0.185)
GSCAD	0.402	**0.126**	0.900	0.799	0.943	0.781
	(0.047)	(0.040)	(0.067)	(0.200)	(0.060)	(0.222)

Experiment 2

	MSE	MSPE	AUC	TPR	TNR	Prec
SSGL	**0.074**	**0.149**	**0.883**	**0.898**	0.885	0.174
	(0.012)	(0.021)	(0.011)	(0.108)	(0.024)	(0.045)
GL	0.077	0.158	0.869	0.706	0.920	0.223
	(0.019)	(0.027)	(0.035)	(0.300)	(0.054)	(0.064)
AdGL	0.077	0.157	0.864	0.70	0.911	0.172
	(0.019)	(0.026)	(0.046)	(0.303)	(0.044)	(0.024)
GMCP	0.082	0.156	0.861	0.596	**0.999**	**0.989**
	(0.201)	(0.008)	(0.016)	(0.201)	(0.001)	(0.074)
GSCAD	0.075	0.153	0.863	0.892	0.964	0.419
	(0.010)	(0.024)	(0.047)	(0.117)	(0.015)	(0.135)

Experiment 3

	MSPE	AUC	TPR	TNR	Prec
SSGL	**0.109**	**0.932**	**1**	0.829	0.126
	(0.015)	(0.017)	(0)	(0.034)	(0.021)
GL	0.112	0.930	**1**	0.816	0.160
	(0.016)	(0.017)	(0)	(0.090)	(0.090)
AdGL	0.110	0.931	**1**	0.816	0.134
	(0.015)	(0.017)	(0)	(0.033)	(0.021)
GMCP	0.157	0.871	**1**	**1**	**1**
	(0.019)	(0.042)	(0)	(0)	(0)
GSCAD	0.150	0.887	**1**	**1**	**1**
	(0.014)	(0.028)	(0)	(0)	(0)

Experiment 4

	MSE	MSPE	AUC	TPR	TNR	Prec
SSGL	**0.007**	**0.180**	**0.809**	**0.840**	0.978	0.138
	(0.001)	(0.022)	(0.032)	(0.167)	(0.010)	(0.032)
GL	0.010	0.192	0.789	0.833	0.972	0.101
	(0.001)	(0.016)	(0.044)	(0.177)	(0.004)	(0.031)
AdGL	0.010	0.191	0.792	0.825	0.979	0.134
	(0.001)	(0.016)	(0.046)	(0.189)	(0.006)	(0.065)
GMCP	0.009	0.183	0.800	0.669	**0.994**	**0.306**
	(0.001)	(0.016)	(0.043)	(0.078)	(0.002)	(0.102)
GSCAD	0.009	0.182	0.802	0.827	0.988	0.207
	(0.001)	(0.014)	(0.056)	(0.186)	(0.002)	(0.035)

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 we decreased the magnitude of the entries in the design matrix $\mathbf{X}$ and the signal sizes in $\boldsymbol{\beta}$ in order to ensure realistic count values. Our four simulations were as follows:

Experiment 5 ($G < n$). 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 $\{-1, -0.75, 0.75, 1\}$. Then we modeled

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

Experiment 6 ($G > n$). We repeated Experiment 1 with $n = 100$, but we increased the number of groups to $G = 200$.

Experiment 7 (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 8 (genetic association study with $G \gg n$). 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 we explained in Section 6.2, we did not report the MSE in Experiment 3.

Our findings for grouped Poisson regression were largely consistent with those for grouped logistic regression. Namely, Table 2 shows that SSGL had the lowest average MSE and MSPE in all experiments. Thus, SSGL gave superior performance in terms of estimation and prediction. SSGL also had the highest average TPR, indicating the highest power to detect the truly nonzero groups. However, just as in grouped logistic regression, the higher average TPR came at a cost of lower average TNR and precision.

