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

Ray Bai*1, Gemma E. Moran*1, Joseph L. Antonelli*1, Yong Chen1, and Mary R. Boland1

1Department of Statistics, University of South Carolina, Columbia, SC; 1Data Science Institute, Columbia University, New York, NY; 1Department of Statistics, University of Florida, Gainesville, FL; 1Department of Biostatistics, Epidemiology, and Informatics, University of Pennsylvania, Philadelphia, PA

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 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 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. Supplementary materials for this article are available online.

1. Introduction

1.1. Regression With Grouped Variables

Group structure arises in many statistical applications. For example, in multifactor analysis of variance, multilevel 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 $Y_{n \times 1}$ can be modeled as a linear regression problem with $G$ groups:

$$Y = \sum_{g=1}^{G} X_{g} \beta_{g} + \epsilon,$$

where $\epsilon \sim N_{n}(0, \sigma^{2} I_{n})$, $\beta_{g}$ is a coefficient vector of length $m_{g}$, and $X_{g}$ is an $n \times m_{g}$ covariate matrix corresponding to group $g = 1, \ldots, G$. Even in the absence of grouping information about the covariates, the model (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 non-linear 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,

$$\arg \min_{\beta} \frac{1}{2} \|Y - \sum_{g=1}^{G} X_{g} \beta_{g}\|_{2}^{2} + \lambda \sqrt{m_{g}} \|\beta_{g}\|_{2},$$

where $\| \cdot \|_{2}$ is the $\ell_{2}$ norm. In the frequentist literature, many variants of model (2) have been introduced, which use some combination of $\ell_{1}$ and $\ell_{2}$ penalties on the coefficients of interest (e.g., Jacob, Obozinski, and Vert 2009; Simon et al. 2013; Li, Nan, and Zhu 2015).

In the Bayesian framework, selection of relevant groups under model (1) is often done by placing spike-and-slab priors on each of the groups $\beta_{g}$ (e.g., Xu and Ghosh 2015; Lique et al. 2017; Ning, Jeong, and Ghosal 2019; Yang and Narisetty 2019). These priors typically take the form,

$$\pi(\beta | \gamma) = \prod_{g=1}^{G} [(1 - \gamma_{g}) \delta_{0}(\beta_{g}) + \gamma_{g} \pi(\beta_{g})],$$

$$\pi(\gamma | \theta) = \prod_{g=1}^{G} \theta_{g}^{\gamma_{g}} (1 - \theta)^{1-\gamma_{g}},$$

$$\theta \sim \pi(\theta),$$

ARTICLE HISTORY
Received March 2019
Accepted April 2020

KEYWORDS
High-dimensional regression; Interaction detection; Maximum a posteriori estimation; Nonparametric regression; Spike-and-slab lasso; Variable selection

CONTACT Ray Bai RBAI@mailbox.sc.edu Department of Statistics, University of South Carolina, Columbia, SC 29208.

© 2020 American Statistical Association
where \( \gamma \) is a binary vector that indexes the \( 2^G \) possible models, \( \theta \in (0, 1) \) is the mixing proportion, \( \delta_0 \) is a point mass at \( 0_{m_G} \in \mathbb{R}^{m_G} \) (the "spike"), and \( \pi(\beta|\gamma) \) is an appropriate "slab" density (typically a multivariate normal distribution or a scale-mixture multivariate normal density). With a well-chosen prior on \( \theta \), 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 (3) are interpretable, but they are computationally intractable in high dimensions, due in large part to the combinatorial complexity of updating the discrete indicators \( \gamma \). As an alternative, fully continuous variants of spike-and-slab models have been developed. For continuous spike-and-slab models, the point mass spike \( \delta_0 \) is replaced by a continuous density heavily concentrated around \( 0_{m_G} \). 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 \( \beta_j \), that is,

\[
\pi(\beta|\theta) = \prod_{j=1}^{p} [(1-\theta)\psi(\beta_j|\lambda_0) + \theta \psi(\beta_j|\lambda_1)].
\]

(4)

where \( \theta \in (0, 1) \) is the mixing proportion and \( \psi(\cdot|\lambda) \) denotes a univariate Laplace density indexed by hyperparameter \( \lambda \), that is, \( \psi(\beta|\lambda) = \frac{1}{2\lambda} e^{-|\beta|/\lambda} \). Typically, we set \( \lambda_0 \gg \lambda_1 \) so that the spike is heavily concentrated about zero. Unlike (3), the SSL model (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 SSL 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 2016b; Moran, Ročková, and George 2019a), multivariate regression (Deshpande, Ročková, and George 2019), covariance precision matrix estimation (Deshpande, Ročková, and George 2019; Gan, Narisetty, and Liang 2019; Li, Mccormick, and Clark 2019), causal inference (Antonelli, Parmigiani, and Dominici 2019), generalized linear models (GLMs) (Tang et al. 2017b, 2018), and Cox proportional hazards models (Tang et al. 2017a).

While the SSL (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) used the univariate SSL prior (4) for grouped data where each group had a group-specific sparsity inducing parameter, \( \theta_g \), instead of a single \( \theta \) 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 article, 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 SSL methodology for situations where the linear model is too inflexible, we extend the SSL to sparse GAMs by introducing the nonparametric spike-and-slab lasso (NPSSL). To our knowledge, our work is the first to apply the SSL 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 \( \beta \) and the unknown variance \( \sigma^2 \) under the SSLG prior.

The rest of the article is structured as follows. In Section 2, we introduce the 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 use ideas from the de-biased lasso to perform inference on the SSGL model. In Section 5, we extend the SSLG to nonparametric settings by proposing the 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 datasets.

### 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 \( a_n \ll b_n \). 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 \( v \in \mathbb{R}^p \), we let \( ||v||_1 := \sum_{i=1}^{p} |v_i| \), \( ||v||_2 := \left( \sum_{i=1}^{p} v_i^2 \right)^{1/2} \), and \( ||v||_\infty := \max_{1 \leq i \leq p} |v_i| \) denote the \( \ell_1, \ell_2, \) and \( \ell_\infty \) norms, respectively. For a symmetric matrix \( A \), we let \( \lambda_{\text{min}}(A) \) and \( \lambda_{\text{max}}(A) \) denote its minimum and maximum eigenvalues.
2. The Spike-and-Slab Group Lasso

Let \( \beta_g \) denote a real-valued vector of length \( m_g \). We define the group lasso density as

\[
\Psi(\beta_g|\lambda) = C_g \lambda^{m_g} \exp(-\lambda \|\beta_g\|_2),
\]

(5)

where \( C_g = 2^{-m_g} \pi^{-(m_g-1)/2} [\Gamma((m_g+1)/2)]^{-1} \). 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). Kyung et al. (2010) considered a single prior (5) on each of the \( \beta_g \)'s, while Xu and Ghosh (2015) employed (5) as the slab in the point-mass mixture (3). These authors implemented their models using Markov chain Monte Carlo (MCMC).

