Spike-and-Slab Group Lassos for Grouped Regression and Sparse Generalized Additive Models

Ray Bai [*] [†], Gemma E. Moran [‡§], Joseph L. Antonelli [¶‖], Yong Chen[**], and Mary R. Boland [**]

July 31, 2020

Abstract

We introduce the spike-and-slab group lasso (SSGL) for Bayesian estimation and variable selection in linear regression with grouped variables. We further extend the SSGL to sparse generalized additive models (GAMs), thereby introducing the first nonparametric variant of the spike-and-slab lasso methodology. Our model simultaneously performs group selection and estimation, while our fully Bayes treatment of the mixture proportion allows for model complexity control and automatic self-adaptivity to different levels of sparsity. We develop theory to uniquely characterize the global posterior mode under the SSGL and introduce a highly efficient block coordinate ascent algorithm for maximum a posteriori (MAP) estimation. We further employ de-biasing methods to provide uncertainty quantification of our estimates. Thus, implementation of our model avoids the computational intensiveness of Markov chain Monte Carlo (MCMC) in high dimensions. We derive posterior concentration rates for both grouped linear regression and sparse GAMs when the number of covariates grows at nearly exponential rate with sample size. Finally, we illustrate our methodology through extensive simulations and data analysis.

Keywords: high-dimensional regression; interaction detection; maximum a posteriori estimation; nonparametric regression; spike-and-slab lasso; variable selection

[*]Department of Statistics, University of South Carolina, Columbia, SC 29208.

[†]Co-first author. Email: RBAI@mailbox.sc.edu

[‡]Data Science Institute, Columbia University, New York, NY 10027.

[§]Co-first author. Email: gm2918@columbia.edu

[¶]Department of Statistics, University of Florida, Gainesville, FL 32611.

[‖]Co-first author. Email: jantonelli@ufl.edu

[**]Department of Biostatistics, Epidemiology, and Informatics, University of Pennsylvania, Philadelphia, PA 19104.

1 Introduction

1.1 Regression with Grouped Variables

Group structure arises in many statistical applications. For example, in multifactor analysis of variance, multi-level categorical predictors are each represented by a group of dummy variables. In genomics, genes within the same pathway may form a group at the pathway or gene set level and act in tandem to regulate a biological system. In each of these scenarios, the response $\boldsymbol{Y}_{n\times 1}$ can be modeled as a linear regression problem with G groups:

$$\boldsymbol{Y} = \sum_{g=1}^{G} \boldsymbol{X}_g \boldsymbol{\beta}_g + \boldsymbol{\varepsilon}, \tag{1.1}$$

where $\boldsymbol{\varepsilon} \sim \mathcal{N}_n(\boldsymbol{0}, \sigma^2 \boldsymbol{I}_n)$, $\boldsymbol{\beta}_g$ is a coefficients vector of length m_g, and $\boldsymbol{X}_g$ is an $n \times m_g$ covariate matrix corresponding to group $g = 1, \dots G$. Even in the absence of grouping information about the covariates, the model (1.1) subsumes a wide class of important nonparametric regression models called *generalized additive models* (GAMs). In GAMs, continuous covariates may be represented by groups of basis functions which have a nonlinear relationship with the response. We defer further discussion of GAMs to Section 5.

It is often of practical interest to select groups of variables that are most significantly associated with the response. To facilitate this group-level selection, Yuan and Lin (2006) introduced the group lasso, which solves the optimization problem,

$$\arg\min_{\boldsymbol{\beta}} \frac{1}{2} \|\boldsymbol{Y} - \sum_{g=1}^{G} \boldsymbol{X}_g \boldsymbol{\beta}_g\|_2^2 + \lambda \sum_{g=1}^{G} \sqrt{m_g} \|\boldsymbol{\beta}_g\|_2, \tag{1.2}$$

where $||\cdot||_2$ is the ℓ_2 norm. In the frequentist literature, many variants of model (1.2) have been introduced, which use some combination of ℓ_1 and ℓ_2 penalties on the coefficients of interest (e.g., Jacob et al. (2009), Li et al. (2015), Simon et al. (2013)).

In the Bayesian framework, selection of relevant groups under model (1.1) is often done by placing spike-and-slab priors on each of the groups $\boldsymbol{\beta}_g$ (e.g., Xu and Ghosh (2015), Liquet et al. (2017), Yang and Narisetty (2019), Ning et al. (2019)). These priors typically take the form,

$$\begin{aligned} \pi(\boldsymbol{\beta}|\boldsymbol{\gamma}) &= \prod_{g=1}^{G}[(1-\gamma_g)\delta_0(\boldsymbol{\beta}_g) + \gamma_g \pi(\boldsymbol{\beta}_g)], \\ \pi(\boldsymbol{\gamma}|\theta) &= \prod_{g=1}^{G} \theta^{\gamma_g}(1-\theta)^{1-\gamma_g}, \\ \theta &\sim \pi(\theta), \end{aligned} \tag{1.3}$$

where $\boldsymbol{\gamma}$ is a binary vector that indexes the 2^G possible models, $\theta \in (0,1)$ is the mixing proportion, δ_0 is a point mass at $\mathbf{0}_{m_g} \in \mathbb{R}^{m_g}$ (the "spike"), and $\pi(\boldsymbol{\beta}_g)$ is an appropriate "slab" density (typically a multivariate normal distribution or a scale-mixture multivariate normal density). With a well-chosen prior on θ, this model will favor parsimonious models in very high dimensions, thus avoiding the curse of dimensionality.

1.2 The Spike-and-Slab Lasso

For Bayesian variable selection, point mass spike-and-slab priors (1.3) are interpretable, but they are computationally intractable in high dimensions, due in large part to the combinatorial complexity of updating the discrete indicators $\boldsymbol{\gamma}$. As an alternative, fully continuous variants of spike-and-slab models have been developed. For continuous spike-and-slab models, the point mass spike δ_0 is replaced by a continuous density heavily concentrated around $\mathbf{0}_{m_g}$. This not only mimics the point mass but it *also* facilitates more efficient computation, as we describe later.

In the context of sparse normal means estimation and univariate linear regression,

Ročková (2018) and Ročková and George (2018) introduced the univariate spike-and-slab lasso (SSL). The SSL places a mixture prior of two Laplace densities on the individual coordinates β_j, i.e.

$$\pi(\boldsymbol{\beta}|\theta) = \prod_{j=1}^{p}[(1-\theta)\psi(\beta_j|\lambda_0) + \theta\psi(\beta_j|\lambda_1)], \tag{1.4}$$

where $\theta \in (0,1)$ is the mixing proportion and $\psi(\cdot|\lambda)$ denotes a univariate Laplace density indexed by hyperparameter λ, i.e. $\psi(\beta|\lambda) = \frac{\lambda}{2}e^{-\lambda|\beta|}$. Typically, we set $\lambda_0 \gg \lambda_1$ so that the spike is heavily concentrated about zero. Unlike (1.3), the SSL model (1.4) does not place any mass on exactly sparse vectors. Nevertheless, the global posterior mode under the SSL prior may be exactly sparse. Meanwhile, the slab stabilizes posterior estimates of the larger coefficients so they are not downward biased. Thus, the SSL posterior mode can be used to perform variable selection and estimation simultaneously.

The spike-and-slab lasso methodology has now been adopted for a wide number of statistical problems. Apart from univariate linear regression, it has been used for factor analysis (Ročková and George (2016), Moran et al. (2019)), multivariate regression (Deshpande et al. (2019)), covariance/precision matrix estimation (Deshpande et al. (2019), Gan et al. (2019), Li et al. (2019)), causal inference (Antonelli et al. (2019)), generalized linear models (GLMs) (Tang et al. (2017b), Tang et al. (2018)), and Cox proportional hazards models (Tang et al. (2017a)).

While the SSL (1.4) induces sparsity on individual coefficients (through the posterior mode), it does not account for group structure of covariates. For inference with structured data in GLMs, Tang et al. (2018) utilized the univariate spike-and-slab lasso prior (1.4) for grouped data where each group had a group-specific sparsity-inducing parameter, θ_g, instead of a single θ for all coefficients. However, this univariate SSL prior does not feature the "all in, all out" selection property of the original group lasso of Yuan and Lin (2006)

or the *grouped* and *multivariate* SSL prior, which we develop in this work.

In this paper, we introduce the *spike-and-slab group lasso* (SSGL) for Bayesian grouped regression and variable selection. Under the SSGL prior, the global posterior mode is exactly sparse, thereby allowing the mode to automatically threshold out insignificant groups of coefficients. To widen the use of spike-and-slab lasso methodology for situations where the linear model is too inflexible, we extend the SSGL to sparse generalized additive models by introducing the *nonparametric spike-and-slab lasso* (NPSSL). To our knowledge, our work is the first to apply the spike-and-slab lasso methodology outside of a parametric setting. Our contributions can be summarized as follows:

1. We propose a new group spike-and-slab prior for estimation and variable selection in both parametric and nonparametric settings. Unlike frequentist methods which rely on separable penalties, our model has a *non*-separable and self-adaptive penalty which allows us to automatically adapt to ensemble information about sparsity.

2. We introduce a highly efficient block coordinate ascent algorithm for global posterior mode estimation. This allows us to rapidly identify significant groups of coefficients, while thresholding out insignificant ones.

3. We show that de-biasing techniques that have been used for the original lasso (Tibshirani, 1996) can be extended to our SSGL model to provide valid inference on the estimated regression coefficients.

4. For both grouped regression and sparse additive models, we derive near-optimal posterior contraction rates for both the regression coefficients $\boldsymbol{\beta}$ *and* the unknown variance σ^2 under the SSGL prior.

The rest of the paper is structured as follows. In Section 2, we introduce the spike-and-slab group lasso (SSGL). In Section 3, we characterize the global posterior mode and introduce efficient algorithms for fast maximum *a posteriori* (MAP) estimation and variable selection. In Section 4, we utilize ideas from the de-biased lasso to perform inference on the SSGL model. In Section 5, we extend the SSGL to nonparametric settings by proposing the nonparametric spike-and-slab lasso (NPSSL). In Section 6, we present asymptotic theory for the SSGL and the NPSSL. Finally, in Sections 7 and 8, we provide extensive simulation studies and use our models to analyze real data sets.

1.3 Notation

We use the following notations. For two nonnegative sequences $\{a_n\}$ and $\{b_n\}$, we write $a_n \asymp b_n$ to denote $0 < \liminf_{n\to\infty} a_n/b_n \leq \limsup_{n\to\infty} a_n/b_n < \infty$. If $\lim_{n\to\infty} a_n/b_n = 0$, we write $a_n = o(b_n)$ or $a_n \prec b_n$. We use $a_n \lesssim b_n$ or $a_n = O(b_n)$ to denote that for sufficiently large n, there exists a constant $C > 0$ independent of n such that $a_n \leq Cb_n$. For a vector $\boldsymbol{v} \in \mathbb{R}^p$, we let $\|\boldsymbol{v}\|_1 := \sum_{i=1}^p |v_i|$, $\|\boldsymbol{v}\|_2 := \sqrt{\sum_{i=1}^p v_i^2}$, and $\|\boldsymbol{v}\|_\infty := \max_{1\leq i\leq p} |v_i|$ denote the ℓ_1, ℓ_2, and ℓ_∞ norms respectively. For a symmetric matrix $\boldsymbol{A}$, we let $\lambda_{\min}(\boldsymbol{A})$ and $\lambda_{\max}(\boldsymbol{A})$ denote its minimum and maximum eigenvalues.

2 The Spike-and-Slab Group Lasso

Let $\boldsymbol{\beta}_g$ denote a real-valued vector of length m_g. We define the *group lasso density* as

$$\boldsymbol{\Psi}(\boldsymbol{\beta}_g|\lambda) = C_g \lambda^{m_g} \exp\left(-\lambda \|\boldsymbol{\beta}_g\|_2\right), \tag{2.1}$$

where $C_g = 2^{-m_g}\pi^{-(m_g-1)/2}\left[\Gamma\left((m_g+1)/2\right)\right]^{-1}$. This prior has been previously considered by Kyung et al. (2010) and Xu and Ghosh (2015) for Bayesian inference in the grouped

regression model (1.1). Kyung et al. (2010) considered a single prior (2.1) on each of the $\boldsymbol{\beta}_g$'s, while Xu and Ghosh (2015) employed (2.1) as the slab in the point-mass mixture (1.3). These authors implemented their models using MCMC.

