Statistica Sinica Preprint No: SS-2025-0071	
Title	Bayesian High-Dimensional Grouped-Regression Using Sparse Projection-posterior
Manuscript ID	SS-2025-0071
URL	http://www.stat.sinica.edu.tw/statistica/
DOI	10.5705/ss.202025.0071
Complete List of Authors	Samhita Pal and Subhashis Ghosal
Corresponding Authors	Samhita Pal
E-mails	samhitapal3896@gmail.com
Notice: Accepted author version.	

Statistica Sinica: Newly accepted Paper

Bayesian High-dimensional Grouped-regression using Sparse Projection-posterior

Samhita Pal and Subhashis Ghosal

North Carolina State University

Abstract: **We present a novel Bayesian approach for high-dimensional grouped regression under sparsity. We leverage a sparse projection method that uses a sparsity-inducing map to induce a posterior on a lower-dimensional parameter space. Our method introduces three distinct projection maps based on popular penalty functions: the Group LASSO projection-posterior, the Group SCAD projection-posterior, and the Adaptive Group LASSO projection-posterior. Each projection map is constructed to immerse posterior samples into a structured, sparse space, allowing for effective group selection and estimation in high-dimensional settings. We derive optimal posterior contraction rates for estimation and prediction, thereby proving that the methods are model-selection consistent. We also propose a Debiased Group LASSO Projection Map that ensures correct asymptotic coverage of credible sets. Our methodology is particularly suited for applications in nonparametric additive models, where we use B-spline expansions to capture complex relationships between covariates and the response. Extensive simulations validate our theoretical findings, demonstrating the robustness of our approach across different settings. Finally, we illustrate the practical utility of our method with an application to brain MRI volume data from the Alzheimer's Disease Neuroimaging Initiative (ADNI), where our model identifies key brain regions associated with Alzheimer's Disease severity.**

Key words and phrases: Grouped-regression; Sparse; High-dimension; Coverage; Penalty; Projection.

1. Introduction

Grouped sparse regression is a powerful tool in statistical modeling that provides a structured approach for selecting and estimating subsets of variables in high-dimensional datasets. In scenarios where very high-dimensional predictors naturally group, such as genes in genomics or pixels in imaging, grouped sparse regression techniques help select relevant groups, thereby improving predictive accuracy. The group-selection approach is also instrumental for high-dimensional additive models by using basis expansions or by representing multilevel categorical variables as a set of dummy variables. There can be another approach in which, along with a subset of groups, it is further assumed that only a minor portion of each group's constituent variables is active, leading to within-group sparsity. However, we do not consider this setup in the current paper.

Assuming we have prior information on the group structure among the covariates, it is interesting to examine the effects of the most relevant groups on the response variable. This model assumes that the underlying sparsity pattern lies within the groups, while keeping all variables within each group. The pioneering work in this direction, the group LASSO, was developed by Yuan and Lin [2006]. This was later extended to the sparse additive models by Ravikumar et al. [2009]. A very broad class of penalties, specifically, a composite absolute penalty for group selection, was developed by Zhao et al. [2009], of which the group LASSO is a special case. Following these initial studies, several other penalization methods of grouped regression under sparsity were proposed in the literature, including but not limited to the group SCAD [Zeng and Xie, 2014], group MCP [Huang et al., 2009], and the adaptive group LASSO [Wang and Leng, 2008]. A complete review of frequentist properties of various group selection and estimation methods can be found in Huang et al. [2012].

In the Bayesian literature, the grouped regression method was first studied by Farcomeni [2010], who extended the SSVS method of George and McCulloch [1995] to the group-sparse setting. Similarly, Raman et al. [2009] extended the Bayesian LASSO [Park and Casella, 2008] to the group setting and proposed the Bayesian group LASSO. Casella et al. [2010] also suggested a multivariate Laplacian prior for the group coefficients along the same line. Xu and Ghosh [2015] proposed the Bayesian group lasso model with spike and slab priors and demonstrated the oracle property of the posterior median estimator for group variable selection and estimation under orthogonal designs. Xu et al. [2016] proposed the Bayesian grouped horseshoe regression, which extends the Bayesian horseshoe estimator to handle grouped variables through a grouped Bayesian horseshoe and a hierarchical Bayesian grouped horseshoe, demonstrating their effectiveness in additive models with both simulated and real data. Yang and Narisetty [2020] proposed a Bayesian hierarchical model with a spike-and-slab prior to perform group selection in high-dimensional linear regression, providing theoretical consistency results. Lai and Chen [2021] thoroughly reviewed and compared Bayesian approaches to group selection under the linear regression model.