In this article, we introduce a continuous spike-and-slab prior with the group lasso density (5) 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 \( \hat{\beta} = (\beta_1^T, \ldots, \beta_G^T)^T \) under model (1), the SSGL is defined as

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

(6)

where \( \Psi(\cdot|\lambda) \) denotes the group lasso density (5) indexed by hyperparameter \( \lambda \), and \( \theta \in (0,1) \) is a mixing proportion. \( \lambda_0 \) corresponds to the spike which shrinks the entire vector \( \beta_g \) toward \( \mathbf{0}_{m_g} \), while \( \lambda_1 \) corresponds to the slab. For shorthand notation, we denote \( \Psi(\beta_g|\lambda_0) \) as \( \Psi_0(\beta_g) \) and \( \Psi(\beta_g|\lambda_1) \) as \( \Psi_1(\beta_g) \) going forward.

Under the grouped regression model (1), we place the SSGL prior (6) on \( \hat{\beta} \). In accordance with the recommendations of Moran, Ročková, and George (2019b), we do not scale our prior by the unknown \( \sigma \). Instead, we place an independent Jeffreys' prior on \( \sigma^2 \), that is,

\[
\pi(\sigma^2) \propto \sigma^{-2}.
\]

(7)

The mixing proportion \( \theta \) in (6) 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, that is, \( p = \sum_{g=1}^G m_g \). Our goal is to find the MAP estimates of the regression coefficients \( \beta \in \mathbb{R}^p \). This optimization problem is equivalent to a penalized likelihood method in which the logarithm of the prior (6) 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 \( \theta \), the prior expected proportion of groups with large coefficients. Ultimately, we will pursue a fully Bayes approach and place a prior on \( \theta \), 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 \( \theta \) is fixed, echoing the development of Ročková and George (2018). In this situation, the regression coefficients \( \beta_g \) are conditionally independent a priori, resulting in a separable SSGL penalty. Later we will consider the fully Bayes approach, which will yield the nonseparable SSGL penalty.

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

\[
\text{pen}_g(\beta|\theta) = \log \left[ \frac{\pi(\beta|\theta)}{\pi(\mathbf{0}_g|\theta)} \right],
\]

(8)

where

\[
p_g^0(\beta_g) = \frac{\theta \Psi_1(\beta_g)}{\theta \Psi_1(\beta_g) + (1-\theta) \Psi_0(\beta_g)}.
\]

(9)

The separable SSGL penalty is almost the logarithm of the original prior (6); the only modification is an additive constant to ensure that \( \text{pen}_g(\mathbf{0}_g|\theta) = 0 \). 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

\[
\frac{\partial \text{pen}_g(\beta|\theta)}{\partial \|\beta_g\|_2} = -\lambda_g^*(\beta_g),
\]

(10)

where

\[
\lambda_g^*(\beta_g) = \lambda_1 p_g^0(\beta_g) + \lambda_0 [1 - p_g^0(\beta_g)].
\]

(11)

Similarly to the SSL, the SSGL penalty is a weighted average of the two regularization parameters, \( \lambda_1 \) and \( \lambda_0 \). The weight \( p_g^0(\beta_g) \) is the conditional probability that \( \beta_g \) 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 (8) 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, Breheny, and Ma (2012), we assume that within each group, covariates are orthonormal, that is, \( X_g^T X_g = n \mathbf{1}_{m_g} \) for \( g = 1, \ldots, G \). If this assumption does not hold, then the \( X_g \) 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 \( \hat{\beta} = (\hat{\beta}_1^T, \ldots, \hat{\beta}_G^T)^T \) to be a global mode are

\[
X_g^T (Y - X \hat{\beta}) = \sigma^2 \lambda_0^* (\hat{\beta}_g) \frac{\beta_g}{\| \beta_g \|_2} \quad \text{for} \quad \beta_g \neq 0_{m_g},
\]

\[
\| X_g^T (Y - X \hat{\beta}) \|_2 \leq \sigma^2 \lambda_0^* (\hat{\beta}_g) \quad \text{for} \quad \beta_g = 0_{m_g}.
\]

Equivalently,

\[
\hat{\beta}_g = \frac{1}{n} \left( 1 - \frac{\sigma^2 \lambda_0^* (\hat{\beta}_g)}{\| \beta_g \|_2} \right) z_g,
\]

where \( z_g = X_g^T \left[ Y - \sum_{l \neq g} X_l \hat{\beta}_l \right] \).

**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 \( g \)-th group, keeping all other groups fixed.

**Proposition 2.** The global mode \( \hat{\beta}_g = 0_{m_g} \) if and only if \( \| z_g \|_2 \leq \Delta \), where

\[
\Delta = \inf_{\beta_g} \left\{ \frac{n \| \beta_g \|_2}{2} - \frac{\sigma^2 \text{pen}_{\text{NS}} (\beta_0)}{\| \beta_g \|_2} \right\}.
\]

The proof for Proposition 2 can be found in Section D.2 of the supplementary materials. Unfortunately, the threshold \( \Delta \) is difficult to compute. We instead find an approximation to this threshold. An upper bound is simply that of the soft-threshold solution (14), with \( \Delta \leq \sigma^2 \lambda_0^* (\beta_g) \). However, when \( \lambda_0 \) is large, this bound may be improved. Similarly to Roˇcková and George (2018), we provide improved bounds on the threshold in Theorem 1. This result requires the function \( h : \mathbb{R}^{m_g} \rightarrow \mathbb{R} \)

\[
h (\beta_g) = [\lambda_0^* (\beta_g) - \lambda_1] \frac{2}{\sigma^2} \log p_0^* (\beta_g).
\]

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

\[
\Delta_L < \Delta < \Delta_U,
\]

where

\[
\Delta_L = \sqrt{2n \sigma^2 \log \left[ 1 / p_0^* (0_{m_g}) \right]} - \sigma^2 d + \sigma^2 \lambda_1,
\]

\[
\Delta_U = \sqrt{2n \sigma^2 \log \left[ 1 / p_0^* (0_{m_g}) \right]} + \sigma^2 \lambda_1,
\]

and

\[
0 < d < \frac{2n}{\sigma^2} - \left( \frac{n}{\sigma^2 (\lambda_0 - \lambda_1)} - \sqrt{2n} / \sigma \right)^2.
\]

When \( \lambda_0 \) is large, \( d \rightarrow 0 \) 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 \( \theta \), the proportion of groups with nonzero coefficients. We now outline this fully Bayes strategy and the resulting nonseparable SSGL penalty. With the inclusion of the prior \( \theta \sim \pi (\theta) \), the marginal prior for the regression coefficients has the following form:

\[
\pi (\beta) = \int_0^1 \prod_{g=1}^G [\theta \Psi_1 (\beta_g) + (1 - \theta) \Psi_0 (\beta_g)] d\pi (\theta) (20)
\]

\[
\pi (\beta) = \int_0^1 \prod_{g=1}^G C_g \lambda_1^m_p \left[ \frac{\theta^G}{\Psi_1 (\beta_g)} \right] e^{-\lambda_1 \sum_{g=1}^G \| \beta_g \|_2} \int_0^1 \theta^G \prod_{g=1}^G p_0^* (\beta_g) d\pi (\theta),
\]

The nonseparable SSGL penalty is then defined similarly to the separable penalty, where again we have centered the penalty to ensure \( \text{pen}_{\text{NS}} (0_p) = 0 \).