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

Experiment 5

	MSE	MSPE	TPR	TNR	Prec
SSGL	**0.043**	**4.864**	**1**	0.914	0.626
	(0.001)	(0.801)	(0)	(0.007)	(0.028)
GL	0.046	6.343	**1**	0.713	0.333
	(0.001)	(0.452)	(0)	(0.017)	(0.009)
AdGL	0.044	5.922	**1**	0.656	0.294
	(0.001)	(0.484)	(0)	(0.012)	(0.006)
GMCP	0.051	8.216	0.799	**0.971**	0.799
	(0.001)	(0.677)	(0.032)	(0.014)	(0.033)
GSCAD	0.051	8.186	0.799	**0.971**	**0.800**
	(0.001)	(0.254)	(0.032)	(0.008)	(0.026)

Experiment 6

	MSE	MSPE	TPR	TNR	Prec
SSGL	**0.008**	**55.32**	**0.998**	0.850	0.150
	(0.001)	(27.30)	(0.020)	(0.025)	(0.041)
GL	0.009	56.30	0.995	0.840	0.144
	(0.001)	(26.530)	(0.051)	(0.033)	(0.033)
AdGL	0.009	55.90	0.997	0.852	0.149
	(0.001)	(27.06)	(0.032)	(0.013)	(0.016)
GMCP	0.044	59.68	0.401	**0.966**	**0.395**
	(0.017)	(25.42)	(0.037)	(0.020)	(0.360)
GSCAD	0.037	60.31	0.402	0.964	0.308
	(0.010)	(25.25)	(0.035)	(0.018)	(0.215)

Experiment 7

	MSPE	TPR	TNR	Prec
SSGL	**16.43**	**1**	0.935	0.284
	(0.453)	(0)	(0.013)	(0.017)
GL	16.58	**1**	0.783	0.106
	(0.228)	(0)	(0.005)	(0.003)
AdGL	16.50	**1**	0.820	0.125
	(0.465)	(0)	(0.004)	(0.002)
GMCP	18.72	**1**	**1**	**1**
	(0.166)	(0)	(0)	(0)
GSCAD	18.72	**1**	0.999	0.998
	(0.185)	(0)	(0.001)	(0.024)

Experiment 8

	MSE	MSPE	TPR	TNR	Prec
SSGL	**0.001**	**7.98**	**1**	**0.999**	**0.988**
	(0)	(0.639)	(0)	(0.001)	(0.088)
GL	0.003	28.34	**1**	0.955	0.077
	(0.001)	(3.69)	(0)	(0.001)	(0.003)
AdGL	0.003	26.60	**1**	0.955	0.078
	(0)	(3.89)	(0)	(0.002)	(0.006)
GMCP	0.002	20.22	0.992	**0.999**	**0.988**
	(0.001)	(2.57)	(0.085)	(0.001)	(0.125)
GSCAD	0.002	19.92	0.992	**0.999**	0.987
	(0.001)	(2.13)	(0.085)	(0.001)	(0.105)

Once again, the SSGL also demonstrated its greatest advantage over the competing methods in the $G \gg n$ setting (Experiment 4), where the average MSE and MSPE were substantially lower for SSGL. In this $G \gg n$ setting, SSGL also had the highest average TPR, TNR, and precision. This suggests that SSGL is especially well-suited for estimation and group selection in count datasets where G is much larger than n.

6.4 Fully Bayesian inference

The simulation studies in Sections 6.2 and 6.3 demonstrated that SSGL often outperforms GL, AdGL, GMCP, and GSCAD in terms of estimation, prediction, and discovery of true nonzero groups, especially when the number of predictors p is larger than sample size n. However, these competing methods do not naturally provide uncertainty quantification of the model parameters. Thus, an additional advantage of SSGL is its ability to quantify uncertainty through its posterior distribution.