When the variables in a high-dimensional linear regression are not individually meaningful but naturally form groups, a statistical method that correctly identifies these groups and accurately estimates the regression coefficients is proper. Our paper solves this problem using a novel Bayesian projection technique and provides an in-depth study of its theoretical properties. The approach used in this paper follows the concept of mapping non-sparse samples from the full posterior obtained by putting a convenient conjugate prior on the vector of regression coefficients. Here, we have proposed three maps. They closely resemble minimizers of the objective functions for the group LASSO, group SCAD, and adaptive group LASSO

procedures. This idea, also called 'immersion posterior' in the literature, has previously been studied in the context of sparse high-dimensional linear regression in Pal and Ghoshal [2024] and in several other contexts in Bhaumik and Ghosal [2015], Chakraborty and Ghosal [2021], Wang and Ghosal [2023] and the references therein. This approach bypasses the need for MCMC, leading to more efficient computations, especially in large-scale problems. The method's flexibility allows different tuning parameters for varying objectives, with smaller values suited for estimation and larger values for consistent model selection. Additionally, this sparse projection-posterior method can be implemented using a distributed computing setup, enabling efficient processing of large datasets and addressing privacy concerns by avoiding the need to store data on a single machine. Although substantial time gains from distributed implementation become prominent only for very large sample sizes (Supplement A), the framework offers a more immediate and practically important benefit in smaller settings: it eliminates the need to share raw data across sites. Since our projection-posterior method operates entirely on aggregated summary statistics $(\boldsymbol{X}^\top\boldsymbol{X}, \boldsymbol{X}^\top\boldsymbol{Y}, \boldsymbol{Y}^\top\boldsymbol{Y})$ for design matrix $\boldsymbol{X}$ and response vector $\boldsymbol{Y}$, it naturally accommodates data-federated environments where privacy, regulatory, or logistical constraints prohibit central data pooling. This feature distinguishes it from existing high-dimensional Bayesian regression approaches, which typically require full data access for MCMC sampling or posterior updates, making our approach particularly suitable for multi-institutional studies and privacy-preserving inference.

Another interpretation is to view the sparse parameter as a projection of a "super-parameter" without sparsity, using a misspecified likelihood for inference. This dual perspective makes the sparse projection-posterior approach computationally efficient and theoretically appealing for high-dimensional inference. For high-dimensional sparse group-regression,

variational inference (VI), as discussed in Subrahmanya and Shin [2013], Babacan et al. [2014], Komodromos et al. [2023], iteratively obtains the Kullback-Leibler projection of the posterior distribution to a tractable class of distributions, giving a random measure that mimics the posterior distribution. The immersion posterior gives a different random measure by transforming posterior samples from an unrestricted posterior distribution via a typically non-iterative optimization. The immersion posterior usually has a more explicit description that helps establish convergence, model selection, and coverage properties. Further discussion of similarities and differences between VI and the immersion posterior is given in Pal and Ghoshal [2024]. Group selection methods can address the high-dimensional additive regression problem via a basis-expansion technique. In this paper, we also develop a Bayesian method for high-dimensional additive regression models using group selection.

The rest of the paper is organized as follows. We delineate the proposed procedure in Section 2, describing the steps leading up to the three sparse projection maps. Following the detailed description, we formally state the asymptotic optimality properties of the three proposed projection-posteriors in Section 3. We show that the sparse projection-posteriors concentrate around the true parameter at rates comparable to those of their frequentist counterparts and that model selection procedure is consistent for all three maps. We also guarantee coverage of credible sets obtained from a specific debiased map of the three group sparse projection-posteriors. In Section 4, we describe the application of grouped regression under sparsity to the nonparametric additive regression and show optimal convergence properties. We conduct extensive simulation studies under both linear and additive models with varying choices of the number of groups, the number of active groups, and sparsity. We also apply our method to MRI brain volume data from the Alzheimer's Disease Neuroimaging

Initiative (ADNI). These numerical results are displayed in Section 5.

2. Methodology

2.1 Model and Notation

Consider the linear model $\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$, with noise $\boldsymbol{\varepsilon} \sim \mathcal{N}(\boldsymbol{0}, \sigma^2 \boldsymbol{I}_n)$, where the response variable $\boldsymbol{Y} \in \mathbb{R}^n$, the design matrix $\boldsymbol{X} \in \mathbb{R}^{n \times p}$, and the coefficient vector $\boldsymbol{\beta} \in \mathbb{R}^p$. Let the true value of the regression coefficient parameter vector be $\boldsymbol{\beta}^0$. Suppose there are K disjoint groups in total, namely, $G_1, G_2, \ldots, G_K$. The group sizes are $|G_k| = p_k, k = 1, 2, \ldots, K$, such that $\sum_{k=1}^{K} p_k = p$. Without loss of generality, we can assume a few adjacent variables to be grouped, that is, $\boldsymbol{X} = [\boldsymbol{X}_{G_1}, \boldsymbol{X}_{G_2} \ldots \boldsymbol{X}_{G_K}]$, where $\boldsymbol{X}_{G_k} \in \mathbb{R}^{n \times p_k}$ for $k = 1, 2, \ldots, K$. Accordingly, the coefficients can also be grouped leading to $\boldsymbol{\beta} = (\boldsymbol{\beta}_1^{\mathrm{T}}, \ldots, \boldsymbol{\beta}_K^{\mathrm{T}})^{\mathrm{T}}$, and $\boldsymbol{\beta}_k \in \mathbb{R}^{p_k}$. The linear model can then be re-written as $\boldsymbol{Y} = \sum_{k=1}^{K} \boldsymbol{X}_{G_k} \boldsymbol{\beta}_k + \boldsymbol{\varepsilon}$. The assumption here is that the setup is group sparse, meaning that there can be more than one group that does not influence the response variable. As a result, all variables within such unimportant groups are removed from the model. Suppose only $s_0 \leq K$ important groups exist, where s_0 is unknown. We seek to identify the active groups and accurately recover the regression parameter. Also, let $p_0 < p$ denote the number of important variables. Let S_0 be the set of group indices corresponding to the signal groups, that is, $S_0 = \{k : 1 \leq k \leq K, \boldsymbol{\beta}_k^0 \neq \boldsymbol{0}_{p_k}\}$, where $\boldsymbol{0}_{p_k}$ is the p_k-dimensional vector of zeroes. Consequently, the cardinality of S_0 is s_0 and moreover, we can write $p_0 = \sum_{k \in S_0} p_k$.