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

\[
\text{pen}_{\text{NS}} (\beta) = \log \left[ \frac{\pi (\beta)}{\pi (0_p)} \right]
\]

\[
= -\lambda_1 \sum_{g=1}^G \| \beta_g \|_2 + \log \left[ \frac{\int_0^1 \theta^G \prod_{g=1}^G p_0^* (\beta_g) d\pi (\theta)}{\int_0^1 \theta^G \prod_{g=1}^G p_0^* (0_{m_g}) d\pi (\theta)} \right].
\]

Although the penalty (21) appears intractable, intuition is again obtained by considering the derivative. Following the same line of argument as Roˇcková and George (2018), the derivative of (21) is given in the following lemma.

**Lemma 2.**

\[
\frac{\partial \text{pen}_{\text{NS}} (\beta)}{\partial \| \beta_g \|_2} \equiv \lambda^* (\beta_g; \beta_\setminus g),
\]

where

\[
\lambda^* (\beta_g; \beta_\setminus g) \equiv p^* (\beta_g; \beta_\setminus g) \lambda_1 + [1 - p^* (\beta_g; \beta_\setminus g)] \lambda_0.
\]

\[
\lambda^* (\beta_g; \beta_\setminus g) \equiv p^*_2 (\beta_g), \quad \text{with} \quad \theta_g = \mathbb{E} (\theta | \beta_g).
\]

That is, the marginal prior from (21) 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 \( \theta \), we will impute the mean of \( \theta \) conditioned on the remaining regression coefficients.

We now consider the form of the conditional mean, \( \mathbb{E} (\theta | \beta_\setminus g) \). As noted by Roˇcková and George (2018), when the number of groups is large, this conditional mean can be replaced by \( \mathbb{E} (\theta | \hat{\beta}_g) \); we will proceed with the same approximation. For the prior
on $\theta$, 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 $\pi(\theta|\tilde{\beta})$. Suppose that the number of groups with nonzero coefficients is $\tilde{q}$, and assume without loss of generality that the first $\tilde{q}$ groups have nonzero coefficients. Then,

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

with $z = 1 - \frac{1}{\lambda_0}$ and $x_g = (1 - \frac{1}{\lambda_0}z_0)/2(\lambda_0 - \lambda_1)$. Similarly to Ročková and George (2018), this distribution is a generalization of the Gauss hypergeometric distribution. Consequently, the expectation may be written as

$$E[\theta|\tilde{\beta}] = \frac{\int_{\theta^a(1 - \theta)^{b-1}(1 - \theta z)G^{-\tilde{q}} \prod_{g=1}^{\tilde{q}}(1 - \theta x_g)d\theta}{\int_{0}^{1}\theta^a(1 - \theta)^{b-1}(1 - \theta z)G^{-\tilde{q}} \prod_{g=1}^{\tilde{q}}(1 - \theta x_g)d\theta}.$$  

(27)

While the above expression (27) appears laborious to compute, it admits a much simpler form when $\lambda_0$ is very large. Using a slight modification to the arguments of Ročková and George (2016a), we obtain this simpler form in Lemma 3.

Lemma 3. Assume $\pi(\theta|\tilde{\beta})$ is distributed according to (26). Let $\tilde{q}$ be the number of groups with nonzero coefficients. Then as $\lambda_0 \to \infty$,

$$E[\theta|\tilde{\beta}] = \frac{a + \tilde{q}}{a + b + G}.$$  

(28)

The proof for Lemma 3 is in Section D.2 of the supplementary materials. We note that the expression (28) is essentially the usual posterior mean of $\theta$ under a beta prior. Intuitively, as $\lambda_0$ diverges, the weights $\rho^0_g(\beta_g)$ concentrate at zero and one, yielding the familiar form for $E[\theta|\tilde{\beta}]$. With this in hand, we are now in a position to outline the block coordinate ascent algorithm for the NS-SSGL.

### 3.3. Optimization

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

$$\hat{\beta}_g = \frac{1}{n} \left(1 - \frac{\sigma^2 \lambda^*_{\hat{\beta}_g}}{\|z_g\|_2^2}\right) z_g + \frac{\sigma^2}{n} \left[Y - \sum_{g \neq g} X_g \hat{\beta}_g \right],$$

where $z_g = X_g^T \left[Y - \sum_{g \neq g} X_g \hat{\beta}_g \right]$ and $\hat{\theta}$ is the mean (28), conditioned on the previous value of $\beta$. As before, (29) is sufficient for a local mode, but not the global mode. When $p \gg n$ and $\lambda_0$ 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 $E[\theta|\tilde{\beta}_g]$ with $E[\theta|\tilde{\beta}]$, we do not require group-specific thresholds. Instead, we can use the threshold given in Proposition 2 and Theorem 1 where $\theta$ is replaced with the current update (28). In particular, we shall use the upper bound $\Delta^U$ in our block coordinate ascent algorithm.

Similarly to Ročková and George (2018), we combine the refined threshold, $\Delta^U$ with the soft thresholding operation (29), to yield the following update for $\beta_g$ at iteration $k$:

$$\beta_g^{(k)} = \frac{1}{n} \left(1 - \frac{\sigma^2 \lambda^*_{\beta_g^{(k-1)}}}{\|z_g\|_2^2}\right) z_g + \frac{\sigma^2}{n} \left[||z_g||_2 > \Delta^U\right],$$

(30)

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

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

$$\sigma^2 = \frac{\|Y - X\hat{\beta}\|_2^2}{n + 2}.$$  

(31)

The complete optimization algorithm is given in Algorithm 1 of Section A.1 of the supplementary materials. The computational complexity of this algorithm is $O(np)$ per iteration, where $p = \sum_{g=1}^G m_g$. It takes $O(nmg)$ operations to compute the partial residual $z_g$ for the $g$th group, for a total cost of $O(n\sum_{g=1}^G m_g) = O(np)$. Similarly, it takes $O(np)$ cost to compute the sum of squared residuals $\|Y - X\hat{\beta}\|_2^2$ to update the variance parameter $\sigma^2$. The computational complexity of our algorithm matches that of the usual gradient descent algorithms for lasso and group lasso (Friedman, Hastie, and Tibshirani 2010).

As a nonconvex 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 materials) enable SSGL to eliminate a number of suboptimal local modes from consideration in a similar manner to Ročková and George (2018). To briefly summarize the initialization strategy, we tune $\lambda_0$ from an increasing sequence of values, and we further scale $\lambda_0$ by $\sqrt{m_g}$ for each $g$th group to ensure that the amount of penalization is on the same scale for groups of potentially different sizes (Huang, Breheny, and Ma 2012). Meanwhile, we keep $\lambda_1$ fixed at a small value so that selected groups have minimal shrinkage. See Section A.2 of the supplementary materials for detailed discussion of choosing $(\lambda_0, \lambda_1)$.