In this manuscript, we introduce a *continuous* spike-and-slab prior with the group lasso density (2.1) for both the spike *and* the slab. The continuous nature of our prior is critical in facilitating efficient coordinate ascent algorithms for MAP estimation that allow us to bypass the use of MCMC. Letting $\boldsymbol{\beta} = (\boldsymbol{\beta}_1^T, \ldots, \boldsymbol{\beta}_G^T)^T$ under model (1.1), the *spike-and-slab group lasso* (SSGL) is defined as:

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

where $\boldsymbol{\Psi}(\cdot|\lambda)$ denotes the group lasso density (2.1) indexed by hyperparameter λ, and $\theta \in (0,1)$ is a mixing proportion. λ_0 corresponds to the spike which shrinks the entire vector $\boldsymbol{\beta}_g$ towards $\mathbf{0}_{m_g}$, while λ_1 corresponds to the slab. For shorthand notation, we denote $\boldsymbol{\Psi}(\boldsymbol{\beta}_g|\lambda_0)$ as $\boldsymbol{\Psi}_0(\boldsymbol{\beta}_g)$ and $\boldsymbol{\Psi}(\boldsymbol{\beta}_g|\lambda_1)$ as $\boldsymbol{\Psi}_1(\boldsymbol{\beta}_g)$ going forward.

Under the grouped regression model (1.1), we place the SSGL prior (2.2) on $\boldsymbol{\beta}$. In accordance with the recommendations of Moran et al. (2019), we do not scale our prior by the unknown σ. Instead, we place an independent Jeffreys prior on σ^2, i.e.

$$\pi(\sigma^2) \propto \sigma^{-2}. \tag{2.3}$$

The mixing proportion θ in (2.2) can either be fixed deterministically or endowed with a prior $\theta \sim \pi(\theta)$. We will discuss this in detail in Section 3.

3 Characterization and Computation of the Global Posterior Mode

Throughout this section, we let p denote the total number of covariates, i.e. $p = \sum_{g=1}^{G} m_g$. Our goal is to find the maximum *a posteriori* estimates of the regression coefficients $\boldsymbol{\beta} \in \mathbb{R}^p$. This optimization problem is equivalent to a penalized likelihood method in which the logarithm of the prior (2.2) may be reinterpreted as a penalty on the regression coefficients. Similarly to Ročková and George (2018), we will leverage this connection between the Bayesian and frequentist paradigms and introduce the SSGL penalty. This strategy combines the adaptivity of the Bayesian approach with the computational efficiency of existing algorithms in the frequentist literature.

A key component of the SSGL model is θ, the prior expected proportion of groups with large coefficients. Ultimately, we will pursue a fully Bayes approach and place a prior on θ, allowing the SSGL to adapt to the underlying sparsity of the data and perform an automatic multiplicity adjustment Scott and Berger (2010). For ease of exposition, however, we will first consider the case where θ is fixed, echoing the development of Ročková and George (2018). In this situation, the regression coefficients $\boldsymbol{\beta}_g$ are conditionally independent *a priori*, resulting in a separable SSGL penalty. Later we will consider the fully Bayes approach, which will yield the *non-separable* SSGL penalty.

Definition 1. *Given $\theta \in (0, 1)$, the separable SSGL penalty is defined as*

$$pen_S(\boldsymbol{\beta}|\theta) = \log\left[\frac{\pi(\boldsymbol{\beta}|\theta)}{\pi(\mathbf{0}_p|\theta)}\right] = -\lambda_1 \sum_{g=1}^{G} \|\boldsymbol{\beta}_g\|_2 + \sum_{g=1}^{G} \log\left[\frac{p^*_\theta(\mathbf{0}_{m_g})}{p^*_\theta(\boldsymbol{\beta}_g)}\right] \tag{3.1}$$

where

$$p^*_\theta(\boldsymbol{\beta}_g) = \frac{\theta \boldsymbol{\Psi}_1(\boldsymbol{\beta}_g)}{\theta \boldsymbol{\Psi}_1(\boldsymbol{\beta}_g) + (1-\theta)\boldsymbol{\Psi}_0(\boldsymbol{\beta}_g)}. \tag{3.2}$$

The separable SSGL penalty is almost the logarithm of the original prior (2.2); the only modification is an additive constant to ensure that $pen_S(\mathbf{0}_p|\theta) = 0$. The connection between the SSGL and penalized likelihood methods is made clearer when considering the derivative of the separable SSGL penalty, given in the following lemma.

Lemma 1. *The derivative of the separable SSGL penalty satisfies*

$$\frac{\partial pen_S(\boldsymbol{\beta}|\theta)}{\partial \|\boldsymbol{\beta}_g\|_2} = -\lambda^*_\theta(\boldsymbol{\beta}_g) \tag{3.3}$$

where

$$\lambda^*_\theta(\boldsymbol{\beta}_g) = \lambda_1 p^*_\theta(\boldsymbol{\beta}_g) + \lambda_0[1 - p^*_\theta(\boldsymbol{\beta}_g)]. \tag{3.4}$$

Similarly to the SSL, the SSGL penalty is a weighted average of the two regularization parameters, λ_1 and λ_0. The weight $p^*_\theta(\boldsymbol{\beta}_g)$ is the conditional probability that $\boldsymbol{\beta}_g$ was drawn from the slab distribution rather than the spike. Hence, the SSGL features an adaptive regularization parameter which applies different amounts of shrinkage to each group, unlike the group lasso which applies the same shrinkage to each group.

3.1 The Global Posterior Mode

Similarly to the group lasso (Yuan and Lin, 2006), the separable nature of the penalty (3.1) lends itself naturally to a block coordinate ascent algorithm which cycles through the groups. In this section, we first outline the group updates resulting from the Karush-Kuhn-Tucker (KKT) conditions. The KKT conditions provide necessary conditions for the global posterior mode. We then derive a more refined condition for the global mode to aid in optimization for multimodal posteriors.

Following Huang et al. (2012), we assume that within each group, covariates are orthonormal, i.e. $\boldsymbol{X}_g^T \boldsymbol{X}_g = n\mathbf{I}_{m_g}$ for $g = 1, \ldots, G$. If this assumption does not hold, then the $\boldsymbol{X}_g$ matrices can be orthonormalized before fitting the model. As noted by Breheny and Huang (2015), orthonormalization can be done without loss of generality since the resulting solution can be transformed back to the original scale.

Proposition 1. *The necessary conditions for* $\widehat{\boldsymbol{\beta}} = (\widehat{\boldsymbol{\beta}}_1^T, \ldots, \widehat{\boldsymbol{\beta}}_G^T)^T$ *to be a global mode are:*

$$\boldsymbol{X}_g^T(\boldsymbol{Y} - \boldsymbol{X}\widehat{\boldsymbol{\beta}}) = \sigma^2 \lambda_\theta^*(\widehat{\boldsymbol{\beta}}_g) \frac{\widehat{\boldsymbol{\beta}}_g}{\|\boldsymbol{\beta}_g\|_2} \quad \textit{for} \quad \widehat{\boldsymbol{\beta}}_g \neq \mathbf{0}_{m_g}, \tag{3.5}$$

$$\|\boldsymbol{X}_g^T(\boldsymbol{Y} - \boldsymbol{X}\widehat{\boldsymbol{\beta}})\|_2 \leq \sigma^2 \lambda_\theta^*(\widehat{\boldsymbol{\beta}}_g) \quad \textit{for} \quad \widehat{\boldsymbol{\beta}}_g = \mathbf{0}_{m_g}. \tag{3.6}$$

Equivalently,

$$\widehat{\boldsymbol{\beta}}_g = \frac{1}{n}\left(1 - \frac{\sigma^2 \lambda_\theta^*(\widehat{\boldsymbol{\beta}}_g)}{\|\boldsymbol{z}_g\|_2}\right)_+ \boldsymbol{z}_g \tag{3.7}$$

where $\boldsymbol{z}_g = \boldsymbol{X}_g^T \left[\boldsymbol{Y} - \sum_{l \neq g} \boldsymbol{X}_l \widehat{\boldsymbol{\beta}}_l\right]$.

Proof. Follows immediately from Lemma 1 and subdifferential Calculus. □

The above characterization for the global mode is necessary, but not sufficient. A more refined characterization may be obtained by considering the group-wise optimization problem, noting that the global mode is also a maximizer of the gth group, keeping all other groups fixed.

Proposition 2. *The global mode* $\widehat{\boldsymbol{\beta}}_g = \mathbf{0}_{m_g}$ *if and only if* $\|\boldsymbol{z}_g\|_2 \leq \Delta$, *where*

$$\Delta = \inf_{\boldsymbol{\beta}_g} \left\{ \frac{n\|\boldsymbol{\beta}_g\|_2}{2} - \frac{\sigma^2 pen_S(\boldsymbol{\beta}|\theta)}{\|\boldsymbol{\beta}_g\|_2} \right\}. \tag{3.8}$$

The proof for Proposition 2 can be found in Section D.2 of the Supplementary Material. Unfortunately, the threshold Δ is difficult to compute. We instead find an approximation to this threshold. An upper bound is simply that of the soft-threshold solution (3.7), with $\Delta \leq \sigma^2\lambda^*(\boldsymbol{\beta}_g)$. However, when λ_0 is large, this bound may be improved. Similarly to Ročková and George (2018), we provide improved bounds on the threshold in Theorem 1. This result requires the function $h : \mathbb{R}^{m_g} \to \mathbb{R}$, defined as:

$$h(\boldsymbol{\beta}_g) = [\lambda_\theta^*(\boldsymbol{\beta}_g) - \lambda_1]^2 + \frac{2n}{\sigma^2}\log p_\theta^*(\boldsymbol{\beta}_g).$$

Theorem 1. *When* $(\lambda_0 - \lambda_1) > 2\sqrt{n}/\sigma$ *and* $h(\mathbf{0}_{m_g}) > 0$*, the threshold* Δ *is bounded by:*

$$\Delta^L < \Delta < \Delta^U \tag{3.9}$$

where

$$\Delta^L = \sqrt{2n\sigma^2\log[1/p_\theta^*(\mathbf{0}_{m_g})] - \sigma^4 d} + \sigma^2\lambda_1, \tag{3.10}$$

$$\Delta^U = \sqrt{2n\sigma^2\log[1/p_\theta^*(\mathbf{0}_{m_g})]} + \sigma^2\lambda_1, \tag{3.11}$$

and

$$0 < d < \frac{2n}{\sigma^2} - \left(\frac{n}{\sigma^2(\lambda_0 - \lambda_1)} - \frac{\sqrt{2n}}{\sigma}\right)^2 \tag{3.12}$$

When λ_0 is large, $d \to 0$ and the lower bound on the threshold approaches the upper bound, yielding the approximation $\Delta = \Delta^U$. We will ultimately use this approximation in our block coordinate ascent algorithm.

3.2 The Non-Separable SSGL penalty

As discussed earlier, a key reason for adopting a Bayesian strategy is that it allows the model to borrow information across groups and self-adapt to the true underlying sparsity

in the data. This is achieved by placing a prior on θ, the proportion of groups with non-zero coefficients. We now outline this fully Bayes strategy and the resulting *non-separable* SSGL penalty. With the inclusion of the prior $\theta \sim \pi(\theta)$, the marginal prior for the regression coefficients has the following form:

$$\pi(\boldsymbol{\beta}) = \int_0^1 \prod_{g=1}^{G} [\theta \boldsymbol{\Psi}_1(\boldsymbol{\beta}_g) + (1-\theta)\boldsymbol{\Psi}_0(\boldsymbol{\beta}_g)] d\pi(\theta) \tag{3.13}$$

$$= \left(\prod_{g=1}^{G} C_g \lambda_1^{m_g} \right) e^{-\lambda_1 \sum_{g=1}^{G} \|\boldsymbol{\beta}_g\|_2} \int_0^1 \frac{\theta^G}{\prod_{g=1}^{G} p^*_\theta(\boldsymbol{\beta}_g)} d\pi(\theta), \tag{3.14}$$

The non-separable SSGL penalty is then defined similarly to the separable penalty, where again we have centered the penalty to ensure $pen_{NS}(\mathbf{0}_p) = 0$.