Throughout this paper, we denote the minimum and maximum group sizes by $p_{\min}$ and $p_{\max}$ respectively. The j-th column of the design matrix is denoted by $\boldsymbol{X}^{(j)}$, whereas the design matrix without the j-th column is denoted by $\boldsymbol{X}^{(-j)}$. Similarly, the design matrix

without the g-th group is denoted by $\boldsymbol{X}_{-g}$. That is, $\boldsymbol{X}_{-g} \in \mathbb{R}^{n \times \sum_{k \neq g} p_k}$ does not contain any variable belonging to the g-th group. The l-th variable within the g-th group is denoted by $\boldsymbol{X}_{G_g}^{(l)}$. Define the gram matrix as $\hat{\boldsymbol{\Sigma}} = n^{-1}\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X}$, so that the inner-product between the j-th and k-th variables forms the (j,k)-th element of this matrix. We define $\hat{\Sigma}_{j,k}$ as the (j,k)-th element of the Gram matrix. Moreover, let $\hat{\boldsymbol{\Sigma}}_{j,-j} = \hat{\boldsymbol{\Sigma}}_{-j,j}^{\mathrm{T}}$ and $\hat{\boldsymbol{\Sigma}}_{-j,-j}$ is the Gram matrix of $\boldsymbol{X}^{(-j)}$. Various norms have been used across the paper. For a d-dimensional vector $\boldsymbol{x} \in \mathbb{R}^d$, we write $\|\boldsymbol{x}\|_q = (\sum_{i=1}^d |a_i|^q)^{1/q}$, $\|\boldsymbol{x}\|_\infty = \max_{1 \leq i \leq d} |x_i|$ and drop the subscript q if $q = 2$. We use the usual O and o notations, and write $a_n \asymp b_n$ if both $a_n = O(b_n)$ and $b_n = O(a_n)$.

2.2 Prior Specification and the Unrestricted Posterior

The central idea here is to initially disregard the group's sparsity, inherent to the regression model, while specifying the prior distribution on the coefficient vector. Instead, the goal here is to specify a prior distribution so that the posterior can be computed efficiently and in a relatively simple form. We thus use the conjugate normal prior for the parameter vector, assuming the error standard deviation $\sigma > 0$ to be given as $\boldsymbol{\beta}|\sigma \sim \mathcal{N}_p(\boldsymbol{0}, \sigma^2 a_n^{-1}\boldsymbol{I}_p)$. Consequently, defining $\hat{\boldsymbol{\beta}}^{\mathrm{R}} = \left(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X} + a_n\boldsymbol{I}_p\right)^{-1}\boldsymbol{X}^{\mathrm{T}}\boldsymbol{Y}$, the ridge regression estimator with penalty a_n^{-1}, the posterior is obtained as $\boldsymbol{\beta}|(\boldsymbol{X}, \boldsymbol{Y}, \sigma) \sim \mathcal{N}_p\big(\hat{\boldsymbol{\beta}}^{\mathrm{R}}, \sigma^2\big(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X} + a_n\boldsymbol{I}_p\big)^{-1}\big)$. The posterior distribution of $\boldsymbol{X}\boldsymbol{\beta}$ given σ is then given by $\boldsymbol{X}\boldsymbol{\beta}|(\sigma, \boldsymbol{Y}) \sim \mathcal{N}_n\big(\boldsymbol{X}\hat{\boldsymbol{\beta}}^{\mathrm{R}}, \sigma^2\boldsymbol{H}(a_n)\big)$, where $\boldsymbol{H}(a_n) = \boldsymbol{X}\big(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X} + a_n\boldsymbol{I}_p\big)^{-1}\boldsymbol{X}^{\mathrm{T}}$. The posterior distribution of the quantity $\boldsymbol{X}\boldsymbol{\beta} - \boldsymbol{X}\boldsymbol{\beta}^0 = \boldsymbol{\eta}$, given the noise variance σ^2, is given by $\mathcal{N}_n\left(\boldsymbol{\mu}, \sigma^2\boldsymbol{H}(a_n)\right)$, where $\boldsymbol{\mu} = \boldsymbol{X}(\hat{\boldsymbol{\beta}}^{\mathrm{R}} - \boldsymbol{\beta}^0)$. However, this posterior is supported on non-sparse vectors, and it does not account for the fact that all groups are irrelevant to the response. To incorporate group-level sparsity in the posterior, we rely on a transformation, immersing a posterior sample without sparsity into