### 4. Approaches to Inference

While the above procedure allows us to find the posterior mode of $\hat{\beta}$, 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 $\hat{\Sigma} = X^TX/n$ and let $\hat{\Theta}$ be an approximate inverse of $\hat{\Sigma}$. We define

$$\hat{\beta}_d = \hat{\beta} + \hat{\Theta}X^T(Y - X\hat{\beta})/n,$$  

(32)

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

$$\sqrt{n}(\hat{\beta}_d - \beta) \sim N(0, \sigma^2\hat{\Sigma} \hat{\Theta}^T).$$  

(33)

For our inference procedure, we replace the population variance $\sigma^2$ in (33) with the modal estimate $\hat{\sigma}^2$ from the SSGL model. To estimate $\hat{\Theta}$, we use the nodewise regression approach developed in Meinsmaun and Buhlmann (2006) and van de Geer et al. (2014). We describe this estimation procedure for $\hat{\Theta}$ in Section A.3 of the supplementary materials.

Let $\hat{\beta}_{dj}$ denote the $j$th coordinate of $\hat{\beta}_d$. We have from (33) that the 100(1 - $\alpha$)% asymptotic pointwise confidence intervals for $\beta_j, j = 1, \ldots, p$, are

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

(34)

where $c(\alpha, n, \hat{\sigma}^2) := \Phi^{-1}(1 - \alpha / 2)\sqrt{\hat{\sigma}^2(\hat{\Theta}\hat{\Theta}^T)_{jj}/n}$ and $\Phi(\cdot)$ denotes the cdf of $N(0, 1)$. 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 (34) 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 0 and an AR(1) covariance structure with correlation $\rho$. The two covariates from each group are the linear and squared term from the original covariates. We set the first seven elements of $\beta$ equal to (0, 0.5, 0.25, 0.1, 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, that is, covariates with a nonzero corresponding $\beta_j$, 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.

| $\rho$ | 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 |

5. Nonparametric Spike-and-Slab Lasso

We now introduce the 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:

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

(35)

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 $B_j = \{g_{j1}, \ldots, g_{jd}\}$, that is,

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

(36)

where $\beta_j = (\beta_{j1}, \ldots, \beta_{jd})^T$ are the unknown weights. Let $\tilde{X}_j$ denote the $n \times d$ matrix with the $(i, k)$th entry $\tilde{X}_j(i, k) = g_{jk}(X_{ij})$. Then, (35) may be represented in matrix form as

$$Y - \delta = \sum_{j=1}^p \tilde{X}_j\beta_j + \epsilon, \quad \epsilon \sim N_n(0, \sigma^2 I_n),$$  

(37)

where $\delta$ is a vector of the lower order truncation bias. Note that we assume the response $Y$ has been centered and so we do not include a grand mean $\mu$ in (37). Thus, we do not require the main effects to integrate to zero as in Wei et al. (2020). We do, however, require the matrices $\tilde{X}_j, j = 1, \ldots, p$, 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 orthonormalizing the $\tilde{X}_j$ matrices before fitting the model.

We assume that $Y$ depends on only a small number of the $p$ covariates so that many of the $f_j$’s have a negligible contribution to (35). This is equivalent to assuming that most of the weight vectors $\beta_j$ have all zero elements. If the $j$th covariate is determined to be predictive of $Y$, then $f_j$ has a nonnegligible contribution to (35). 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, $\tilde{X}_j, j = 1, \ldots, p$. We will utilize the NS-SSGL penalty developed in Section 3.2 to enforce this group-sparsity behavior in the model (37). More specifically, we seek to maximize the objective function with respect to $\beta = (\beta_1^T, \ldots, \beta_p^T)^T \in \mathbb{R}^{pd}$ and $\sigma^2$:

$$L(\beta, \sigma^2) = -\frac{1}{2\sigma^2}||Y - \sum_{j=1}^p \tilde{X}_j\beta_j||_2^2 - (n + 2)\log\sigma + \text{pen}_{NS}(\beta).$$  

(38)
To find the estimators of $\beta$ and $\sigma^2$, we use Algorithm 1 in Section A.1 of the supplementary materials. 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 $\sigma^2$ to be unknown, unlike Ravikumar et al. (2009). Accurate estimates of $\sigma^2$ 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) use MCMC. Finally, our SSSG 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 (35) 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:

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

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:

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

where $\beta_{kl} = (\beta_{kl1}, \ldots, \beta_{kl, d^*}, \beta_{kld^*, 1}, \ldots, \beta_{kld^*, d^*})^T \in \mathbb{R}^{d^2}$ is the vector of unknown weights. We let $\tilde{X}_{kl}$ denote the $n \times d^2$ matrix with rows

$$ \tilde{X}_{kl}(i, :) = \text{vec}(g_k(X_{ik}) g_l(X_{il})^T), $$

where $g_k(X_{ik}) = (g_{k1}(X_{ik}), \ldots, g_{kd_k}(X_{ik}))^T$. Then, (39) may be represented in matrix form as

$$ Y - \delta = \sum_{j=1}^{p} \tilde{X}_j \beta_j + \sum_{k=1}^{p} \sum_{l=k+1}^{p} \tilde{X}_{kl} \beta_{kl} + \varepsilon, \quad \varepsilon \sim N_n(0, \sigma^2 I_n), $$

where $\delta$ is a vector of the lower order truncation bias. We again assume $Y$ has been centered and so do not include a grand mean in (41). We do not constrain $\tilde{f}_{kl}$ 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 $\tilde{X}_l$ and $\tilde{X}_k$ to be orthogonal to the “interaction” matrix $\tilde{X}_{kl}$. 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 $\tilde{X}_k \circ \tilde{X}_l$ on $\tilde{X}_k$ and $\tilde{X}_l$.

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), that is, the “interaction” matrices in (41) would be $\tilde{X}_{kl}^{\text{aug}} = [\tilde{X}_k, \tilde{X}_l, \tilde{X}_{kl}]$, instead of simply $\tilde{X}_{kl}$. 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 $\beta_{kl}^{\text{aug}} = 0$). 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 $\beta_{kl}$ in the model (41) if there is a nonnegligible interaction between the $k$th and $l$th covariates, or we estimate $\beta_{kl} = 0$ if such an interaction is negligible. With the NS-SSGL penalty, the objective function is

$$ L(\beta, \sigma^2) = \frac{1}{2\sigma^2} \| Y - \sum_{j=1}^{p} \tilde{X}_j \beta_j \|_2^2 + \text{pen}_{\text{NS}}(\beta) + (n + 2) \log \sigma, $$

where $\beta = (\beta_1^T, \ldots, \beta_p^T, \beta_{p+1}^T, \ldots, \beta_{p(p-1)/2}^T)^T \in \mathbb{R}^{p^2/2}$. We can again use Algorithm 1 in Section A.1 of the supplementary materials to find the modal estimates of $\beta$ and $\sigma^2$.