Definition 2. *The non-separable SSGL (NS-SSGL) penalty with $\theta \sim \pi(\theta)$ is defined as*

$$pen_{NS}(\boldsymbol{\beta}) = \log \left[\frac{\pi(\boldsymbol{\beta})}{\pi(\mathbf{0}_p)} \right] = -\lambda_1 \sum_{g=1}^{G} \|\boldsymbol{\beta}_g\|_2 + \log \left[\frac{\int_0^1 \theta^G / \prod_{g=1}^{G} p^*_\theta(\boldsymbol{\beta}_g) d\pi(\theta)}{\int_0^1 \theta^G / \prod_{g=1}^{G} p^*_\theta(\mathbf{0}_{m_g}) d\pi(\theta)} \right]. \tag{3.15}$$

Although the penalty (3.14) appears intractable, intuition is again obtained by considering the derivative. Following the same line of argument as Ročková and George (2018), the derivative of (3.14) is given in the following lemma.

Lemma 2.

$$\frac{\partial pen_{NS}(\boldsymbol{\beta})}{\partial \|\boldsymbol{\beta}_g\|_2} \equiv \lambda^*(\boldsymbol{\beta}_g; \boldsymbol{\beta}_{\setminus g}), \tag{3.16}$$

where

$$\lambda^*(\boldsymbol{\beta}_g; \boldsymbol{\beta}_{\setminus g}) = p^*(\boldsymbol{\beta}_g; \boldsymbol{\beta}_{\setminus g})\lambda_1 + [1 - p^*(\boldsymbol{\beta}_g; \boldsymbol{\beta}_{\setminus g})]\lambda_0 \tag{3.17}$$

and

$$p^*(\boldsymbol{\beta}_g; \boldsymbol{\beta}_{\setminus g}) \equiv p^*_{\theta_g}(\boldsymbol{\beta}_g), \quad \textit{with} \quad \theta_g = \mathbb{E}[\theta | \boldsymbol{\beta}_{\setminus g}]. \tag{3.18}$$

That is, the marginal prior from (3.14) is rendered tractable by considering each group of regression coefficients separately, conditional on the remaining coefficients. Such a conditional strategy is motivated by the group-wise updates for the separable penalty considered in the previous section. Thus, our optimization strategy for the non-separable penalty will be very similar to the separable case, except instead of a fixed value for θ, we will impute the mean of θ conditioned on the remaining regression coefficients.

We now consider the form of the conditional mean, $\mathbb{E}[\theta|\widehat{\boldsymbol{\beta}}_{\setminus g}]$. As noted by Ročková and George (2018), when the number of groups is large, this conditional mean can be replaced by $\mathbb{E}[\theta|\widehat{\boldsymbol{\beta}}]$; we will proceed with the same approximation. For the prior on θ, we will use the standard beta prior $\theta \sim \mathcal{B}(a, b)$. With the choices $a = 1$ and $b = G$ for these hyperparameters, this prior results in an automatic multiplicity adjustment for the regression coefficients (Scott and Berger (2010)).

We now examine the conditional distribution $\pi(\theta|\widehat{\boldsymbol{\beta}})$. Suppose that the number of groups with non-zero coefficients is $\widehat{q}$, and assume without loss of generality that the first $\widehat{q}$ groups have non-zero coefficients. Then,

$$\pi(\theta|\widehat{\boldsymbol{\beta}}) \propto \theta^{a-1}(1-\theta)^{b-1}(1-\theta z)^{G-\widehat{q}} \prod_{g=1}^{\widehat{q}}(1-\theta x_g), \tag{3.19}$$

with $z = 1 - \frac{\lambda_1}{\lambda_0}$ and $x_g = (1 - \frac{\lambda_1}{\lambda_0} e^{\|\widehat{\boldsymbol{\beta}}_g\|_2(\lambda_0 - \lambda_1)})$. Similarly to Ročková and George (2018), this distribution is a generalization of the Gauss hypergeometric distribution. Consequently, the expectation may be written as

$$\mathbb{E}[\theta|\widehat{\boldsymbol{\beta}}] = \frac{\int_0^1 \theta^a(1-\theta)^{b-1}(1-\theta z)^{G-\widehat{q}} \prod_{g=1}^{\widehat{q}}(1-\theta x_g)d\theta}{\int_0^1 \theta^{a-1}(1-\theta)^{b-1}(1-\theta z)^{G-\widehat{q}} \prod_{g=1}^{\widehat{q}}(1-\theta x_g)d\theta}. \tag{3.20}$$

While the above expression (3.20) appears laborious to compute, it admits a much simpler form when λ_0 is very large. Using a slight modification to the arguments of Ročková and

George (2016), we obtain this simpler form in Lemma 3.

Lemma 3. *Assume $\pi(\theta|\widehat{\boldsymbol{\beta}})$ is distributed according to* (3.19)*. Let $\widehat{q}$ be the number of groups with non-zero coefficients. Then as $\lambda_0 \to \infty$,*

$$\mathbb{E}[\theta|\widehat{\boldsymbol{\beta}}] = \frac{a+\widehat{q}}{a+b+G}. \tag{3.21}$$

The proof for Lemma 3 is in Section D.2 of the Supplementary Material. We note that the expression (3.21) is essentially the usual posterior mean of θ under a beta prior. Intuitively, as λ_0 diverges, the weights $p^*_\theta(\boldsymbol{\beta}_g)$ concentrate at zero and one, yielding the familiar form for $\mathbb{E}[\theta|\widehat{\boldsymbol{\beta}}]$. With this in hand, we are now in a position to outline the block coordinate ascent algorithm for the non-separable SSGL.

3.3 Optimization

The KKT conditions for the non-separable SSGL penalty yield the following necessary condition for the global mode:

$$\widehat{\boldsymbol{\beta}}_g \leftarrow \frac{1}{n}\left(1 - \frac{\sigma^2 \lambda^*_{\widehat{\theta}}(\widehat{\boldsymbol{\beta}}_g)}{\|\boldsymbol{z}_g\|_2}\right)_+ \boldsymbol{z}_g, \tag{3.22}$$

where $\boldsymbol{z}_g = \boldsymbol{X}_g^T\left[\boldsymbol{Y} - \sum_{l\neq g}\boldsymbol{X}_l\widehat{\boldsymbol{\beta}}_l\right]$ and $\widehat{\theta}$ is the mean (3.21), conditioned on the previous value of $\boldsymbol{\beta}$. As before, (3.22) is sufficient for a local mode, but not the global mode. When $p \gg n$ and λ_0 is large, the posterior will be highly multimodal. As in the separable case, we require a refined thresholding scheme that will eliminate some of these suboptimal local modes from consideration. In approximating the group-wise conditional mean $\mathbb{E}[\theta|\widehat{\boldsymbol{\beta}}_{\setminus g}]$ with $\mathbb{E}[\theta|\widehat{\boldsymbol{\beta}}]$, we do not require group-specific thresholds. Instead, we can use the threshold

given in Proposition 2 and Theorem 1 where θ is replaced with the current update (3.21). In particular, we shall use the upper bound Δ^U in our block coordinate ascent algorithm.

Similarly to Ročková and George (2018), we combine the refined threshold, Δ^U with the soft thresholding operation (3.22), to yield the following update for $\widehat{\boldsymbol{\beta}}_g$ at iteration k:

$$\boldsymbol{\beta}_g^{(k)} \leftarrow \frac{1}{n}\left(1 - \frac{\sigma^{2(k)}\lambda^*(\boldsymbol{\beta}_g^{(k-1)}; \theta^{(k)})}{\|\boldsymbol{z}_g\|_2}\right)_+ \boldsymbol{z}_g \, \mathbb{I}(\|\boldsymbol{z}_g\|_2 > \Delta^U) \tag{3.23}$$

where $\theta^{(k)} = \mathbb{E}[\theta|\boldsymbol{\beta}^{(k-1)}]$. Technically, θ should be updated after each group $\boldsymbol{\beta}_g$ is updated. In practice, however, there will be little change after one group is updated and so we will update both θ and Δ^U after every M iterations with a default value of $M = 10$.

With the Jeffreys prior $\pi(\sigma^2) \propto \sigma^{-2}$, the error variance σ^2 also has a closed form update:

$$\sigma^{2(k)} \leftarrow \frac{\|\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta}^{(k-1)}\|_2^2}{n+2}. \tag{3.24}$$

The complete optimization algorithm is given in Algorithm 1 of Section A.1 of the Supplementary Material. The computational complexity of this algorithm is $\mathcal{O}(np)$ per iteration, where $p = \sum_{g=1}^G m_g$. It takes $\mathcal{O}(nm_g)$ operations to compute the partial residual $\boldsymbol{z}_g$ for the gth group, for a total cost of $\mathcal{O}(n\sum_{g=1}^G m_g) = \mathcal{O}(np)$. Similarly, it takes $\mathcal{O}(np)$ cost to compute the sum of squared residuals $\|\boldsymbol{Y} - \boldsymbol{X}\widehat{\boldsymbol{\beta}}\|_2^2$ to update the variance parameter σ^2. The computational complexity of our algorithm matches that of the usual gradient descent algorithms for lasso and group lasso (Friedman et al., 2010).

As a non-convex method, it is not guaranteed that SSGL will find the global posterior mode, only a local mode. However, the refined thresholding scheme (Theorem 1) and a warm start initialization strategy (described in detail in Section A.2 of the Supplementary Material) enable SSGL to eliminate a number sub-optimal local modes from consideration in a similar manner to Ročková and George (2018). To briefly summarize the initialization

strategy, we tune λ_0 from an increasing sequence of values, and we further scale λ_0 by $\sqrt{m_g}$ for each gth group to ensure that the amount of penalization is on the same scale for groups of potentially different sizes (Huang et al., 2012). Meanwhile, we keep λ_1 fixed at a small value so that selected groups have minimal shrinkage. See Section A.2 of the Supplementary Material for detailed discussion of choosing (λ_0, λ_1).

4 Approaches to Inference

While the above procedure allows us to find the posterior mode of $\boldsymbol{\beta}$, providing a measure of uncertainty around our estimate is a challenging task. One possible solution is to run MCMC where the algorithm is initialized at the posterior mode. By starting the MCMC chain at the mode, the algorithm should converge faster. However, this is still not ideal, as it can be computationally burdensome in high dimensions. Instead, we will adopt ideas from a recent line of research (van de Geer et al. (2014), Javanmard and Montanari (2018)) based on de-biasing estimates from high-dimensional regression. These ideas were derived in the context of lasso regression, and we will explore the extent to which they work for the SSGL penalty. Define $\widehat{\boldsymbol{\Sigma}} = \boldsymbol{X}^T\boldsymbol{X}/n$ and let $\widehat{\boldsymbol{\Theta}}$ be an approximate inverse of $\widehat{\boldsymbol{\Sigma}}$. We define

$$\widehat{\boldsymbol{\beta}}_d = \widehat{\boldsymbol{\beta}} + \widehat{\boldsymbol{\Theta}}\boldsymbol{X}^T(\boldsymbol{Y} - \boldsymbol{X}\widehat{\boldsymbol{\beta}})/n. \tag{4.1}$$

where $\widehat{\boldsymbol{\beta}}$ is the MAP estimator of $\boldsymbol{\beta}$ under the SSGL model. By van de Geer et al. (2014), this quantity $\widehat{\boldsymbol{\beta}}_d$ has the following asymptotic distribution:

$$\sqrt{n}(\widehat{\boldsymbol{\beta}}_d - \boldsymbol{\beta}) \sim \mathcal{N}(\mathbf{0}, \sigma^2\widehat{\boldsymbol{\Theta}}\widehat{\boldsymbol{\Sigma}}\widehat{\boldsymbol{\Theta}}^T). \tag{4.2}$$

For our inference procedure, we replace the population variance σ^2 in (4.2) with the modal estimate $\widehat{\sigma}^2$ from the SSGL model. To estimate $\widehat{\boldsymbol{\Theta}}$, we utilize the nodewise regression

approach developed in Meinshausen and Bühlmann (2006) and van de Geer et al. (2014). We describe this estimation procedure for $\widehat{\boldsymbol{\Theta}}$ in Section A.3 of the Supplementary Material.