the fuzzy set of sparse vectors with a lower effective dimension. The induced *Immersion Posterior*, a term coined by Wang and Ghosal [2023], will be generically referred to as the sparse projection-posterior in our context. In the next section, we discuss the strategy we adopted in this paper to obtain the group-sparse induced posterior. Specific names will be used for the sparse projection-posterior corresponding to specific group sparsity-generating maps.

For the sparse projection-posterior approach, conditional posterior draws are made given σ^2, requiring the posterior distribution of σ^2 to be consistent to ensure the desired asymptotic frequentist properties. However, when p grows at a rate comparable to or faster than n, the marginal vanilla posterior for σ^2, obtained using a standard inverse-gamma prior in the regression model, becomes inconsistent. Specifically, with the prior for $\sigma^2 \propto 1/\sigma^2$ (or equivalently, $p(\upsilon) \propto 1/\upsilon$), the posterior distribution is $\upsilon \mid \boldsymbol{Y} \sim \text{Gamma}(n/2,\, n\hat{\sigma}_n^2/2)$, where $\hat{\sigma}_n^2 = n^{-1}\boldsymbol{Y}^{\mathrm{T}}(\boldsymbol{I}_n - \boldsymbol{H}(a_n))\boldsymbol{Y}$. Hence, $\mathbb{E}[\upsilon \mid \boldsymbol{Y}] = n/\{\boldsymbol{Y}^{\mathrm{T}}(\boldsymbol{I}_n - \boldsymbol{H}(a_n))\boldsymbol{Y}\}$, which diverges as $n \to \infty$. This issue can be addressed by applying a data-dependent map to the precision υ. Using a non-informative Gamma prior on υ and the map $\iota : \upsilon \mapsto \kappa\upsilon = \upsilon^*$, we obtain the induced posterior $\upsilon^*|\boldsymbol{Y} \sim \text{Gamma}(n/2, \kappa^{-1}n\hat{\sigma}_n^2/2)$, with mean $\kappa/\hat{\sigma}_n^2$, where $\hat{\sigma}_n^2 = n^{-1}\boldsymbol{Y}^{\mathrm{T}}(\boldsymbol{I}_n - \boldsymbol{H}(a_n))\boldsymbol{Y}$. Now, choosing $\kappa = \hat{\sigma}_n^2/\tilde{\sigma}^2$, where $\tilde{\sigma}^2$ is a consistent estimator of σ^2 (for example, from Sun and Zhang [2012]), ensures that the posterior mean of the induced posterior becomes a consistent for the true precision. This is because, the posterior variance of this induced posterior distribution is $2/(n\tilde{\sigma}^4) \to 0$ as $n \to \infty$, since $\tilde{\sigma}^2$ converges to the finite positive constant σ_0^2 by the consistency of $\tilde{\sigma}^2$ for σ^2. By the smooth one-to-one correspondence between precision υ and noise variance σ^2, we can also ensure that the induced posterior $\sigma^{*2} = 1/\upsilon^*$ is consistent for the true error variance σ^{0^2}.

2.3 Sparse Projection Maps

To transform non-sparse posterior samples into a space that reflects the desired sparsity, for $\boldsymbol{u} \in \mathbb{R}^p$, where $\boldsymbol{u} = (\boldsymbol{u}_1, \ldots, \boldsymbol{u}_K)^{\mathrm{T}}$, we apply a mapping with L as a loss function and $\mathcal{P}_{\lambda_n}(\cdot)$ as a penalty function, given by:

$$\iota : \boldsymbol{\beta} \mapsto \boldsymbol{\beta}^* = \underset{u \in \mathbb{R}^p}{\arg\min} \{ L(\boldsymbol{\beta}, \boldsymbol{u}) + \sum_{k=1}^{K} \mathcal{P}_{\lambda_n}(\|\boldsymbol{u}_k\|) \}, \tag{2.1}$$

designed to enforce sparsity in the solution, using a tuning parameter λ_n depending on n to control the level of sparsity. The loss function is designed to capture the discrepancy between $\boldsymbol{X}\boldsymbol{\beta}^*$ from $\boldsymbol{X}\boldsymbol{\beta}$, which we take to be the squared error loss $L(\boldsymbol{\theta}, \boldsymbol{u}) = n^{-1}\|\boldsymbol{X}\boldsymbol{\theta} - \boldsymbol{X}\boldsymbol{u}\|^2$. In this paper, we demonstrate a variety of penalty options, including group LASSO, group SCAD, and adaptive group LASSO, representing a broad spectrum of convex and non-convex penalty functions.