6. Asymptotic Theory for the SSSG and NPSSL

In this section, we derive asymptotic properties for the separable SSSG 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 $\beta$ and the unknown $\sigma^2$, whereas Ročková and George (2018) proved their result only for $\beta$, 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 $\epsilon_n$ for the full posterior only. Our rate $\epsilon_n$ satisfies $\epsilon_n \to 0$ as $n \to \infty$ (i.e., the full posterior collapses to the true $(\beta, \sigma^2)$ almost surely as $n \to \infty$), and hence, it automatically follows that the posterior mode is a consistent estimator of $(\beta, \sigma^2)$. 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 materials.
6.1. Grouped Linear Regression

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

\[ Y = \sum_{g=1}^{G} X_g \beta_{0g} + \varepsilon, \quad \varepsilon \sim N_n(0, \sigma_0 I_n), \tag{42} \]

where \( \beta_0 = (\beta_{01}, \ldots, \beta_{0G})^T \) and \( \sigma_0^2 \in (0, \infty) \). Denote \( X = [X_1, \ldots, X_G] \) and \( \beta = (\beta_1^T, \ldots, \beta_G^T)^T \). Suppose we endow \( (\beta, \sigma^2) \) under model (42) with the following prior:

\[
\begin{align*}
\pi(\beta, \sigma^2) &\sim \prod_{g=1}^{G} \left[ (1-\theta) \Psi(\beta_g | \lambda_{0} + \theta \Psi(\beta_g | \lambda_1) \right], \\
\theta &\sim B(a, b), \\
\sigma^2 &\sim IG(c_0, d_0),
\end{align*}
\tag{43}
\]

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

Remark 1. In our implementation of the SSGL model, we endowed \( \sigma^2 \) with an improper prior, \( \pi(\sigma^2) \propto \sigma^{-2} \). This can be viewed as a limiting case of the \( IG(c_0, d_0) \) 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 \( (\beta, \sigma^2) \) to be proper.

6.1.1. Posterior Contraction Rates

Let \( m_{\max} = \max_{1 \leq g \leq G} m_g \) 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 \( \lambda_{\max}(X^T X) \leq kn^p \), for some \( \alpha \in [1, \infty) \).

(A4) Let \( \xi \in \{ 1, \ldots, G \} \) and let \( X_\xi \) denote the submatrix of \( X \) that contains the submatrices with groups indexed by \( \xi \). There exist constants \( v_1 > 0, v_2 > 0 \), and an integer \( \hat{p} \) satisfying \( s_0 = o(\hat{p}) \) and \( \hat{p} = o(s_0 \log n) \), so that \( n v_1 \leq \lambda_{\min}(X_\xi^T X_\xi) \leq \lambda_{\max}(X_\xi^T X_\xi) \leq n v_2 \) for any model of size \( |\xi| \leq \hat{p} \).

(A5) \( \|\beta_0\|_2 = O(\log G) \).

Assumption 6.1.1 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 6.1.1 specifies the growth rate for the true model size \( s_0 \). Assumption 6.1.1 bounds the eigenvalues of \( X^T X \) from above and is less stringent than requiring all the eigenvalues of the Gram matrix \( (X^T X) / n \) to be bounded away from infinity. Assumption 6.1.1 ensures that \( X^T X \) is locally invertible over sparse sets. In general, conditions 6.1.1 and 6.1.1 are difficult to verify, but they can be shown to hold with high probability for certain classes of matrices where the rows of \( X \) are independent and sub-Gaussian (Mendelson and Pajor 2006; Raskutti, Wainwright, and Yu 2010). Finally, Assumption 6.1.1 places a restriction on the growth rate of the maximum signal size for the true \( \beta_0 \).

We now state our main theorem on the posterior contraction rates for the SSGL prior (43) under model (42). Let \( \mathbb{P}_0 \) denote the probability measure underlying the truth (42) and \( \Pi(\cdot | Y) \) denote the posterior distribution under the prior (43) for \( (\beta, \sigma^2) \).

**Theorem 2 (Posterior contraction rates).** Let \( \epsilon_n = \sqrt{s_0 \log G / n} \), and suppose that Assumptions 6.1.1–6.1.1 hold. Under model (42), suppose that we endow \( (\beta, \sigma^2) \) with the prior (43). For the hyperparameters in the \( B(a, b) \) prior on \( \theta \), we choose \( a = 1, b = G^r, c > 2 \). Further, we set \( \lambda_0 = (1 - \theta) / \theta \) and \( \lambda_1 \asymp 1 / n \) in the SSGL prior. Then

\[
\begin{align*}
\Pi(\beta : \|\beta - \beta_0\|_2 \geq M_1 \sigma_0 \epsilon_n | Y) &\to 0 \text{ a.s. } \mathbb{P}_0 \text{ as } n \to \infty, \\
\Pi(\beta : \|X \beta - X \beta_0\|_2 \geq M_2 \sigma_0 \sqrt{\epsilon_n} | Y) &\to 0 \text{ a.s. } \mathbb{P}_0 \text{ as } n \to \infty, \\
\Pi(\sigma^2 : \sigma^2 - \sigma_0^2 \geq 4M_0^2 \epsilon_n | Y) &\to 0 \text{ as } n \to \infty, \text{ a.s. } \mathbb{P}_0 \text{ as } n \to \infty,
\end{align*}
\tag{46}
\]

for some \( M_1 > 0, M_2 > 0 \).

Remark 2. In the case where \( G = p = m_1 = \cdots = m_G = 1 \), the \( \epsilon \) and prediction error rates in (44) and (45) reduce to the familiar optimal rates of \( \sqrt{s_0 \log p / n} \) and \( \sqrt{s_0 \log p} \), respectively.

Remark 3. Equation (46) demonstrates that our model also consistently estimates the unknown variance \( \sigma^2 \), therefore providing further theoretical justification for placing an independent prior on \( \sigma^2 \), as advocated by Moran, Roˇcková, and George (2019b).

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 \( \omega_0 > 0 \), we define the generalized inclusion indicator and generalized dimensionality, respectively, as

\[
\gamma_{\omega_0}(\beta_g) = I(\|\beta_g\|_2 > \omega_0) \text{ and } |\gamma(\beta)| = \sum_{g=1}^{G} \gamma_{\omega_0}(\beta_g). \tag{47}
\]

In contrast to Bhattacharya et al. (2015) and Roˇcková and George (2018), we allow the threshold \( \omega_0 \) 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 \( \omega_0 \)'s, \( g = 1, \ldots, G \), should still tend toward zero as \( n \) increases, so that \( |\gamma(\beta)| \) provides a good approximation to \( \#(g : \beta_g \neq 0_{m_g}) \).