Let $\widehat{\beta}_{dj}$ denote the jth coordinate of $\widehat{\boldsymbol{\beta}}_d$. We have from (4.2) that the $100(1-\alpha)\%$ asymptotic pointwise confidence intervals for $\beta_j, j = 1, \ldots, p$, are

$$[\widehat{\beta}_{dj} - c(\alpha, n, \widehat{\sigma}^2), \widehat{\beta}_{dj} + c(\alpha, n, \widehat{\sigma}^2)], \tag{4.3}$$

where $c(\alpha, n, \widehat{\sigma}^2) := \Phi^{-1}(1-\alpha/2)\sqrt{\widehat{\sigma}^2(\widehat{\boldsymbol{\Theta}}\widehat{\boldsymbol{\Sigma}}\widehat{\boldsymbol{\Theta}}^T)_{jj}/n}$ and $\Phi(\cdot)$ denotes the cdf of $\mathcal{N}(0,1)$. It should be noted that our posterior mode estimates should have less bias than existing estimates such as the group lasso. Therefore, the goal of the de-biasing procedure is less about de-biasing the posterior mode estimates, and more about providing an estimator with an asymptotic normal distribution from which we can perform inference.

To assess the ability of this procedure to obtain accurate confidence intervals (4.3) with $\alpha = 0.05$, we run a small simulation study with $n = 100$, $G = 100$ or $n = 300, G = 300$, and each of the G groups having $m = 2$ covariates. We generate the covariates from a multivariate normal distribution with mean $\mathbf{0}$ and an AR(1) covariance structure with correlation ρ. The two covariates from each group are the linear and squared term from the original covariates. We set the first seven elements of $\boldsymbol{\beta}$ equal to $(0, 0.5, 0.25, 0.1, 0, 0, 0.7)$ and the remaining elements equal to zero. Lastly, we try $\rho = 0$ and $\rho = 0.7$. Table 1 shows the coverage probabilities across 1000 simulations for all scenarios looked at. We see that important covariates, i.e. covariates with a nonzero corresponding β_j, have coverage near 0.85 when $n = 100$ under either correlation structure, though this increases to nearly the nominal rate when $n = 300$. The remaining covariates (null covariates) achieve the nominal level regardless of the sample size or correlation present.

	ρ	Important covariates	Null covariates
$n = 100, G = 100$	0.0	0.83	0.93
	0.7	0.85	0.94
$n = 300, G = 300$	0.0	0.93	0.95
	0.7	0.92	0.95

Table 1: Coverage probabilities for de-biasing simulation.

5 Nonparametric Spike-and-Slab Lasso

We now introduce the nonparametric spike-and-slab lasso (NPSSL). The NPSSL allows for flexible modeling of a response surface with minimal assumptions regarding its functional form. We consider two cases for the NPSSL: (i) a main effects only model, and (ii) a model with both main and interaction effects.

5.1 Main Effects

We first consider the main effects NPSSL model. Here, we assume that the response surface may be decomposed into the sum of univariate functions of each of the p covariates. That is, we have the following model:

$$y_i = \sum_{j=1}^{p} f_j(X_{ij}) + \varepsilon_i, \quad \varepsilon_i \sim \mathcal{N}(0, \sigma^2). \tag{5.1}$$

Following Ravikumar et al. (2009), we assume that each f_j, $j = 1, \ldots, p$, may be approximated by a linear combination of basis functions $\mathcal{B}_j = \{g_{j1}, \ldots, g_{jd}\}$, i.e.,

$$f_j(X_{ij}) \approx \sum_{k=1}^{d} g_{jk}(X_{ij})\beta_{jk} \tag{5.2}$$

where $\boldsymbol{\beta}_j = (\beta_{j1}, \ldots, \beta_{jd})^T$ are the unknown weights. Let $\widetilde{\boldsymbol{X}}_j$ denote the $n \times d$ matrix with the (i,k)th entry $\widetilde{\boldsymbol{X}}_j(i,k) = g_{jk}(X_{ij})$. Then, (5.1) may be represented in matrix form as

$$\boldsymbol{Y} - \boldsymbol{\delta} = \sum_{j=1}^{p} \widetilde{\boldsymbol{X}}_j \boldsymbol{\beta}_j + \boldsymbol{\varepsilon}, \quad \boldsymbol{\varepsilon} \sim \mathcal{N}_n(\boldsymbol{0}, \sigma^2 \boldsymbol{I}_n), \tag{5.3}$$

where $\boldsymbol{\delta}$ is a vector of the lower-order truncation bias. Note that we assume the response $\boldsymbol{Y}$ has been centered and so we do not include a grand mean $\boldsymbol{\mu}$ in (5.3). Thus, we do not require the main effects to integrate to zero as in Wei et al. (2020). We do, however, require the matrices $\widetilde{\boldsymbol{X}}_j, j = 1, \ldots, p$, to be orthogonal, as discussed in Section 3. Note that the entire design matrix does not need to be orthogonal; only the group-specific matrices need to be. We can enforce this in practice by either using orthonormal basis functions or by orthornormalizing the $\widetilde{\boldsymbol{X}}_j$ matrices before fitting the model.

We assume that $\boldsymbol{Y}$ depends on only a small number of the p covariates so that many of the f_j's have a negligible contribution to (5.1). This is equivalent to assuming that most of the weight vectors $\boldsymbol{\beta}_j$ have all zero elements. If the jth covariate is determined to be predictive of $\boldsymbol{Y}$, then f_j has a non-negligible contribution to (5.1). In this case, we want to include the *entire* basis function approximation to f_j in the model.

The above situation is a natural fit for the SSGL. We have p groups where each group is either included as a whole or not included in the model. The design matrices for each group are exactly the matrices of basis functions, $\widetilde{\boldsymbol{X}}_j, j = 1, \ldots, p$. We will utilize the non-separable SSGL penalty developed in Section 3.2 to enforce this group-sparsity behavior in the model (5.3). More specifically, we seek to maximize the objective function with respect to $\boldsymbol{\beta} = (\boldsymbol{\beta}_1^T, \ldots, \boldsymbol{\beta}_p^T)^T \in \mathbb{R}^{pd}$ and σ^2:

$$L(\boldsymbol{\beta}, \sigma^2) = -\frac{1}{2\sigma^2} \|\boldsymbol{Y} - \sum_{j=1}^{p} \widetilde{\boldsymbol{X}}_j \boldsymbol{\beta}_j\|_2^2 - (n+2) \log \sigma + pen_{NS}(\boldsymbol{\beta}). \tag{5.4}$$

To find the estimators of $\boldsymbol{\beta}$ and σ^2, we use Algorithm 1 in Section A.1 of the Supplementary Material. Similar additive models have been proposed by a number of authors including Ravikumar et al. (2009) and Wei et al. (2020). However, our proposed NPSSL method has a number of advantages. First, we allow the noise variance σ^2 to be unknown, unlike Ravikumar et al. (2009). Accurate estimates of σ^2 are important to avoid overfitting the noise beyond the signal. Secondly, we use a block-descent algorithm to quickly target the modes of the posterior, whereas Wei et al. (2020) utilize MCMC. Finally, our SSGL algorithm automatically thresholds negligible groups to zero, negating the need for a post-processing thresholding step.

5.2 Main and Interaction Effects

The main effects model (5.1) allows for each covariate to have a nonlinear contribution to the model, but assumes a linear relationship *between* the covariates. In some applications, this assumption may be too restrictive. For example, in the environmental exposures data which we analyze in Section 8.2, we may expect high levels of two toxins to have an even more adverse effect on a person's health than high levels of either of the two toxins. Such an effect may be modeled by including interaction effects between the covariates.

Here, we extend the NPSSL to include interaction effects. We consider only second-order interactions between the covariates, but our model can easily be extended to include even higher-order interactions. We assume that the interaction effects may be decomposed into the sum of bivariate functions of each pair of covariates, yielding the model:

$$y_i = \sum_{j=1}^{p} f_j(X_{ij}) + \sum_{k=1}^{p-1} \sum_{l=k+1}^{p} f_{kl}(X_{ik}, X_{il}) + \varepsilon_i, \quad \varepsilon_i \sim \mathcal{N}(0, \sigma^2). \tag{5.5}$$

For the interaction terms, we follow Wei et al. (2020) and approximate f_{kl} using the

outer product of the basis functions of the interacting covariates:

$$f_{kl}(X_{ik}, X_{il}) \approx \sum_{s=1}^{d^*} \sum_{r=1}^{d^*} g_{ks}(X_{ik}) g_{lr}(X_{il}) \beta_{klsr} \tag{5.6}$$

where $\boldsymbol{\beta}_{kl} = (\beta_{kl11}, \ldots, \beta_{kl1d^*}, \beta_{kl21}, \ldots, \beta_{kld^*d^*})^T \in \mathbb{R}^{d^{*2}}$ is the vector of unknown weights. We let $\widetilde{\boldsymbol{X}}_{kl}$ denote the $n \times d^{*2}$ matrix with rows

$$\widetilde{\boldsymbol{X}}_{kl}(i, \cdot) = \text{vec}(\boldsymbol{g}_k(X_{ik}) \boldsymbol{g}_l(X_{il})^T),$$

where $\boldsymbol{g}_k(X_{ik}) = (g_{k1}(X_{ik}), \ldots, g_{kd^*}(X_{ik}))^T$. Then, (5.5) may be represented in matrix form as

$$\boldsymbol{Y} - \boldsymbol{\delta} = \sum_{j=1}^{p} \widetilde{\boldsymbol{X}}_j \boldsymbol{\beta}_j + \sum_{k=1}^{p-1} \sum_{l=k+1}^{p} \widetilde{\boldsymbol{X}}_{kl} \boldsymbol{\beta}_{kl} + \boldsymbol{\varepsilon}, \quad \boldsymbol{\varepsilon} \sim \mathcal{N}_n(\boldsymbol{0}, \sigma^2 \boldsymbol{I}_n), \tag{5.7}$$

where $\boldsymbol{\delta}$ is a vector of the lower-order truncation bias. We again assume $\boldsymbol{Y}$ has been centered and so do not include a grand mean in (5.7). We do not constrain f_{kl} to integrate to zero as in Wei et al. (2020). However, we do ensure that the main effects are not in the linear span of the interaction functions. That is, we require the "main effect" matrices $\widetilde{\boldsymbol{X}}_l$ and $\widetilde{\boldsymbol{X}}_k$ to be orthogonal to the "interaction" matrix $\widetilde{\boldsymbol{X}}_{kl}$. This condition is needed to maintain identifiability for both the main and interaction effects in the model. In practice, we enforce this condition by setting the interaction design matrix to be the residuals of the regression of $\widetilde{\boldsymbol{X}}_k \circ \widetilde{\boldsymbol{X}}_l$ on $\widetilde{\boldsymbol{X}}_k$ and $\widetilde{\boldsymbol{X}}_l$.

Note that the current representation does not enforce strong hierarchy. That is, interaction terms can be included even if their corresponding main effects are removed from the model. However, the NPSSL model can be easily modified to accommodate strong hierarchy. If hierarchy is desired, the "interaction" matrices can be augmented to contain both main and interaction effects, as in Lim and Hastie (2015), i.e. the "interaction" matrices

in (5.7) would be $\widetilde{\boldsymbol{X}}_{kl}^{\text{aug}} = [\widetilde{\boldsymbol{X}}_k, \widetilde{\boldsymbol{X}}_l, \widetilde{\boldsymbol{X}}_{kl}]$, instead of simply $\widetilde{\boldsymbol{X}}_{kl}$. This augmented model is overparameterized since the main effects still have their own separate design matrices as well (to ensure that main effects can still be selected even if $\boldsymbol{\beta}_{kl}^{\text{aug}} = \mathbf{0}$). However, this ensures that interaction effects are only selected if the corresponding main effects are also in the model.