2.4 Sparse Projection-posteriors under Different Maps

The Group LASSO is one of the most popular models for grouped regression problems in which not all groups are relevant to the study. However, if a group is deemed important, all the variables within that group are considered relevant to the response of interest. We use this group LASSO penalty $\mathcal{P}_{\lambda_n}(\|\boldsymbol{u}_k\|) = \lambda_n \sqrt{p_k} \|\boldsymbol{u}_k\|$ for $\boldsymbol{u}_k \in \mathbb{R}^{p_k}$ to transform a posterior sample to $\boldsymbol{\beta}^*$ corresponding to the tuning parameter λ_n. The induced posterior distribution of $\boldsymbol{\beta}^*$ is called the Group LASSO projection-posterior.

The Group SCAD penalty extends the Smoothly Clipped Absolute Deviation (SCAD) penalty [Fan and Li, 2001] to the group setting. It is a non-convex penalty function that promotes sparsity while mitigating the bias typically introduced by convex penalties such as LASSO. The group SCAD penalty $\mathcal{P}_{\lambda_n}(|t|)\}$ is given by $\lambda_n |t| \mathbb{1}(|t| \leq \lambda_n) + \frac{|t|^2 - 2\tau\lambda_n|t| + \lambda_n^2}{2(\tau - 1)} \mathbb{1}(\lambda_n <$

$|t| \leq \tau\lambda_n) + \frac{\lambda_n^2(\tau+1)}{2}\mathbb{1}(|t| > \tau\lambda_n)$, where $\mathbb{1}(\cdot)$ is the indicator function of a set. for $t > 0$ and $\tau > 2$. The resulting induced posterior of $\boldsymbol{\beta}^*$ is called the Group SCAD projection-posterior.

While the group LASSO penalty ensures that larger groups do not unduly dominate the model just because they contain more variables, the adaptive lasso extends this concept by applying a penalty that is inversely proportional to the estimated ℓ_2-norm of the coefficients within each group, typically derived from an initial group LASSO fit. This approach introduces an additional layer of flexibility, allowing the model to apply different penalties to different groups based on their preliminary importance. Let the group LASSO estimator of $\boldsymbol{\beta}$ be $\hat{\boldsymbol{\beta}}^{\text{GL}} = \arg\min\left\{n^{-1}\|\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{u}\|^2 + \tilde{\lambda}_n \sum_{k=1}^{K} \|\boldsymbol{u}_k\| : \boldsymbol{u} \in \mathbb{R}^p\right\}$ where the tuning parameter $\tilde{\lambda}_n$ should be such that the group LASSO estimator $\hat{\boldsymbol{\beta}}^{\text{GL}}$ is consistent with rate r_n. Then, the Adaptive Group LASSO projection-posterior is based on the adaptive group LASSO penalty function, $\mathcal{P}(\|\boldsymbol{u}_k\|) = w_k\|\boldsymbol{u}_k\|$, where the weights are obtained by using the group LASSO estimator $\hat{\boldsymbol{\beta}}^{\text{GL}}$, that is $w_k = \|\hat{\boldsymbol{\beta}}_k^{\text{GL}}\|^{-1}$, with a thresholded weight for zero groups under the initial group LASSO fit. The induced posterior distribution of $\boldsymbol{\beta}^*$ is called the Group Adaptive LASSO projection-posterior.

Below, we rewrite the proposed algorithm with the three penalty choices studied across the scenarios in this paper. Although we have chosen three penalization techniques to project our initial posterior samples into the required lower-dimensional space to impose sparsity, we would like to emphasize that our proposed approach is not restricted to these three penalties. Besides these, there is another popular non-convex penalty, the minimax concave penalty (MCP) introduced in Zhang [2010], which tries to mitigate the bias due to shrinkage. The group MCP penalty [Huang et al., 2012] is, thus, another viable option for the sparse projection-posterior. We provide theoretical guarantees for estimation and model selection

corresponding to the group MCP penalty in Supplement C.

Algorithm 1: Sparse Projection-Posterior Sampler

Input: $\boldsymbol{X} \in \mathbb{R}^{n\times p}$, $\boldsymbol{Y} \in \mathbb{R}^n$, prior scale a_n, penalty P_{λ_n}, tuning λ_n, draws M

Output: $\{\boldsymbol{\beta}^{*(m)}\}_{m=1}^M$

$\hat{\boldsymbol{\beta}}^{\mathrm{R}} \leftarrow (\boldsymbol{X}^\top\boldsymbol{X} + a_n I_p)^{-1}\boldsymbol{X}^\top\boldsymbol{Y}, \quad \hat{\sigma}_n^2 \leftarrow n^{-1}\boldsymbol{Y}^\top(I_n - H(a_n))\boldsymbol{Y}$

for $m = 1, \ldots, M$ **do**

Draw $\sigma^{*2(m)} \mid \boldsymbol{Y} \sim \mathrm{InvGamma}\Big(\frac{n}{2}, \kappa^{-1}\frac{n\hat{\sigma}_n^2}{2}\Big)$, with $\kappa = \hat{\sigma}_n^2/\tilde{\sigma}^2$