In this section, we assess the performance of Gibbs sampling algorithm for SSGL introduced in Section 5.3. This Gibbs sampler is implemented in the R package `SSGL`. We conducted the following experiments:

Experiment 9 (grouped logistic regression). With $n = 400$ and $G = 250$ groups of size four (i.e. $p = 1000$), we simulated $\mathbf{X}$ as in Experiment 1 of Section 6.2. We randomly chose $s_0 = 10$ of the vectors $\boldsymbol{\beta}_g$ to be nonzero with $\boldsymbol{\beta}_g = (-1, -0.8, 0.8, 1)^\top$, and the remaining groups were set to zero. Then we simulated binary responses $y_i \mid \mathbf{x}_i \sim \text{Bernoulli}(1/(1 + \exp(-\theta_i))), i = 1, \ldots, n$, where $\log(\theta_i/(1-\theta_i)) = \mathbf{x}_i^\top \boldsymbol{\beta}$.

Experiment 10 (grouped Poisson regression). With $n = 300$ and $G = 200$ groups of size four (i.e. $p = 800$), we simulated $\mathbf{X}$ as in Experiment 5 of Section 6.2. We randomly chose $s_0 = 10$ of the vectors $\boldsymbol{\beta}_g$ to be nonzero with $\boldsymbol{\beta}_g = (-0.9, -0.7, 0.7, 0.9)^\top$, and the remaining groups were set to zero. Then we simulated count responses $y_i \mid \mathbf{x}_i \sim \text{Poisson}(\exp(\theta_i)), i = 1, \ldots, n$, where $\log(\theta_i) = \mathbf{x}_i^\top \boldsymbol{\beta}$.

After generating the data, we used the Gibbs sampling algorithm of Section 5.3 to draw posterior samples. We used the same hyperparameters as those suggested in Section 5.2. In particular, the spike parameter λ_0 was first tuned for the SSGL MAP estimator using the EM algorithm and then fixed at this value for the MCMC algorithm. We ran the Gibbs sampler for a total of 3000 iterations, discarding the first 1000 samples as burnin. The remaining 2000 samples were used to approximate the marginal posterior distributions for the regression coefficients. To assess the quality of the SSGL posterior approximation, we recorded the following metrics:

1. the coverage probability (CP) of the 95% posterior credible intervals (CIs) for the 40 true non-null coefficients in $\boldsymbol{\beta}$, i.e. the proportion of CIs for the nonzero regression coefficients $\{\beta_j : \beta_j \neq 0\}$ that contained the true β_j.

Table 3: Simulation results for the SSGL Gibbs sampling algorithm, averaged across the 200 replicates. The empirical standard error is reported in parentheses.

	Experiment 9	**Experiment 10**
CP	0.960 (0.032)	0.937 (0.039)
Width	2.44 (0.238)	1.81 (0.048)
ESS	1523.29 (13.91)	1255.78 (53.92)
MCSE($q_{0.025}$)	0.034 (0.007)	0.033 (0.004)
MCSE($q_{0.975}$)	0.030 (0.008)	0.033 (0.008)

These intervals were constructed using the 0.025 and 0.975 sample quantiles $[q_{0.025}(\beta_j), q_{0.975}(\beta_j)]$ of the saved 2000 MCMC samples for each β_j.