In the interaction model, we either include $\boldsymbol{\beta}_{kl}$ in the model (5.7) if there is a non-negligible interaction between the kth and lth covariates, or we estimate $\widehat{\boldsymbol{\beta}}_{kl} = \mathbf{0}_{d^*2}$ if such an interaction is negligible. With the non-separable SSGL penalty, the objective function is:

$$L(\boldsymbol{\beta}, \sigma^2) = -\frac{1}{2\sigma^2}\|\boldsymbol{Y} - \sum_{j=1}^{p} \widetilde{\boldsymbol{X}}_j\boldsymbol{\beta}_j - \sum_{k=1}^{p-1}\sum_{l=k+1}^{p} \widetilde{\boldsymbol{X}}_{kl}\boldsymbol{\beta}_{kl}\|_2^2 + pen_{NS}(\boldsymbol{\beta}) - (n+2)\log\sigma, \tag{5.8}$$

where $\boldsymbol{\beta} = (\boldsymbol{\beta}_1^T, \ldots, \boldsymbol{\beta}_p^T, \boldsymbol{\beta}_{12}^T, \ldots \boldsymbol{\beta}_{(p-1)p}^T)^T \in \mathbb{R}^{pd+p(p-1)d^{*2}/2}$. We can again use Algorithm 1 in Section A.1 of the Supplementary Material to find the modal estimates of $\boldsymbol{\beta}$ and σ^2.

6 Asymptotic Theory for the SSGL and NPSSL

In this section, we derive asymptotic properties for the separable SSGL and NPSSL models. We first note some differences between our theory and the theory in Ročková and George (2018). First, we prove *joint* consistency in estimation of both the unknown $\boldsymbol{\beta}$ *and* the unknown σ^2, whereas Ročková and George (2018) proved their result only for $\boldsymbol{\beta}$, assuming known variance $\sigma^2 = 1$. Secondly, Ročková and George (2018) established convergence rates for the global posterior mode and the full posterior separately, whereas we establish a contraction rate ϵ_n for the full posterior only. Our rate ϵ_n satisfies $\epsilon_n \to 0$ as $n \to \infty$

(i.e. the full posterior collapses to the true $(\boldsymbol{\beta}, \sigma^2)$ almost surely as $n \to \infty$), and hence, it automatically follows that the posterior mode is a consistent estimator of $(\boldsymbol{\beta}, \sigma^2)$. Finally, we also derive a posterior contraction rate for nonparametric additive regression, not just linear regression. All proofs for the theorems in this section can be found in Section D.3 of the Supplementary Material.

6.1 Grouped Linear Regression

We work under the frequentist assumption that there is a true model,

$$\boldsymbol{Y} = \sum_{g=1}^{G} \boldsymbol{X}_g \boldsymbol{\beta}_{0g} + \boldsymbol{\varepsilon}, \quad \boldsymbol{\varepsilon} \sim \mathcal{N}_n(\boldsymbol{0}, \sigma_0^2 \mathbf{I}_n), \tag{6.1}$$

where $\boldsymbol{\beta}_0 = (\boldsymbol{\beta}_{01}^T, \ldots, \boldsymbol{\beta}_{0G}^T)^T$ and $\sigma_0^2 \in (0, \infty)$. Denote $\boldsymbol{X} = [\boldsymbol{X}_1, \ldots, \boldsymbol{X}_G]$ and $\boldsymbol{\beta} = (\boldsymbol{\beta}_1^T, \ldots, \boldsymbol{\beta}_G^T)^T$. Suppose we endow $(\boldsymbol{\beta}, \sigma^2)$ under model (6.1) with the following prior:

$$\begin{aligned} \pi(\boldsymbol{\beta}|\theta) \sim & \prod_{g=1}^{G} \left[(1-\theta) \boldsymbol{\Psi}(\boldsymbol{\beta}_g | \lambda_0) + \theta \boldsymbol{\Psi}(\boldsymbol{\beta}_g | \lambda_1) \right], \\ \theta \sim & \ \mathcal{B}(a, b), \\ \sigma^2 \sim & \ \mathcal{IG}(c_0, d_0), \end{aligned} \tag{6.2}$$

where $c_0 > 0$ and $d_0 > 0$ are fixed constants and the hyperparameters (a, b) in the prior on θ are to be chosen later.

Remark 1. *In our implementation of the SSGL model, we endowed σ^2 with an improper prior, $\pi(\sigma^2) \propto \sigma^{-2}$. This can be viewed as a limiting case of the $\mathcal{IG}(c_0, d_0)$ prior with $c_0 \to 0, d_0 \to 0$. This improper prior is fine for implementation since it leads to a proper posterior, but for our theoretical investigation, we require the priors on $(\boldsymbol{\beta}, \sigma^2)$ to be proper.*

6.1.1 Posterior Contraction Rates

Let $m_{\max} = \max_{1 \leq j \leq G} m_g$ and let $p = \sum_{g=1}^{G} m_g$. Let S_0 be the set containing the indices of the true nonzero groups, where $S_0 \subseteq \{1, \ldots, G\}$ with cardinality $s_0 = |S_0|$. We make the following assumptions:

(A1) Assume that $G \gg n$, $\log(G) = o(n)$, and $m_{\max} = O(\log G / \log n)$.

(A2) The true number of nonzero groups satisfies $s_0 = o(n / \log G)$.

(A3) There exists a constant $k > 0$ so that $\lambda_{\max}(\boldsymbol{X}^T \boldsymbol{X}) \leq k n^{\alpha}$, for some $\alpha \in [1, \infty)$.

(A4) Let $\xi \subset \{1, \ldots, G\}$, and let $\boldsymbol{X}_\xi$ denote the submatrix of $\boldsymbol{X}$ that contains the submatrices with groups indexed by ξ. There exist constants $\nu_1 > 0$, $\nu_2 > 0$, and an integer $\bar{p}$ satisfying $s_0 = o(\bar{p})$ and $\bar{p} = o(s_0 \log n)$, so that $n\nu_1 \leq \lambda_{\min}(\boldsymbol{X}_\xi^T \boldsymbol{X}_\xi) \leq \lambda_{\max}(\boldsymbol{X}_\xi^T \boldsymbol{X}_\xi) \leq n\nu_2$ for any model of size $|\xi| \leq \bar{p}$.

(A5) $\|\boldsymbol{\beta}_0\|_\infty = O(\log G)$.

Assumption (A1) allows the number of groups G and total number of covariates p to grow at nearly exponential rate with sample size n. The size of each individual group may also grow as n grows, but should grow at a slower rate than $n / \log n$. Assumption (A2) specifies the growth rate for the true model size s_0. Assumption (A3) bounds the eigenvalues of $\boldsymbol{X}^T \boldsymbol{X}$ from above and is less stringent than requiring all the eigenvalues of the Gram matrix $(\boldsymbol{X}^T \boldsymbol{X}/n)$ to be bounded away from infinity. Assumption (A4) ensures that $\boldsymbol{X}^T \boldsymbol{X}$ is locally invertible over sparse sets. In general, conditions (A3)-(A4) are difficult to verify, but they can be shown to hold with high probability for certain classes of matrices where the rows of $\boldsymbol{X}$ are independent and sub-Gaussian (Mendelson and Pajor (2006), Raskutti

et al. (2010)). Finally, Assumption (A5) places a restriction on the growth rate of the maximum signal size for the true $\boldsymbol{\beta}_0$.

We now state our main theorem on the posterior contraction rates for the SSGL prior (6.2) under model (6.1). Let $\mathbb{P}_0$ denote the probability measure underlying the truth (6.1) and $\Pi(\cdot|\boldsymbol{Y})$ denote the posterior distribution under the prior (6.2) for $(\boldsymbol{\beta}, \sigma^2)$.

Theorem 2 (posterior contraction rates)**.** *Let $\epsilon_n = \sqrt{s_0 \log G/n}$, and suppose that Assumptions (A1)-(A5) hold. Under model (6.1), suppose that we endow $(\boldsymbol{\beta}, \sigma^2)$ with the prior (6.2). For the hyperparameters in the $\mathcal{B}(a, b)$ prior on θ, we choose $a = 1, b = G^c$, $c > 2$. Further, we set $\lambda_0 = (1-\theta)/\theta$ and $\lambda_1 \asymp 1/n$ in the SSGL prior. Then*

$$\Pi\left(\boldsymbol{\beta} : \|\boldsymbol{\beta} - \boldsymbol{\beta}_0\|_2 \geq M_1 \sigma_0 \epsilon_n | \boldsymbol{Y}\right) \to 0 \text{ a.s. } \mathbb{P}_0 \text{ as } n, G \to \infty, \tag{6.3}$$

$$\Pi\left(\boldsymbol{\beta} : \|\boldsymbol{X}\boldsymbol{\beta} - \boldsymbol{X}\boldsymbol{\beta}_0\|_2 \geq M_2 \sigma_0 \sqrt{n} \epsilon_n | \boldsymbol{Y}\right) \to 0 \text{ a.s. } \mathbb{P}_0 \text{ as } n, G \to \infty, \tag{6.4}$$

$$\Pi\left(\sigma^2 : |\sigma^2 - \sigma_0^2| \geq 4\sigma_0^2 \epsilon_n | \boldsymbol{Y}\right) \to 0 \text{ as } n \to \infty, \text{ a.s. } \mathbb{P}_0 \text{ as } n, G \to \infty, \tag{6.5}$$

for some $M_1 > 0, M_2 > 0$.

Remark 2. *In the case where $G = p$ and $m_1 = \ldots = m_G = 1$, the ℓ_2 and prediction error rates in (6.3)-(6.4) reduce to the familiar optimal rates of $\sqrt{s_0 \log p/n}$ and $\sqrt{s_0 \log p}$ respectively.*

Remark 3. *Eq. (6.5) demonstrates that our model also consistently estimates the unknown variance σ^2, therefore providing further theoretical justification for placing an independent prior on σ^2, as advocated by Moran et al. (2019).*

6.1.2 Dimensionality Recovery

Although the posterior mode is exactly sparse, the SSGL prior is absolutely continuous so it assigns zero mass to exactly sparse vectors. To approximate the model size under the

SSGL model, we use the following generalized notion of sparsity (Bhattacharya et al., 2015). For $\omega_g > 0$, we define the generalized inclusion indicator and generalized dimensionality, respectively, as

$$\gamma_{\omega_g}(\boldsymbol{\beta}_g) = I(\|\boldsymbol{\beta}_g\|_2 > \omega_g) \text{ and } |\boldsymbol{\gamma}(\boldsymbol{\beta})| = \sum_{g=1}^{G} \gamma_{\omega_g}(\boldsymbol{\beta}_g). \tag{6.6}$$

In contrast to Bhattacharya et al. (2015) and Ročková and George (2018), we allow the threshold ω_g to be different for each group, owing to the fact that the group sizes m_g may not necessarily all be the same. However, the ω_g's, $g = 1, \ldots, G$, should still tend towards zero as n increases, so that $|\boldsymbol{\gamma}(\boldsymbol{\beta})|$ provides a good approximation to $\#\{g : \boldsymbol{\beta}_g \neq \mathbf{0}_{m_g}\}$.

Consider as the threshold,

$$\omega_g \equiv \omega_g(\lambda_0, \lambda_1, \theta) = \frac{1}{\lambda_0 - \lambda_1} \log\left[\frac{1-\theta}{\theta} \frac{\lambda_0^{m_g}}{\lambda_1^{m_g}}\right] \tag{6.7}$$

Note that for large λ_0, this threshold rapidly approaches zero. Analogous to Ročková (2018) and Ročková and George (2018), any vectors $\boldsymbol{\beta}_g$ that satisfy $\|\boldsymbol{\beta}_g\|_2 = \omega_g$ correspond to the intersection points between the two group lasso densities in the separable SSGL prior (2.2), or when the second derivative $\partial^2 pen_S(\boldsymbol{\beta}|\theta)/\partial\|\boldsymbol{\beta}_g\|_2^2 = 0.5$. The value ω_g represents the turning point where the slab has dominated the spike, and thus, the sharper the spike (when λ_0 is large), the smaller the threshold.

Using the notion of generalized dimensionality (6.6) with (6.7) as the threshold, we have the following theorem.