Draw $\boldsymbol{\beta}^{(m)} \sim \mathcal{N}_p\Big(\hat{\boldsymbol{\beta}}^{\mathrm{R}}, \sigma^{*2(m)}(\boldsymbol{X}^\top\boldsymbol{X} + a_n I_p)^{-1}\Big)$

$\boldsymbol{\beta}^{*(m)} \leftarrow \arg\min_{\boldsymbol{u}\in\mathbb{R}^p}\Big\{n^{-1}\|\boldsymbol{X}\boldsymbol{\beta}^{(m)} - \boldsymbol{X}\boldsymbol{u}\|^2 + \sum_{k=1}^K P_{\lambda_n}\big(\|\boldsymbol{u}_k\|_2\big)\Big\}$

return $\{\boldsymbol{\beta}^{*(m)}\}_{m=1}^M$

Penalty choices $P_{\lambda_n}(\cdot)$ Let $t = \|\boldsymbol{u}_k\|_2$ and recall the group size p_k.

Group LASSO : $P_{\lambda_n}^{\mathrm{GL}}(t) = \lambda_n\sqrt{p_k}\,t.$

Group SCAD : $\lambda_n|t|\mathbb{1}(|t| \le \lambda_n) + \dfrac{|t|^2 - 2\tau\lambda_n|t| + \lambda_n^2}{2(\tau-1)}\mathbb{1}(\lambda_n < |t| \le \tau\lambda_n) + \dfrac{\lambda_n^2(\tau+1)}{2}\mathbb{1}(|t| > \tau\lambda_n).$

Adaptive Group LASSO : $P_{\lambda_n}^{\mathrm{AGL}}(t) = \lambda_n\, w_k\, t, \qquad w_k = \big\|\hat{\boldsymbol{\beta}}_k^{\mathrm{GL}}\big\|_2^{-1},$

where $\hat{\boldsymbol{\beta}}^{\mathrm{GL}}$ is a preliminary group LASSO fit with its own penalty $\tilde{\lambda}_n$ (chosen so that $\hat{\boldsymbol{\beta}}^{\mathrm{GL}}$ is consistent at rate r_n). The optimization in Step (2c) was carried out in R using the package `grpreg`. We used `penalty = "grLasso"`, `"grSCAD"`, and `"grMCP"` for the group LASSO, group SCAD, and group MCP maps, respectively; the adaptive group LASSO map was implemented via `penalty = "grLasso"` with the `group.multiplier` option based on an initial group-LASSO fit. In all runs we set `family = "gaussian"`, `intercept = FALSE`, standardized the design matrix, and used the default convergence settings `eps = 1e-4` and `max.iter = 10000`. A single cross-validated value of λ_n was selected and then reused for all posterior draws.

3. Main Results

The theoretical properties of the group LASSO have been studied extensively. Bühlmann and van de Geer [2011] provided oracle rates for estimation and prediction. Nardi and Rinaldo [2008] proved selection consistency using the irrepresentability condition [Zhao and Yu, 2006, Javanmard and Montanari, 2013] and the estimation and prediction consistency using the restricted eigenvalue condition [Bickel et al., 2009], whereas Wei and Huang [2010] showed group selection to be consistent by making use of the sparse Riesz condition [Zhang and Huang, 2008]. Theoretical properties of group SCAD and group adaptive LASSO penalization techniques were derived under the high-dimensional setting by Guo et al. [2015] and Wei and Huang [2010], respectively. In this section, we study the main assumptions required to establish the estimation and selection consistency of our three proposed sparse projection-posterior methods, followed by the corresponding theorems. We also study a debiased projection-posterior that gives asymptotically accurate coverage.

3.1 Main Assumptions

This section outlines the main assumptions required for the theoretical development and justification of the proposed sparse projection-posterior methods. These assumptions ensure that the regression model behaves appropriately in sparse and high-dimensional settings and enable us to derive consistency and convergence results. Each assumption addresses a specific aspect of the model, from the properties of the design matrix to signal strength requirements, ensuring that the proposed methods achieve optimal rates of recovery, prediction, and variable selection. The conditions required for establishing the theoretical guarantees of our approach are exactly the same as their frequentist counterparts, including the order

of the tuning parameter and the minimal signal strength requirements. In addition to these assumptions introduced in the frequentist penalization approaches, we only require a condition on the the prior parameter a_n (Assumption 3.3). Thus, our assumptions on the penalty level, design, and minimal signal are fully consistent with the frequentist high-dimensional literature for the corresponding penalties. Consequently, the projection-posterior achieves the same rates and selection guarantees while providing Bayesian uncertainty quantification, and the recommended λ_n choices align with standard frequentist guidance.