Consider as the threshold,

\[
\omega_0 \equiv \omega_0(\lambda_0, \lambda_1, \theta) = \frac{1}{\lambda_0 - \lambda_1} \log \left[ \frac{1 - \theta \lambda_0}{\theta} \frac{m_g}{\lambda_1} \right]. \tag{48}
\]

Note that for large \( \lambda_0 \), this threshold rapidly approaches zero. Analogous to Roˇcková (2018) and Roˇcková and George (2018),
any vectors $\beta_k$ that satisfy $\|\beta_k\|_2 = \omega_k$ correspond to the intersection points between the two group lasso densities in the separable SSGL prior (6), or when the second derivative $\partial^2 \text{pen}_s(\beta(\theta))/\partial \beta_k^2 = 0.5$. The value $\omega_k$ represents the turning point where the slab has dominated the spike, and thus, the sharper the spike (when $\lambda_0$ is large), the smaller the threshold.

Using the notion of generalized dimensionality (47) with (48) as the threshold, we have the following theorem.

**Theorem 3 (Dimensionality).** Suppose that the same conditions as those in Theorem 2 hold. Then under (42), for sufficiently large $M_3 > 0$,

$$\sup_{\beta_0} \mathbb{E}_{\beta_0} \mathbb{P}(\beta : |y(\beta)| > M_3 \sigma_0 |Y|) \to 0 \text{ as } n, G \to \infty.$$  

(49)

Theorem 3 shows that the expected posterior probability that the generalized dimension is a constant multiple larger than the true model size $\sigma_0$ 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,

$$y_i = \sum_{j=1}^{p} f_{ij}(X_{ij}) + \epsilon_i, \quad \epsilon_i \sim \mathcal{N}(0, \sigma_i^2),$$  

(50)

where $\sigma_i^2 \in (0, \infty)$. Throughout this section, we assume that all the covariates $X_i = (X_{i1}, \ldots, X_{ip})^T$ have been standardized to lie in $[0, 1]^p$ and that $f_{ij} \in C^k[0,1], j = 1, \ldots, p$. That is, the true functions are all at least $k$-times continuously differentiable over $[0, 1]$, for some $k \in \mathbb{N}$. Suppose further that each $f_{ij}$ can be approximated by a linear combination of basis functions $\{g_{i1}, \ldots, g_{id}\}$. In matrix notation, (50) can then be written as

$$Y = \sum_{j=1}^{p} \tilde{X}_j \beta_{ij} + \delta + \epsilon, \quad \epsilon \sim \mathcal{N}(0, \sigma_0^2 I_n),$$  

(51)

where $\tilde{X}_j$ denotes an $n \times d$ matrix where the $(i,k)$th entry is $\tilde{X}_{ik} = g_{ik}(X_{ij})$, the $\beta_{ij}$’s are $d \times 1$ vectors of basis coefficients, and $\delta$ denotes an $n \times 1$ vector of lower order bias.

Denote $\tilde{X} = [\tilde{X}_1, \ldots, \tilde{X}_p]$ and $\beta = (\beta_{i1}^T, \ldots, \beta_{ip}^T)^T$. Under (50), suppose that we endow $(\beta, \sigma^2)$ in (51) with the prior (43). We have the following assumptions:

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

(B2) The number of true nonzero functions satisfies $s_0 = o_{\text{max}}(n/\log p, n^{2/(2k+1)})$.

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

(B4) Let $\xi \subset \{1, \ldots, p\}$, and let $\tilde{X}_\xi$ denote the submatrix of $\tilde{X}$ that contains the submatrices indexed by $\xi$. There exists a constant $v_1 > 0$ and an integer $\tilde{p}$ satisfying $s_0 = o(\tilde{p})$ and $\tilde{p} = o(s_0 \log n)$, so that $\lambda_{\text{min}}(\tilde{X}_\xi^T \tilde{X}_\xi) \geq n v_1$ for any model of size $|\xi| \leq \tilde{p}$.

(B5) $\|\beta_0\|_\infty = O(\log p)$.

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

Assumptions 6.2–6.2 are analogous to assumptions 6.1.1–6.1.1. Assumptions 6.2 and 6.2 are difficult to verify but can be shown to hold if appropriate basis functions for the $g_{ik}$’s are used, for example, cubic B-splines (Yoo and Ghosal 2016; Wei et al. 2020). Finally, Assumption 6.2 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, Shen, and Wolfe 1998; Wei et al. 2020).

Let $\tilde{P}_0$ denote the probability measure underlying the truth (50) and $\Pi(\cdot | Y)$ denote the posterior distribution under NPSSL with the prior (43) for $(\beta, \sigma^2)$ in (51). Further, let $f(X_i) = \sum_{j=1}^{p} f_j(X_{ij})$ and $f_0(X_i) = \sum_{j=1}^{p} f_{0j}(X_{ij})$, and define the empirical norm $\|f\|_n$ as

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

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, Wainwright, and Yu (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 $\sigma_0^2 \asymp \log p/n + s_0 n^{-2k/(2k+1)}$. The next theorem establishes that the NPSSL model achieves this minimax posterior contraction rate.

**Theorem 4 (Posterior contraction rates).** Suppose that Assumptions 6.2–6.2 hold. Under model (51), suppose that we endow $(\beta, \sigma^2)$ with the prior (43) (replacing $G$ with $p$). For the hyperparameters in the $B(a,b)$ prior on $\theta$, we choose $a = 1, b = p', c > 2$. Further, we set $\lambda_0 = (1 - \theta)/\theta$ and $\lambda_1 \asymp 1/n$ in the SSGL prior. Then

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

(52)

$$\mathbb{P} \left( \sigma^2 : |\sigma^2 - \sigma_0^2| \geq \tilde{M}_2 \epsilon_n \right) \to 0 \text{ as } n \to \infty, \text{ a.s. as } \tilde{P}_0 \text{ as } n, p \to \infty,$$  

(53)

for some $\tilde{M}_1 > 0$.

Let the generalized dimensionality $|y(\beta)|$ be defined as before in (47) (replacing $G$ with $p$), with $\omega_\delta$ from (48) as the threshold (replacing $m_\delta$ 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 $\sigma_0$ asymptotically vanishes.

**Theorem 5 (Dimensionality).** Suppose that the same conditions as those in Theorem 4 hold. Then under (51), for sufficiently large $M_2 > 0$,

$$\sup_{\beta_0} \mathbb{E}_{\beta_0} \mathbb{P}(\beta : |y(\beta)| > M_2 \sigma_0 |Y|) \to 0 \text{ as } n, p \to \infty.$$  

(54)

**7. Simulation Studies**

In this section, we will evaluate our method in a number of settings. For the SSGL approach, we fix $\lambda_0 = 1$ and use cross-validation to choose from $\lambda_0 \in \{1, 2, \ldots, 100\}$. For the prior $\theta \sim B(a,b)$, we set $a = 1, b = G$ so that $\theta$ 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, Polley, and Hubbard 2007)
6. GroupSpike: point-mass spike-and-slab priors (3) placed on groups of coefficients.