Theorem 3 (dimensionality)**.** *Suppose that the same conditions as those in Theorem 2 hold. Then under* (6.1)*, for sufficiently large* $M_3 > 0$*,*

$$\sup_{\boldsymbol{\beta}_0} \mathbb{E}_{\boldsymbol{\beta}_0} \Pi\left(\boldsymbol{\beta} : |\boldsymbol{\gamma}(\boldsymbol{\beta})| > M_3 s_0 | \boldsymbol{Y}\right) \to 0 \text{ as } n, G \to \infty. \tag{6.8}$$

Theorem 3 shows that the expected posterior probability that the generalized dimension is a constant multiple larger than the true model size s_0 is asymptotically vanishing. In other words, the SSGL posterior concentrates on sparse sets.

6.2 Sparse Generalized Additive Models (GAMs)

Assume there is a true model,

$$y_i = \sum_{j=1}^{p} f_{0j}(X_{ij}) + \varepsilon_i, \quad \varepsilon_i \sim \mathcal{N}(0, \sigma_0^2). \tag{6.9}$$

where $\sigma_0^2 \in (0, \infty)$. Throughout this section, we assume that all the covariates $\boldsymbol{X}_i = (X_{i1}, \ldots, X_{ip})^T$ have been standardized to lie in $[0,1]^p$ and that $f_{0j} \in \mathcal{C}^\kappa[0,1], j = 1, \ldots, p$. That is, the true functions are all at least κ-times continuously differentiable over $[0,1]$, for some $\kappa \in \mathbb{N}$. Suppose further that each f_{0j} can be approximated by a linear combination of basis functions $\{g_{j1}, \ldots, g_{jd}\}$. In matrix notation, (6.9) can then be written as

$$\boldsymbol{Y} = \sum_{j=1}^{p} \widetilde{\boldsymbol{X}}_j \boldsymbol{\beta}_{0j} + \boldsymbol{\delta} + \boldsymbol{\varepsilon}, \quad \boldsymbol{\varepsilon} \sim \mathcal{N}_n(\boldsymbol{0}, \sigma_0^2 \mathbf{I}_n), \tag{6.10}$$

where $\widetilde{\boldsymbol{X}}_j$ denotes an $n \times d$ matrix where the (i,k)th entry is $\widetilde{\boldsymbol{X}}_j(i,k) = g_{jk}(X_{ij})$, the $\boldsymbol{\beta}_{0j}$'s are $d \times 1$ vectors of basis coefficients, and $\boldsymbol{\delta}$ denotes an $n \times 1$ vector of lower-order bias.

Denote $\widetilde{\boldsymbol{X}} = [\widetilde{\boldsymbol{X}}_1, \ldots, \widetilde{\boldsymbol{X}}_p]$ and $\boldsymbol{\beta} = (\boldsymbol{\beta}_1^T, \ldots, \boldsymbol{\beta}_p^T)^T$. Under (6.9), suppose that we endow $(\boldsymbol{\beta}, \sigma^2)$ in (6.10) with the prior (6.2). We have the following assumptions:

(B1) Assume that $p \gg n$, $\log p = o(n)$, and $d \asymp n^{1/(2\kappa+1)}$.

(B2) The number of true nonzero functions satisfies $s_0 = o(\min\{n/\log p, n^{2\kappa/(2\kappa+1)}\})$.

(B3) There exists a constant $k_1 > 0$ so that for all n, $\lambda_{\max}(\widetilde{\boldsymbol{X}}^T \widetilde{\boldsymbol{X}}) \leq k_1 n$.

(B4) Let $\xi \subset \{1, \ldots, p\}$, and let $\widetilde{\boldsymbol{X}}_\xi$ denote the submatrix of $\widetilde{\boldsymbol{X}}$ that contains the submatrices indexed by ξ. There exists a constant $\nu_1 > 0$ and an integer $\bar{p}$ satisfying $s_0 = o(\bar{p})$ and $\bar{p} = o(s_0 \log n)$, so that $\lambda_{\min}(\widetilde{\boldsymbol{X}}_\xi^T \widetilde{\boldsymbol{X}}_\xi) \geq n\nu_1$ for any model of size $|\xi| \leq \bar{p}$.

(B5) $\|\boldsymbol{\beta}_0\|_\infty = O(\log p)$.

(B6) The bias $\boldsymbol{\delta}$ satisfies $\|\boldsymbol{\delta}\|_2 \lesssim \sqrt{s_0 n} d^{-\kappa}$.

Assumptions (B1)-(B5) are analogous to assumptions (A1)-(A5). Assumptions (B3)-(B4) are difficult to verify but can be shown to hold if appropriate basis functions for the g_{jk}'s are used, e.g. cubic B-splines (Yoo and Ghosal (2016), Wei et al. (2020)). Finally, Assumption (B6) bounds the approximation error incurred by truncating the basis expansions to be of size d. This assumption is satisfied, for example, by B-spline basis expansions (Zhou et al. (1998), Wei et al. (2020)).

Let $\widetilde{\mathbb{P}}_0$ denote the probability measure underlying the truth (6.9) and $\Pi(\cdot|\boldsymbol{Y})$ denote the posterior distribution under NPSSL model with the prior (6.2) for $(\boldsymbol{\beta}, \sigma^2)$ in (6.10). Further, let $f(\boldsymbol{X}_i) = \sum_{j=1}^p f_j(X_{ij})$ and $f_0(\boldsymbol{X}_i) = \sum_{j=1}^p f_{0j}(X_{ij})$, and define the empirical norm $\|\cdot\|_n$ as

$$\|f - f_0\|_n^2 = \frac{1}{n} \sum_{i=1}^n \left[f(\boldsymbol{X}_i) - f_0(\boldsymbol{X}_i) \right]^2 .$$

Let $\mathcal{F}$ denote the infinite-dimensional set of all possible additive functions $f = \sum_{j=1}^p f_j$, where each f_j can be represented by a d-dimensional basis expansion. In Raskutti et al. (2012), it was shown that the minimax estimation rate for $f_0 = \sum_{j=1}^p f_{0j}$ under squared ℓ_2 error loss is $\epsilon_n^2 \asymp s_0 \log p/n + s_0 n^{-2\kappa/(2\kappa+1)}$. The next theorem establishes that the NPSSL model achieves this minimax posterior contraction rate.

Theorem 4 (posterior contraction rates)**.** *Let* $\epsilon_n^2 = s_0 \log p/n + s_0 n^{-2\kappa/(2\kappa+1)}$*. Suppose that Assumptions (B1)-(B6) hold. Under model (6.10), suppose that we endow* $(\boldsymbol{\beta}, \sigma^2)$ *with the prior (6.2) (replacing* G *with* p*). For the hyperparameters in the* $\mathcal{B}(a, b)$ *prior on* θ*, we choose* $a = 1, b = p^c$*,* $c > 2$*. Further, we set* $\lambda_0 = (1-\theta)/\theta$ *and* $\lambda_1 \asymp 1/n$ *in the SSGL prior. Then*

$$\Pi\left(f \in \mathcal{F} : \|f - f_0\|_n \geq \widetilde{M}_1 \epsilon_n | \boldsymbol{Y}\right) \to 0 \text{ a.s. } \widetilde{\mathbb{P}}_0 \text{ as } n, p \to \infty, \tag{6.11}$$

$$\Pi\left(\sigma^2 : |\sigma^2 - \sigma_0^2| \geq 4\sigma_0^2 \epsilon_n | \boldsymbol{Y}\right) \to 0 \text{ as } n \to \infty, \text{ a.s. } \widetilde{\mathbb{P}}_0 \text{ as } n, p \to \infty, \tag{6.12}$$

for some $\widetilde{M}_1 > 0$*.*

Let the generalized dimensionality $|\boldsymbol{\gamma}(\boldsymbol{\beta})|$ be defined as before in (6.6) (replacing G with p), with ω_g from (6.7) as the threshold (replacing m_g with d). The next theorem shows that under the NPSSL, the expected posterior probability that the generalized dimension size is a constant multiple larger than the true model size s_0 asymptotically vanishes.

Theorem 5 (dimensionality)**.** *Suppose that the same conditions as those in Theorem 4 hold. Then under* (6.10)*, for sufficiently large* $\widetilde{M}_2 > 0$*,*

$$\sup_{\boldsymbol{\beta}_0} \widetilde{\mathbb{E}}_{\boldsymbol{\beta}_0} \Pi\left(\boldsymbol{\beta} : |\boldsymbol{\gamma}(\boldsymbol{\beta})| > \widetilde{M}_2 s_0 | \boldsymbol{Y}\right) \to 0 \text{ as } n, p \to \infty. \tag{6.13}$$

7 Simulation Studies

In this section, we will evaluate our method in a number of settings. For the SSGL approach, we fix $\lambda_1 = 1$ and use cross-validation to choose from $\lambda_0 \in \{1, 2, \ldots, 100\}$. For the prior $\theta \sim \mathcal{B}(a, b)$, we set $a = 1, b = G$ so that θ is small with high probability. We will compare our SSGL approach with the following methods:

1. GroupLasso: the group lasso (Yuan and Lin, 2006)

2. BSGS: Bayesian sparse group selection (Chen et al., 2016)

3. SoftBart: soft Bayesian additive regression tree (BART) (Linero and Yang, 2018)

4. RandomForest: random forests (Breiman, 2001)

5. SuperLearner: super learner (van der Laan et al., 2007)

6. GroupSpike: point-mass spike-and-slab priors (1.3) placed on groups of coefficients[1]

In our simulations, we will look at the mean squared error (MSE) for estimating $f(\boldsymbol{X}_{\text{new}})$ averaged over a new sample of data $\boldsymbol{X}_{\text{new}}$. We will also evaluate the variable selection properties of the different methods using precision and recall, where precision = TP/(TP + FP), recall = TP/(TP+FN), and TP, FP, and FN denote the number of true positives, false positives, and false negatives respectively. Note that we will not show precision or recall for the SuperLearner, which averages over different models and different variable selection procedures and therefore does not have one set of variables that are deemed significant.

7.1 Sparse Semiparametric Regression

Here, we will evaluate the use of our proposed SSGL procedure in sparse semiparametric regression with p continuous covariates. Namely, we implement the NPSSL main effects model described in Section 5.1. In Section B of the Supplementary Material, we include more simulation studies of the SSGL approach under both sparse and dense settings, as

[1]Code to implement GroupSpike is included in the Supplementary data. Due to the discontinuous prior, GroupSpike is not amenable to a MAP finding algorithm and has to be implemented using MCMC.

well as a simulation study showing that we are accurately estimating the residual variance σ^2.

We let $n = 100, p = 300$. We generate independent covariates from a standard uniform distribution, and we let the true regression surface take the following form:

$$\mathbb{E}(Y|\boldsymbol{X}) = 5\sin(\pi X_1) + 2.5(X_3^2 - 0.5) + e^{X_4} + 3X_5,$$

with variance $\sigma^2 = 1$.

To implement the SSGL approach, we estimate the mean response as

$$\mathbb{E}(\boldsymbol{Y}|\boldsymbol{X}) = \widetilde{\boldsymbol{X}}_1\boldsymbol{\beta}_1 + \cdots + \widetilde{\boldsymbol{X}}_p\boldsymbol{\beta}_p,$$

where $\widetilde{\boldsymbol{X}}_j$ is a design matrix of basis functions used to capture the possibly nonlinear effect of X_j on Y. For the basis functions in $\widetilde{\boldsymbol{X}}_j, j = 1, \ldots, p$, we use natural splines with degrees of freedom d chosen from $d \in \{2, 3, 4\}$ using cross-validation. Thus, we are estimating a total of between 600 and 1200 unknown basis coefficients.

We run 1000 simulations and average all of the metrics considered over each simulated data set. Figure 1 shows the results from this simulation study. The GroupSpike approach has the best performance in terms of MSE, followed closely by SSGL, with the next best approach being SoftBart. In terms of recall, the SSGL and GroupLasso approaches perform the best, indicating the highest power in detecting the significant groups. This comes with a loss of precision as the GroupSpike and SoftBart approaches have the best precision among all methods.