Assumption 3.1 (Bounded design condition)**.** *The predictor variables* $|x_{ij}| \leq M_1$ *for all* $i = 1, \ldots, n$, $j = 1, \ldots, p$, *for some* $M_1 > 0$ *free of* n.

Assumption 3.2 (Bounded true mean condition)**.** *The maximum of the absolute value of the expected response under the truth,* $\max_{1 \leq i \leq n} |\mathbb{E}_{\boldsymbol{\theta}^0}(Y_i)|$, *is bounded.*

Assumptions 3.1 and 3.2 help control the complexity of the model, and are commonly assumed in the literature.

Assumption 3.3 (Non-collinearity condition)**.** *When* $p > n$, $\text{rank}(\boldsymbol{X}) = n$ *and the singular values* $d_1, \ldots, d_n$ *of* $\boldsymbol{X}$ *satisfy* $\min\{d_j^2 : j = 1, \ldots, n\}/(na_n) \to 0$.

The assumption avoids instability problems arising from multicollinearity by requiring that the design matrix's rank equals the sample size and its singular values remain somewhat large relative to the prior parameter a_n, which may be chosen as small as required.

Assumption 3.4 (Compatibility condition)**.** *Let* $S \subset \{1, 2, \ldots, K\}$ *denote the index set of active groups and let* $0 < L < \infty$. *The design matrix* $\boldsymbol{X} \in \mathbb{R}^{n \times p}$ *is said to satisfy the* group compatibility condition *on the cone* $\mathcal{C}(S, L) = \left\{\boldsymbol{v} \in \mathbb{R}^p : \sum_{k \in S^c} \sqrt{p_k}\, \|\boldsymbol{v}_k\| \leq L \sum_{k \in S} \sqrt{p_k}\, \|\boldsymbol{v}_k\|\right\}$, *with compatibility constant* $\phi(S) = \phi(S, L) > 0$, *if for all* $\boldsymbol{v} \in \mathcal{C}(S, L)$ *it holds that* $\left(\sum_{k \in S} \sqrt{p_k}\, \|\boldsymbol{v}_k\|\right)^2 \leq \frac{\sum_{k \in S} p_k}{\phi^2(S)} \cdot \frac{\|\boldsymbol{X}\boldsymbol{v}\|^2}{n}$.

This assumption ensures that the set of predictors satisfies a sparsity-related geometric property, needed to link the sparsity structure of the coefficients to the observed data. We also make explicit that the inequality is required only on the cone $\mathcal{C}(S, L)$ which is the set of directions delivered by the basic-inequality/KKT steps in our proofs. When $p > n$, we have $\text{null}(\boldsymbol{X}) \neq 0$. If there exists a nonzero $\boldsymbol{v} \in \mathcal{C}(S, L) \cap \text{null}(\boldsymbol{X})$, then $|\boldsymbol{X}\boldsymbol{v}| = 0$ while $\sum_{k \in S} \sqrt{p_k}, |\boldsymbol{v}_k| > 0$, which would force $\phi(S, L) = 0$. Assumption 3.4 explicitly rules out this pathology by requiring $\phi(S, L) > 0$; equivalently, it requires that the empirical Gram matrix be strictly positive on the cone $\mathcal{C}(S, L)$. This is the standard compatibility/restricted-eigenvalue requirement in high-dimensional analysis: we do not ask for invertibility on $\mathbb{R}^p$ (impossible when $p > n$), only on the restricted set of directions relevant for sparse recovery.

Assumption 3.5 (Non-singularity of the Gram matrix)**.** *Let $\boldsymbol{C}_{S_0} = n^{-1}\boldsymbol{X}_{S_0}^{\mathrm{T}}\boldsymbol{X}_{S_0}$ be the Gram matrix formed by the columns in the true active set S_0. Assume $\boldsymbol{C}_{S_0}$ is positive definite with its smallest eigenvalue bounded away from zero, i.e. $\lambda_{\min}(\boldsymbol{C}_{S_0}) \geq M_2 > 0$, equivalently, for all $\boldsymbol{u} \in \mathbb{R}^{p_0}$, $\boldsymbol{u}^{\mathrm{T}}\boldsymbol{C}_{S_0}\boldsymbol{u} \geq M_2\|\boldsymbol{u}\|^2$.*

This ensures that the relevant subset of the design matrix has eigenvalues bounded away from zero, allowing stability of estimates for the active groups and consistent model selection.

Assumption 3.6 (Beta-min condition)**.** *The magnitude of a signal (coefficient of an active predictor) is bounded below as follows:*

(A) *(Group LASSO projection-posterior)* $\min_{k \in S_0} \|\boldsymbol{\beta}_k^0\|_\infty \geq \max\left\{ \dfrac{\sqrt{\log p_0}}{n}, \dfrac{\lambda_n \sqrt{s_0} p_{\max}^{3/2}}{\lambda_{\min}(\boldsymbol{C}_{n(11)})} \right\}$;

(B) *(Group SCAD projection-posterior)* $\min_{k \in S_0} \|\boldsymbol{\beta}_k^0\| \geq M_3 n^{-(1-c_3)/2}$ *for constants* $c_3, M_3 > 0$;

(C) *(Group adaptive LASSO projection-posterior) Let r_n be the convergence rate of an initial estimator $\hat{\boldsymbol{\beta}}^{\mathrm{GL}}$. Then* $\min_{k \in S_0} \|\boldsymbol{\beta}_k^0\| > \max\left\{ p_{\max}^{3/4}\sqrt{\lambda_n s_0}, \dfrac{\sqrt{p_{\max} \log s_0}}{n}, \dfrac{p_{\max}^{5/2} s_0^2}{r_n p_{\min}} \right\}$.