In our simulations, we will look at the mean squared error (MSE) for estimating \( f(X) \) averaged over a new sample of data \( X \). 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 materials, we include more simulation studies of the SSGL approach under both sparse and dense settings, as well as a simulation study showing that we are accurately estimating the residual variance \( \sigma^2 \).

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:

\[
E(Y|X) = 5\sin(\pi X_1) + 2.5(X_2^2 - 0.5) + e^{X_4} + 3X_5,
\]

with variance \( \sigma^2 = 1 \).

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

\[
E(Y|X) = \tilde{X}_1\beta_1 + \cdots + \tilde{X}_p\beta_p,
\]

where \( \tilde{X}_j \) is a design matrix of basis functions used to capture the possibly nonlinear effect of \( X_j \) on \( Y \). For the basis functions in \( \tilde{X}_j, j = 1, \ldots, p \), 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 dataset. 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.

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 \( f(X_1) \) for SSGL. The MSE for BSGS is not displayed as it lies outside of the plot area.

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 materials, when \( p = 4000 \), fitting the SSGL model with a sufficiently large \( \lambda_0 \) takes around three seconds to run. This is almost 50 times faster than running 100 MCMC iterations of the GroupSpike method (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:

\[
E(Y|X) = 2.5\sin(\pi X_1 X_2) + 2\cos(\pi (X_2 + X_3)) + 2(X_6 - 0.5) + 2.5X_7,
\]

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...
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.

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 datasets.

8. Real Data Analysis

Here, we will illustrate the SSGL procedure in two distinct settings: (1) evaluating the SSGL’s performance on a dataset where \(n = 120\) and \(p = 15,000\), and (2) identifying important (nonlinear) main effects and interactions of environmental exposures. In Section C of the supplementary materials, we evaluate the predictive performance of our approach on benchmark datasets 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 dataset consisting of gene expression measurements from the eye tissue of 120 laboratory rats. 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 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 5000 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 \(\lambda_0\) 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 materials. 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. 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 materials.

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. (2020). We aim to

**Table 2.** Results for SSGL and group lasso on the Bardet–Biedl syndrome gene expression dataset.

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

NOTE: In parentheses, we report the standard errors for the CV prediction error.
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 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 dataset 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, for example, Antonelli et al. (2020), which could lead to more targeted future research in understanding how these pollutants affect telomere length. Additional discussion and analysis of the NHANES dataset can be found in Section C.3 of the supplementary materials.

9. Discussion

We have introduced the SSGL model for variable selection and linear regression with grouped variables. We also extended the SSGL model to GAMs with the 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 nonseparable penalty. For computation, we introduced highly efficient coordinate ascent algorithms for MAP estimation and employed debiasing 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 GAM. 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.

Supplementary Materials

The supplementary materials 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 datasets where $p < n$ and additional discussion and analysis of the two datasets 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.

Acknowledgments

Dr. Ray Bai, Dr. Gemma Moran, and Dr. Joseph Antonelli contributed equally and wrote this article 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 article. 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.).

References

Antonelli, J., Mazumdar, M., Bellinger, D., Christiani, D. C., Wright, R., and Coull, B. A. (2020), “Estimating the Health Effects of Environmental Mixtures Using Bayesian Semiparametric Regression and Sparsity Inducing Priors,” The Annals of Applied Statistics, 14, 257–275. [194,195]
Antonelli, J., Parmigiani, G., and Dominici, F. (2019), “High-Dimensional Confounding Adjustment Using Continuous Spike and Slab Priors,” Bayesian Analysis, 14, 805–828. [185]
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, 1479–1490. [191]
Breheyn, P., and Huang, J. (2015), “Group Descent Algorithms for Non-convex Penalized Linear and Logistic Regression Models With Grouped Predictors,” Statistics and Computing, 25, 173–187. [187,194]
Breiman, L. (2001), "Random Forests," Machine Learning, 45, 5–32. [193]
Chen, R.-B., Chu, C.-H., Yuan, S., and Wu, Y. N. (2016), "Bayesian Sparse Group Selection," Journal of Computational and Graphical Statistics, 25, 665–683. [193]
Chiang, A. P., Beck, J. S., Yen, H.-J., Tayeh, M. K., Scheetz, T. E., Swiderski, R. E., Nishimura, D. Y., Braun, T. A., Kim, K.-Y. A., Huang, J., and Elbedour, K. (2006), “Homozygosity Mapping With SNP Arrays Identities TRIM32, an E3 Ubiquitin Ligase, as a Bardet–Biedl Syndrome Gene (BBS11),” Proceedings of the National Academy of Sciences of the United States of America, 103, 6287–6292. [194]
Deshpande, S. K., Ročková, V., and George, E. I. (2019), “Simultaneous Variable and Covariance Selection With the Multivariate Spike-and-Slab Lasso,” Journal of Computational and Graphical Statistics, 28, 921–931. [185]
Friedman, J., Hastie, T., and Tibshirani, R. (2010), “Regularization Paths for Generalized Linear Models via Coordinate Descent,” Journal of Statistical Software, 33, 1–22. [188]
Gan, L., Narisetty, N. N., and Liang, F. (2019), “Bayesian Regularization for Graphical Models With Unequal Shrinkage,” Journal of the American Statistical Association, 114, 1218–1231. [185]
Haycock, P. C., Heydon, E. E., Kaptoge, S., Butterworth, A. S., Thompson, A. A., and Willeit, P. (2014), “Leucocyte Telomere Length and Risk of Cardiovascular Disease: Systematic Review and Meta-Analysis,” BMJ, 349, g4227. [195]
Huang, J., Breheyn, P., and Ma, S. (2012), “A Selective Review of Group Selection in High-Dimensional Models,” Statistical Science, 27, 481–499. [186,188]
Jacob, L., Obozinski, G., and Vert, J.-P. (2009), “Group Lasso With Overlap and Slab Lasso,” in Proceedings of the 26th Annual International Conference on Machine Learning, ICML ’09, ACM, New York, NY, USA, pp. 433–440. [184]
Javanmard, A., and Montanari, A. (2018), “Debiasing the Lasso: Optimal Sample Size for Gaussian Designs,” The Annals of Statistics, 46, 2593–2622. [189]
Kyung, M., Gill, J., Ghosh, M., and Casella, G. (2010), “Penalized Regression, Standard Errors, and Bayesian Lassos,” Bayesian Analysis, 5, 369–411. [186]
Li, Y., Nan, B., and Zhu, J. (2015), “Multivariate Sparse Group Lasso for the Multivariate Multiple Linear Regression With An Arbitrary Group Structure,” Biometrika, 71, 354–363. [184]
Li, Z., Mccormick, T., and Clark, S. (2019), “Bayesian Joint Spike-and-Slab Graphical Lasso,” in Proceedings of the 36th International Conference on Machine Learning, Proceedings of Machine Learning Research, PMLR, Long Beach, CA, USA (Vol. 97), eds. K. Chaudhuri and R. Salakhutdinov, pp. 3877–3885. [185]
Lim, M., and Hastie, T. (2015), “Learning Interactions via Hierarchical Group-Lasso Regularization,” Journal of Computational and Graphical Statistics, 24, 627–654. [190]
Linero, A. R., and Yang, Y. (2018), “Bayesian Regression Tree Ensembles That Adapt to Smoothness and Sparsity,” Journal of the Royal Statistical Society, Series B, 80, 1087–1110. [193]
Liquet, B., Mengersen, K., Pettitt, A. N., and Sutton, M. (2017), “Bayesian Variable Selection Regression of Multivariate Responses for Group Data,” Bayesian Analysis, 12, 1039–1067. [184]
Meinshausen, N., and Bühlmann, P. (2006), “High-Dimensional Graphs and Variable Selection With the Lasso,” The Annals of Statistics, 34, 1436–1462. [189]
Mendelson, S., and Pajor, A. (2006), “On Singular Values of Matrices With Independent Rows,” Bernoulli, 12, 761–773. [191]
Mitro, S. D., Birnbaum, L. S., Needham, B. L., and Zota, A. R. (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, 651–658. [195]
Moran, G. E., Ročková, V., and George, E. I. (2019a), “Splice-and-Slab Lasso Biclustering,” Preprint. [185]
Moran, G. E., Ročková, V., and George, E. I. (2019b), “Variance Prior Forms for High-Dimensional Bayesian Variable Selection,” Bayesian Analysis, 14, 1091–1119. [186,191]
Ning, B., Jeong, S., and Ghosal, S. (2019), “Bayesian Linear Regression for Multivariate Responses Under Group Sparsity,” Bernoulli (to appear). [184]
Raskutti, G., Wainwright, M. J., and Yu, B. (2010), "Restricted Eigenvalue Properties for Correlated Gaussian Designs," *Journal of Machine Learning Research*, 11, 2241–2259. [191]