Although the GroupSpike method performed best in this scenario, the SSGL method was much faster. As we show in Section B.5 of the Supplementary Material, when $p = 4000$, fitting the SSGL model with a sufficiently large λ_0 takes around three seconds to run. This is almost 50 times faster than running 100 MCMC iterations of the GroupSpike method

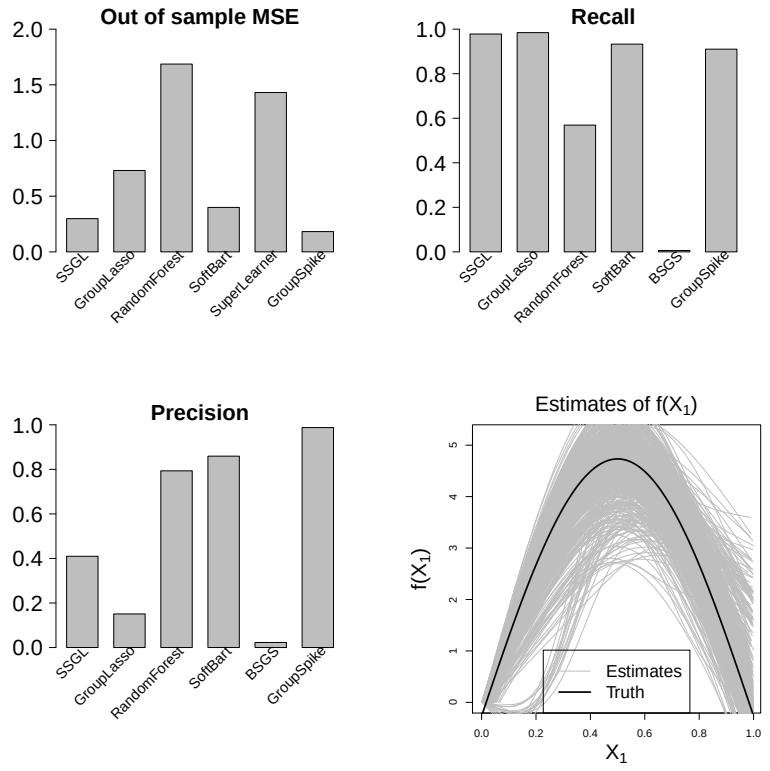

Figure 1: Simulation results for semiparametric regression. The top left panel presents the out-of-sample mean squared error, the top right panel shows the recall score to evaluate variable selection, the bottom left panel shows the precision score, and the bottom right panel shows the estimates from each simulation of $f_1(X_1)$ for SSGL. The MSE for BSGS is not displayed as it lies outside of the plot area.

(never mind the total time it takes for the GroupSpike model to converge). Our experiments demonstrate that the SSGL model gives comparable performance to the "theoretically ideal" point mass spike-and-slab in a fraction of the computational time.

7.2 Interaction Detection

We now explore the ability of the SSGL approach to identify important interaction terms in a nonparametric regression model. To this end, we implement the NPSSL model with interactions from Section 5.2. We generate 25 independent covariates from a standard uniform distribution with a sample size of 300. Data is generated from the model:

$$\mathbb{E}(Y|\boldsymbol{X}) = 2.5\sin(\pi X_1 X_2) + 2\cos(\pi(X_3 + X_5)) + 2(X_6 - 0.5) + 2.5X_7,$$

with variance $\sigma^2 = 1$. While this may not seem like a high-dimensional problem, we will consider all two-way interactions, and there are 300 such interactions. The important two-way interactions are between X_1 and X_2 and between X_3 and X_5. We evaluate the performance of each method and examine the ability of SSGL to identify important interactions while excluding all of the remaining interactions. Figure 2 shows the results for this simulation setting. The SSGL, GL, GroupSpike, and SoftBart approaches all perform well in terms of out-of-sample mean squared error, with GroupSpike slightly outperforming the competitors. The SSGL also does a very good job at identifying the two important interactions. The (X_1, X_2) interaction is included in 97% of simulations, while the (X_3, X_5) interaction is included 100% of the time. All other interactions are included in only a small fraction of simulated data sets.

8 Real Data Analysis

Here, we will illustrate the SSGL procedure in two distinct settings: 1) evaluating the SSGL's performance on a data set where $n = 120$ and $p = 15,000$, and 2) identifying important (nonlinear) main effects and interactions of environmental exposures. In Section

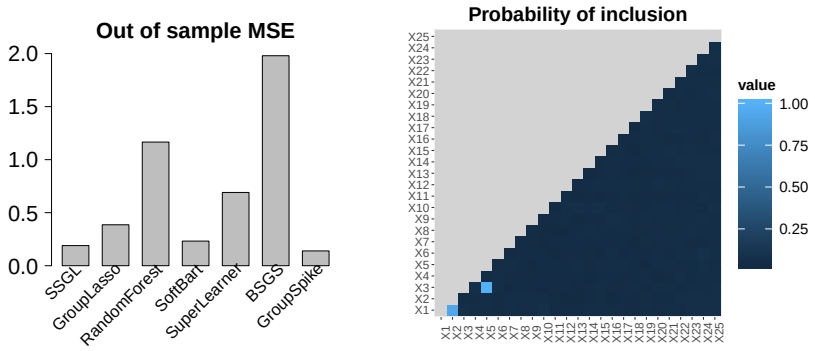

Figure 2: Simulation results from the interaction setting. The left panel shows out-of-sample MSE for each approach, while the right panel shows the probability of a two-way interaction being included into the SSGL model for all pairs of covariates.

C of the Supplementary Material, we evaluate the predictive performance of our approach on benchmark data sets where $p < n$, compared to several other state-of-the-other methods. Our results show that in both the $p \gg n$ and $p < n$ settings, the SSGL maintains good predictive accuracy.

8.1 Bardet-Biedl Syndrome Gene Expression Study

We now analyze a microarray data set consisting of gene expression measurements from the eye tissue of 120 laboratory rats[2]. The data was originally studied by Scheetz et al. (2006) to investigate mammalian eye disease, and later analyzed by Breheny and Huang (2015) to demonstrate the performance of their group variable selection algorithm. In this

[2]Data accessed from the Gene Expression Omnibus `www.ncbi.nlm.nih.gov/geo` (accession no. GSE5680).

	SSGL	Group Lasso
# groups selected	12	83
10-fold CV error	0.012 (0.003)	0.017 (0.008)

Table 2: Results for SSGL and Group Lasso on the Bardet-Biedl syndrome gene expression data set. In parentheses, we report the standard errors for the CV prediction error.

data, the goal is to identify genes which are associated with the gene TRIM32. TRIM32 has previously been shown to cause Bardet-Biedl syndrome (Chiang et al., 2006), a disease affecting multiple organs including the retina.

The original data consists of 31,099 probe sets. Following Breheny and Huang (2015), we included only the 5,000 probe sets with the largest variances in expression (on the log scale). For these probe sets, we considered a three-term natural cubic spline basis expansion, resulting in a grouped regression problem with $n = 120$ and $p = 15,000$. We implemented SSGL with regularization parameter values $\lambda_1 = 1$ and λ_0 ranging on an equally spaced grid from 1 to 500. We compared SSGL with the group lasso (Yuan and Lin, 2006), implemented using the R package `gglasso` (Yang and Zou, 2015).

As shown in Table 2, SSGL selected much fewer groups than the group lasso. Namely, SSGL selected 12 probe sets, while the group lasso selected 83 probe sets. Moreover, SSGL achieved a smaller 10-fold cross-validation error than the group lasso, albeit within range of random variability (Table 2). These results demonstrate that the SSGL achieves strong predictive accuracy, while *also* achieving the most parsimony. The probe IDs and gene symbols for the groups selected by both SSGL and the group lasso are displayed in Table 2 of Section C.2 of the Supplementary Material. Interestingly, only four of the 12 probes selected by SSGL were also selected by the group lasso.

We next conducted gene ontology enrichment analysis on the group of genes found by each of the methods using the R package `clusterProfiler` (Yu et al., 2012). This software determines whether subsets of genes known to act in a biological process are overrepresented in a group of genes, relative to chance. If such a subset is significant, the group of genes is said to be "enriched" for that biological process. With a false discovery rate of 0.01, SSGL had five enriched terms, while the group lasso had none. The terms for which SSGL was enriched included RNA binding, a biological process with which the response gene TRIM32 is associated.[3] These findings show the ability of SSGL to find biologically meaningful signal in the data. Additional details for our gene ontology enrichment analysis can be found in in Section C.2 of the Supplementary Material.

8.2 Environmental Exposures in the NHANES Data

Here, we analyze data from the 2001-2002 cycle of the National Health and Nutrition Examination Survey (NHANES), which was previously analyzed by Antonelli et al. (2019). We aim to identify which organic pollutants are associated with changes in leukocyte telomere length (LTL) levels. Telomeres are segments of DNA that help to protect chromosomes, and LTL levels are commonly used as a proxy for overall telomere length. LTL levels have previously been shown to be associated with adverse health effects (Haycock et al., 2014), and recent studies within the NHANES data have found that organic pollutants can be associated with telomere length (Mitro et al., 2015).

We use the SSGL approach to evaluate whether any of 18 organic pollutants are associated with LTL length and whether there are any significant interactions among the pollutants also associated with LTL length. In addition to the 18 exposures, there are 18

[3]https://www.genecards.org/cgi-bin/carddisp.pl?gene=TRIM32 (accessed 03/01/20)

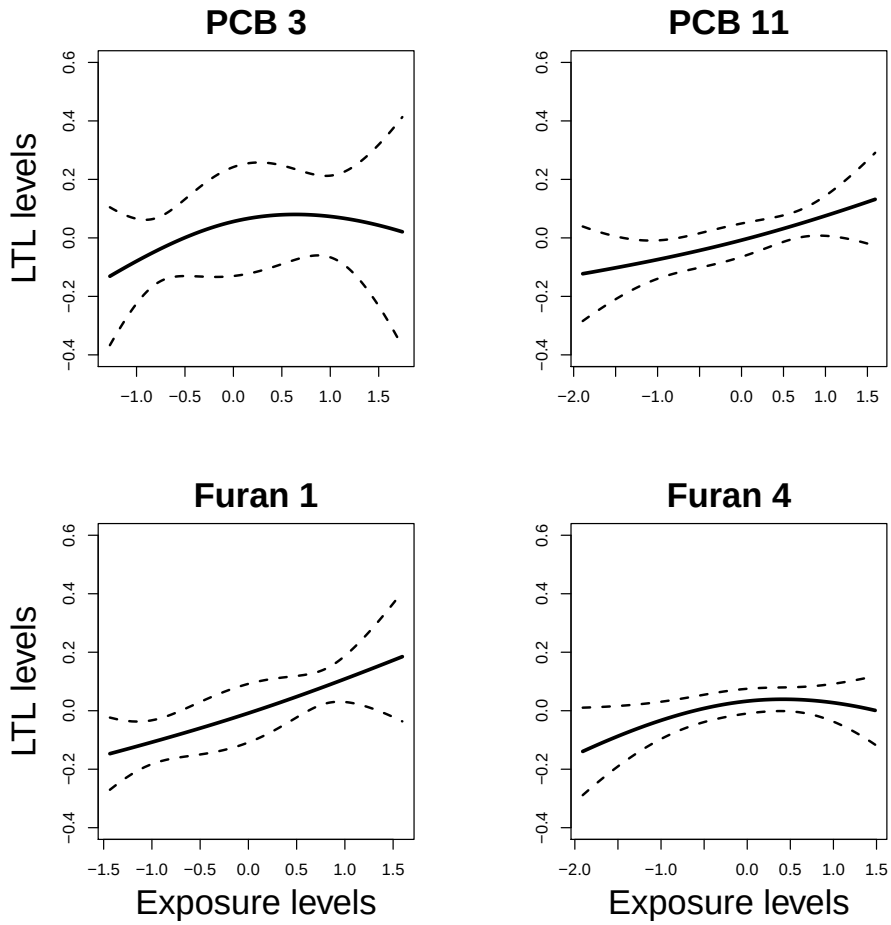

Figure 3: Exposure response curves for each of the four exposures with significant main effects identified by the model.

additional demographic variables which we adjust for in our model. We model the effects of the 18 exposures on LTL length using spline basis functions with two degrees of freedom. For the interaction terms, this leads to four terms for each pair of interactions, and we orthogonalize these terms with respect to the main effects. In total, this leads to a data

set with $n = 1003$ and $p = 666$.