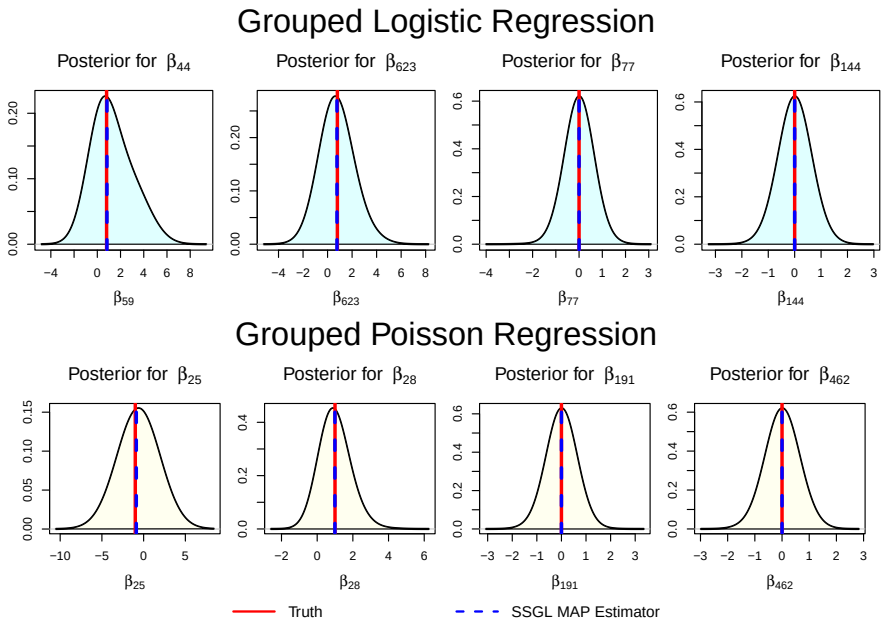
2. the average width of the 95% CIs for the 40 non-null regression coefficients;
3. the average effective sample size (ESS) of the 2000 saved MCMC samples for the β_j's, $j = 1, \ldots, p$;
4. the maximum Monte Carlo standard error (MCSE) of $q_{0.025}(\beta_j)$ and $q_{0.975}(\beta_j)$, $j = 1, \ldots, p$. We denote these quantities by MCSE($q_{0.025}$) and MCSE($q_{0.975}$) respectively.

We repeated Experiments 9 and 10 for 200 replications. Table 3 reports the CP, Width, ESS, MCSE($q_{0.025}$) and MCSE($q_{0.975}$) averaged across these 200 replicates. We see that in both experiments, the 95% CIs covered the true non-null regression coefficients in $\boldsymbol{\beta}$ at close to the nominal rate. Table 3 also shows that the average widths of the 95% CIs for the true nonzero coefficients were not overwhelmingly conservative. Thus, this coverage was not achieved by the CIs being too wide. Our results verify the usefulness of our SSGL Gibbs sampling algorithm for quantifying uncertainty of $\boldsymbol{\beta}$.

Table 3 shows that in Experiment 9 (grouped logistic regression), the average ESS for the 2000 saved MCMC samples was 1523.29, suggesting high efficiency of our Gibbs sampling algorithm. In Experiment 10 (grouped Poisson regression), the average ESS for the 2000 saved MCMC samples was slightly lower (1255.78), and the CIs had average coverage slightly below the nominal level. This suggests that there was some loss of efficiency from using the approximation (5.11) for the Poisson log-likelihood in our Gibbs sampler. However, the CP and ESS were still acceptable in Experiment 10. Finally, in both experiments, MCSE($q_{0.025}$) and MCSE($q_{0.975}$) indicated an acceptable level of precision for the 0.025 and 0.975 MCMC sample quantiles.

Figure 1 plots the kernel-smoothed marginal posterior densities for four regression coefficients from one replication of Experiment 9 (top panel) and four regression coefficients from one replication of Experiment 10 (bottom panel). The true parameter values are plotted as solid red vertical lines, and the SSGL MAP estimates are plotted as dashed vertical lines. Figure 1 shows that in both grouped logistic regression and grouped Poisson regression, the SSGL posterior was able to capture the ground truth. Moreover, the SSGL MAP estimates were close to their true values for the true nonzero coefficients,

Fig. 1: Estimated marginal posterior densities for four regression coefficients from one replication of Experiment 9 (top panel) and four regression coefficients from one replication of Experiment 10 (bottom panel). The true parameter values are plotted as solid red vertical lines, and the SSGL MAP estimators are plotted as dashed blue lines. The left four plots depict the results for four nonzero coefficients, while the right four plots depict the results for four null coefficients where the SSGL MAP estimate is also exactly zero.



and they were *exactly* equal to zero for the true null coefficients. This is due to the exact sparsity of the SSGL MAP estimate (2.6).

As discussed, the SSGL posterior mean and median are *not* exactly sparse. Therefore, we generally recommend using the EM algorithm of Section 5.1 to find a sparse local mode which can be used for group selection and estimation. If uncertainty quantification for $\boldsymbol{\beta}$ is also desired, then the Gibbs sampling algorithm of Section 5.3 can be used for fully Bayesian inference.