(2012), "Minimax-Optimal Rates for Sparse Additive Models Over Kernel Classes via Convex Programming," *Journal of Machine Learning Research*, 13, 389–427. [192]

Ravikumar, P., Lafferty, J., Liu, H., and Wasserman, L. (2009), "Sparse Additive Models," *Journal of the Royal Statistical Society*, Series B, 71, 1009–1030. [189,190]

Roˇcková, V. (2018), "Bayesian Estimation of Sparse Signals With a Continuous Spike-and-Slab Prior," *The Annals of Statistics*, 46, 401–437. [185,191]

Roˇcková, V., and George, E. I. (2016a), "Bayesian Penalty Mixing: The Case of a Non-Separable Penalty," in *Statistical Analysis for High-Dimensional Data: The Abel Symposium 2014* (Vol. 11), Springer, p. 233. [188]

(2016b), "Fast Bayesian Factor Analysis via Automatic Rotations to Sparsity," *Journal of the American Statistical Association*, 111, 1608–1622. [185]

(2018), "The Spike-and-Slab Lasso," *Journal of the American Statistical Association*, 113, 431–444. [185,186,187,188,190,191]

Scheetz, T. E., Kim, K.-Y. A., Swiderski, R. E., Philp, A. R., Braun, T. A., Knudtson, K. L., Dorrance, A. M., DiBona, G. F., Huang, J., Casavant, T. L., and Sheffield, V. C. (2006), "Regulation of Gene Expression in the Mammalian Eye and Its Relevance to Eye Disease," *Proceedings of the National Academy of Sciences of the United States of America*, 103, 14429–14434. [194]

Scott, J. G., and Berger, J. O. (2010), "Bayes and Empirical-Bayes Multiplicity Adjustment in the Variable-Selection Problem," *The Annals of Statistics*, 38, 2587–2619. [186,188]

Simon, N., Friedman, J., Hastie, T., and Tibshirani, R. (2013), "A Sparse-Group Lasso," *Journal of Computational and Graphical Statistics*, 22, 231–245. [184]

Tang, Z., Shen, Y., Li, Y., Zhang, X., Wen, J., Qian, C., Zhuang, W., Shi, X., and Yi, N. (2018), "Group Spike-and-Slab Lasso Generalized Linear Models for Disease Prediction and Associated Genes Detection by Incorporating Pathway Information," *Bioinformatics*, 34, 901–910. [185]

Tang, Z., Shen, Y., Zhang, X., and Yi, N. (2017a), "The Spike-and-Slab Lasso Cox Model for Survival Prediction and Associated Genes Detection," *Bioinformatics*, 33, 2799–2807. [185]

(2017b), "The Spike-and-Slab Lasso Generalized Linear Models for Prediction and Associated Genes Detection," *Genetics*, 205, 77–88. [185]

Tibshirani, R. (1996), "Regression Shrinkage and Selection via the Lasso," *Journal of the Royal Statistical Society*, Series B, 58, 267–288. [185]

van de Geer, S., Bühlmann, P., Ritov, Y., and Dezeure, R. (2014), "On Asymptotically Optimal Confidence Regions and Tests for High-Dimensional Models," *The Annals of Statistics*, 42, 1166–1202. [188,189]

van der Laan, M. J., Polley, E. C., and Hubbard, A. E. (2007), "Super Learner," *Statistical Applications in Genetics and Molecular Biology*, 6, 1544–6115. [193]

Wei, R., Reich, B. J., Hoppin, J. A., and Ghosal, S. (2020), "Sparse Bayesian Additive Nonparametric Regression With Application to Health Effects of Pesticides Mixtures," *Statistica Sinica*, 30, 55–79. [189,190,192]

Xu, X., and Ghosh, M. (2015), "Bayesian Variable Selection and Estimation for Group Lasso," *Bayesian Analysis*, 10, 909–936. [184,186]

Yang, X., and Narisetty, N. N. (2019), "Consistent Group Selection With Bayesian High Dimensional Modeling." *Bayesian Analysis* (to appear). [184]

Yang, Y., and Zou, H. (2015), "A Fast Unified Algorithm for Solving Group-Lasso Penalize Learning Problems," *Statistics and Computing*, 25, 1129–1141. [194]

Yoo, W. W., and Ghosal, S. (2016), "Supremum Norm Posterior Contraction and Credible Sets for Nonparametric Multivariate Regression," *The Annals of Statistics*, 44, 1069–1102. [192]

Yu, G., Wang, L.-G., Han, Y., and He, Q.-Y. (2012), "clusterProfiler: An R Package for Comparing Biological Themes Among Gene Clusters," *OMICS*, 16, 284–287. [194]

Yuan, M., and Lin, Y. (2006), "Model Selection and Estimation in Regression With Grouped Variables," *Journal of the Royal Statistical Society*, Series B, 68, 49–67. [184,185,186,193,194,196]

Zhou, S., Shen, X., and Wolfe, D. (1998), "Local Asymptotics for Regression Splines and Confidence Regions," *The Annals of Statistics*, 26, 1760–1782. [192]