Our model selects four significant main effects and six significant interaction terms. In particular, PCB 3, PCB 11, Furan 1, and Furan 4 are identified as the important main effects in the model. Figure 3 plots the exposure response curves for these exposures. We see that each of these four exposures has a positive association with LTL length, which agrees with results seen in Mitro et al. (2015) that saw positive relationships between persistent organic pollutants and telomere length. Further, our model identifies more main effects and more interactions than previous analyses of these data, e.g. Antonelli et al. (2019), which could lead to more targeted future research in understanding how these pollutants affect telomere length. Additional discussion and analysis of the NHANES data set can be found in Section C.3 of the Supplementary Material.

9 Discussion

We have introduced the spike-and-slab group lasso (SSGL) model for variable selection and linear regression with grouped variables. We also extended the SSGL model to generalized additive models with the nonparametric spike-and-slab lasso (NPSSL). The NPSSL can efficiently identify both nonlinear main effects *and* higher-order nonlinear interaction terms. Moreover, our prior performs an automatic multiplicity adjustment and self-adapts to the true sparsity pattern of the data through a *non*-separable penalty. For computation, we introduced highly efficient coordinate ascent algorithms for MAP estimation and employed de-biasing methods for uncertainty quantification. An R package implementing the SSGL model can be found at `https://github.com/jantonelli111/SSGL`.

Although our model performs group selection, it does so in an "all-in-all-out" manner,

similar to the original group lasso (Yuan and Lin, 2006). Future work will be to extend our model to perform both group selection and within-group selection of individual coordinates. We are currently working to extend the SSGL to perform bilevel selection.

We are also working to extend the nonparametric spike-and-slab lasso so it can adapt to even more flexible regression surfaces than the generalized additive model. Under the NPSSL model, we used cross-validation to tune a single value for the degrees of freedom. In reality, different functions can have vastly differing degrees of smoothness, and it will be desirable to model anisotropic regression surfaces while avoiding the computational burden of tuning the individual degrees of freedom over a p-dimensional grid.

Acknowledgments

Dr. Ray Bai, Dr. Gemma Moran, and Dr. Joseph Antonelli contributed equally and wrote this manuscript together, with input and suggestions from all other listed co-authors. The bulk of this work was done when the first listed author was a postdoc at the Perelman School of Medicine, University of Pennsylvania, under the mentorship of the last two authors. The authors are grateful to three anonymous reviewers, the Associate Editor, and the Editor whose thoughtful comments and suggestions helped to improve this manuscript. The authors would also like to thank Ruoyang Zhang, Peter Bühlmann, and Edward George for helpful discussions.

Funding

Ray Bai and Mary Boland were funded in part by generous funding from the Perelman School of Medicine, University of Pennsylvania. The work of Ray Bai and Yong Chen was supported in part by National Institutes of Health grants 1R01LM012607 (R.B., Y.C.) and 1R01AI130460 (Y.C.).

SUPPLEMENTARY MATERIAL

The Supplementary Material Section A contains the entire block-coordinate ascent algorithm and additional remarks on implementation of the SSGL model. Section B contains additional simulation studies. Section C contains additional data analysis on benchmark data sets where $p < n$ and additional discussion and analysis of the two data sets introduced in Section 8. Section D contains all the proofs for the theoretical results in this article. Code for reproducing the results in the simulation studies and data analysis is also available for download in the supplementary data.

References

Antonelli, J., M. Mazumdar, D. Bellinger, D. C. Christiani, R. Wright, and B. A. Coull (2019). Estimating the health effects of environmental mixtures using Bayesian semi-parametric regression and sparsity inducing priors. *The Annals of Applied Statistics (to appear)*.

Antonelli, J., G. Parmigiani, and F. Dominici (2019). High-dimensional confounding adjustment using continuous spike and slab priors. *Bayesian Analysis 14*(3), 805–828.

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

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

Breiman, L. (2001). Random forests. *Machine Learning 45*(1), 5–32.

Chen, R.-B., C.-H. Chu, S. Yuan, and Y. N. Wu (2016). Bayesian sparse group selection. *Journal of Computational and Graphical Statistics 25*(3), 665–683.

Chiang, A. P., J. S. Beck, H.-J. Yen, M. K. Tayeh, T. E. Scheetz, R. E. Swiderski, D. Y. Nishimura, T. A. Braun, K.-Y. A. Kim, J. Huang, et al. (2006). Homozygosity mapping with SNP arrays identifies TRIM32, an E3 ubiquitin ligase, as a Bardet–Biedl syndrome gene (BBS11). *Proceedings of the National Academy of Sciences 103*(16), 6287–6292.

Deshpande, S. K., V. Ročková, and E. I. George (2019). Simultaneous variable and covariance selection with the multivariate spike-and-slab lasso. *Journal of Computational and Graphical Statistics 28*(4), 921–931.

Friedman, J., T. Hastie, and R. Tibshirani (2010). Regularization paths for generalized linear models via coordinate descent. *Journal of Statistical Software 33*(1), 1–22.

Gan, L., N. N. Narisetty, and F. Liang (2019). Bayesian regularization for graphical models with unequal shrinkage. *Journal of the American Statistical Association 114*(527), 1218–1231.

Haycock, P. C., E. E. Heydon, S. Kaptoge, A. S. Butterworth, A. Thompson, and P. Willeit (2014). Leucocyte telomere length and risk of cardiovascular disease: systematic review and meta-analysis. *Bmj 349*, g4227.

Huang, J., P. Breheny, and S. Ma (2012). A selective review of group selection in high-dimensional models. *Statistical Science: A Review Journal of the Institute of Mathematical Statistics 27*(4).

Jacob, L., G. Obozinski, and J.-P. Vert (2009). Group lasso with overlap and graph lasso. In *Proceedings of the 26th Annual International Conference on Machine Learning*, ICML '09, New York, NY, USA, pp. 433–440. ACM.

Javanmard, A. and A. Montanari (2018). Debiasing the lasso: Optimal sample size for Gaussian designs. *The Annals of Statistics 46*(6A), 2593–2622.

Kyung, M., J. Gill, M. Ghosh, and G. Casella (2010). Penalized regression, standard errors, and Bayesian lassos. *Bayesian Analysis 5*(2), 369–411.

Li, Y., B. Nan, and J. Zhu (2015). Multivariate sparse group lasso for the multivariate multiple linear regression with an arbitrary group structure. *Biometrics 71*(2), 354–363.

Li, Z., T. Mccormick, and S. Clark (2019). Bayesian joint spike-and-slab graphical lasso. In K. Chaudhuri and R. Salakhutdinov (Eds.), *Proceedings of the 36th International Conference on Machine Learning*, Volume 97 of *Proceedings of Machine Learning Research*, Long Beach, California, USA, pp. 3877–3885. PMLR.

Lim, M. and T. Hastie (2015). Learning interactions via hierarchical group-lasso regularization. *Journal of Computational and Graphical Statistics 24*(3), 627–654.

Linero, A. R. and Y. Yang (2018). Bayesian regression tree ensembles that adapt to smoothness and sparsity. *Journal of the Royal Statistical Society: Series B (Statistical Methodology) 80*(5), 1087–1110.

Liquet, B., K. Mengersen, A. N. Pettitt, and M. Sutton (2017). Bayesian variable selection regression of multivariate responses for group data. *Bayesian Analysis 12*(4), 1039–1067.

Meinshausen, N. and P. Bühlmann (2006). High-dimensional graphs and variable selection with the lasso. *The Annals of Statistics 34*(3), 1436–1462.

Mendelson, S. and A. Pajor (2006). On singular values of matrices with independent rows. *Bernoulli 12*(5), 761–773.

Mitro, S. D., L. S. Birnbaum, B. L. Needham, and A. R. Zota (2015). Cross-sectional associations between exposure to persistent organic pollutants and leukocyte telomere length among US adults in NHANES, 2001–2002. *Environmental Health Perspectives 124*(5), 651–658.

Moran, G. E., V. Ročková, and E. I. George (2019). Spike-and-slab lasso biclustering. *preprint*.

Moran, G. E., V. Ročková, and E. I. George (2019). Variance prior forms for high-dimensional Bayesian variable selection. *Bayesian Analysis 14*(4), 1091–1119.

Ning, B., S. Jeong, and S. Ghosal (2019). Bayesian linear regression for multivariate responses under group sparsity. *Bernoulli (to appear)*.

Raskutti, G., M. J. Wainwright, and B. Yu (2010). Restricted eigenvalue properties for correlated gaussian designs. *Journal of Machine Learning Research 11*, 2241–2259.

Raskutti, G., M. J. Wainwright, and B. Yu (2012). Minimax-optimal rates for sparse additive models over kernel classes via convex programming. *Journal of Machine Learning Research 13*, 389–427.

Ravikumar, P., J. Lafferty, H. Liu, and L. Wasserman (2009). Sparse additive models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology) 71*(5), 1009–1030.

Ročková, V. and E. I. George (2016). Bayesian penalty mixing: The case of a non-separable penalty. In *Statistical Analysis for High-Dimensional Data: The Abel Symposium 2014*, Volume 11, pp. 233. Springer.

Ročková, V. (2018). Bayesian estimation of sparse signals with a continuous spike-and-slab prior. *The Annals of Statistics 46*(1), 401–437.

Ročková, V. and E. I. George (2016). Fast Bayesian factor analysis via automatic rotations to sparsity. *Journal of the American Statistical Association 111*(516), 1608–1622.

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

Scheetz, T. E., K.-Y. A. Kim, R. E. Swiderski, A. R. Philp, T. A. Braun, K. L. Knudtson, A. M. Dorrance, G. F. DiBona, J. Huang, T. L. Casavant, et al. (2006). Regulation of gene expression in the mammalian eye and its relevance to eye disease. *Proceedings of the National Academy of Sciences 103*(39), 14429–14434.

Scott, J. G. and J. O. Berger (2010). Bayes and empirical-Bayes multiplicity adjustment in the variable-selection problem. *The Annals of Statistics 38*(5), 2587–2619.

Simon, N., J. Friedman, T. Hastie, and R. Tibshirani (2013). A sparse-group lasso. *Journal of Computational and Graphical Statistics 22*(2), 231–245.

Tang, Z., Y. Shen, Y. Li, X. Zhang, J. Wen, C. Qian, W. Zhuang, X. Shi, and N. Yi (2018). 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., Y. Shen, X. Zhang, and N. Yi (2017a). The spike-and-slab lasso Cox model for survival prediction and associated genes detection. *Bioinformatics 33*(18), 2799–2807.

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

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society: Series B (Statistical Methodology) 58*, 267–288.

van de Geer, S., P. Bühlmann, Y. Ritov, and R. Dezeure (2014). On asymptotically optimal confidence regions and tests for high-dimensional models. *The Annals of Statistics 42*(3), 1166–1202.

van der Laan, M. J., E. C. Polley, and A. E. Hubbard (2007). Super learner. *Statistical Applications in Genetics and Molecular Biology 6*(1), 1544–6115.

Wei, R., B. J. Reich, J. A. Hoppin, and S. Ghosal (2020). Sparse Bayesian additive nonparametric regression with application to health effects of pesticides mixtures. *Statistica Sinica 30*, 55–79.

Xu, X. and M. Ghosh (2015). Bayesian variable selection and estimation for group lasso. *Bayesian Analysis 10*(4), 909–936.

Yang, X. and N. N. Narisetty (2019). Consistent group selection with Bayesian high dimensional modeling. *Bayesian Analysis (to appear)*.

Yang, Y. and H. Zou (2015). A fast unified algorithm for solving group-lasso penalize learning problems. *Statistics and Computing 25*(6), 1129–1141.

Yoo, W. W. and S. Ghosal (2016). Supremum norm posterior contraction and credible sets for nonparametric multivariate regression. *The Annals of Statistics 44*(3), 1069–1102.

Yu, G., L.-G. Wang, Y. Han, and Q.-Y. He (2012). clusterprofiler: an r package for comparing biological themes among gene clusters. *OMICS: A Journal of Integrative Biology 16*(5), 284–287.

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

Zhou, S., X. Shen, and D. Wolfe (1998). Local asymptotics for regression splines and confidence regions. *The Annals of Statistics 26*(5), 1760–1782.