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. (2006), these isolates were used to predict resistance to 16 antiretroviral drugs used in

Table 4: Results for SSGL, GL, AdGL, 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	54	**0.087** (0.004)	**0.951** (0.004)
GL	54	0.088 (0.002)	0.945 (0.003)
AdGL	58	0.089 (0.003)	0.944 (0.003)
GMCP	5	0.102 (0.003)	0.925 (0.004)
GSCAD	11	0.100 (0.025)	0.936 (0.003)

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. (2006) 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 the group $g = 13$ would be a "1" instead of a "0."

In Rhee et al. (2006), 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 (Rhee et al. 2006). 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 GL, AdGL, GMCP, and GSCAD. The hyperparameters, tuning parameters, and weights were all chosen the same way as they were in Section 6.

To perform group selection, we fit the five 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.

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

Our results are shown in Table 4. SSGL selected 54 positions. In contrast to SSGL, GL, and AdGL, the methods GMCP and GSCAD selected much more parsimonious models. However, the average out-of-sample MSPE was higher and the average AUC was lower for GMCP and GSCAD than for SSGL, GL, or AdGL.

Despite the fact that the SSGL selected a less sparse model, SSGL still achieved 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 group selection and predictive accuracy.

Supplementary data

The Online Supplementary Material contains proofs of the theoretical results in Sections 3 and 4, additional details for the EM algorithm and the Gibbs sampling algorithms introduced in Section 5, and additional simulation studies for negative binomial regression with a log link.

An R package `SSGL` to implement the methodology in this paper is publicly available on the Comprehensive R Archive Network (CRAN).

Acknowledgements The author is grateful to the anonymous reviewers whose thoughtful feedback helped to greatly improve this paper. The author also thanks Dr. Seonghyun Jeong for helpful discussions which led to derivations of the theoretical results in this paper.

References

Albert, J. H. and Chib, S. (1993). Bayesian analysis of binary and polychotomous response data. *Journal of the American Statistical Association*, 88(422):669–679.

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

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.

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.

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

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.

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.

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

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

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.

Donoho, D. L., Johnstone, I. M., Hoch, J. C., and Stern, A. S. (1992). Maximum entropy and the nearly black object. *Journal of the Royal Statistical Society. Series B (Methodological)*, 54(1):41–81.

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

Fang, K.-T., Kotz, S., and Ng, K. W. (1990). *Symmetric Multivariate and Related Distributions*. Chapman and Hall.

Ghosal, S., Ghosh, J. K., and van der Vaart, A. W. (2000). Convergence rates of posterior distributions. *The Annals of Statistics*, 28(2):500–531.

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

Huang, J., Ma, S., and Zhang, C.-H. (2008). Adaptive lasso for sparse high-dimensional regression models. *Statistica Sinica*, 18(4):1603–1618.

Huang, J. and Zhang, T. (2010). The benefit of group sparsity. *The Annals of Statistics*, 38(4):1978 – 2004.

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

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.

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.

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

Li, Z., Chang, C., Kundu, S., and Long, Q. (2018). Bayesian generalized biclustering analysis via adaptive structured shrinkage. *Biostatistics*, 21(3):610–624.

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

Loh, P.-L. and Wainwright, M. J. (2017). Support recovery without incoherence: A case for nonconvex regularization. *The Annals of Statistics*, 45(6):2455 – 2482.

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

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.

Nardi, Y. and Rinaldo, A. (2008). On the asymptotic properties of the group lasso estimator for linear models. *Electronic Journal of Statistics*, 2:605 – 633.

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

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.

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

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.

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

Tang, Y. and Martin, R. (2024). Empirical Bayes inference in sparse high-dimensional generalized linear models. *Electronic Journal of Statistics*, 2(18):3212–3246.

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.

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.

Wang, H. and Leng, C. (2008). A note on adaptive group lasso. *Computational Statistics & Data Analysis*, 52(12):5277–5286.

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.

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.

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