The Beta-min condition is necessary to separate signals from noise.

Assumption 3.7 (Irrepresentability condition)**.** *For some* $0 < \nu < 1$ *and for every* $k \in S_0^c$*, we have* $\|\boldsymbol{X}_{G_k}^{\mathrm{T}} \boldsymbol{X}_{S_0} (\boldsymbol{X}_{S_0}^{\mathrm{T}} \boldsymbol{X}_{S_0})^{-1}\| \leq 1 - \nu$.

The condition restricts the total effect of inactive groups of predictors compared to active ones and is needed to establish model selection consistency [Zhao and Yu, 2006].

3.2 Recovery and Prediction Rate

In this section, we analyze the recovery and prediction consistency properties of the group LASSO, group SCAD, and group adaptive LASSO maps. We focus on establishing the conditions under which each posterior is consistent for estimation and demonstrates superior predictive performance in high-dimensional settings with group structure among the covariates. We study the following three properties:

$$\Pi\big(\sum_{k=1}^{K} \sqrt{p_k} \|\boldsymbol{\beta}_k^* - \boldsymbol{\beta}_k^0\| \geq c_1 r_{1n} \big| \boldsymbol{Y}\big) \to 0; \tag{3.1}$$

$$\Pi\big(\|\boldsymbol{\beta}^* - \boldsymbol{\beta}^0\|^2 \geq c_1^2 r_{2n} \big| \boldsymbol{Y}\big) \to 0; \tag{3.2}$$

$$\Pi\big(n^{-1} \|\boldsymbol{X}(\boldsymbol{\beta}^* - \boldsymbol{\beta}^0)\|^2 \geq c_2 r_{3n} \big| \boldsymbol{Y}\big) \to 0, \tag{3.3}$$

in probability under the true distribution for some rates r_{1n}, r_{2n}, r_{3n}. We identify the applicable rates (r_{1n}, r_{2n}, and r_{3n}) for the group LASSO, group SCAD, and group adaptive LASSO projection-posteriors. Note that the following theorems also apply to the ungrouped case by setting $p_k = 1$ for $k = 1, \ldots, K$. For instance, taking $\lambda_n \asymp \sqrt{(\log p)/n}$, we obtain $r_{2n} = (s_0^2 \log p)/n$ and $r_{3n} = (s_0 \log p)/n$, which are optimal for sparse high-dimensional linear regression and are also attained by the ungrouped LASSO.

To prove that the induced posterior from the SCAD penalized map is estimation and prediction consistent with the optimal rate, we first need to show an oracle property of this map. This requires an additional assumption about the signal strength to show that the group SCAD projection-posterior has an oracle property, as if we knew the locations of the zero coefficients in advance. Similar assumptions will later be used to show that the sparse projection-posteriors are model-selection consistent. We do not, however, need to make any assumptions on the minimum signal strength to obtain rates for the adaptive group LASSO projection-posterior. Scaling the penalty by the group LASSO estimator $\hat{\boldsymbol{\beta}}^{\mathrm{GL}}$ guarantees the same convergence rate under reduced conditions. We state the optimal recovery and prediction consistency rates for the three sparse projection-posteriors below.

Theorem 1 (Recovery and prediction rates)**.** *For $\lambda_n \geq \lambda_0$, where*

$$\lambda_0^2(x, p) = \frac{8}{n}\left(1 + \sqrt{\frac{4x + 4\log p}{p_{\min}}} + \frac{4x + 4\log p}{p_{\min}}\right) \text{ with } x > 0, \tag{3.4}$$

the following assertions hold with posterior probability at least $1 - \exp(-x)$:

1. *(Group LASSO Projection-posterior) Under Assumptions 3.1, 3.2, 3.3, 3.4 and 3.5, we have for $c_1, c_2 > 0$, (3.1) holds with $r_{1n} = \lambda_n \sum_{k=1}^{s_0} p_k$, (3.2) holds with $r_{2n} = \lambda_n^2 \big(\sum_{k=1}^{s_0} p_k\big)^2 / p_{\min}$ and (3.3) holds with $r_{3n} = \lambda_n^2 \sum_{k=1}^{s_0} p_k$.*

2. *(Group SCAD Projection-posterior) If Assumptions 3.1, 3.2, 3.3, 3.4, 3.5 and 3.6(B) hold, then for $c_1, c_2 > 0$, (3.1) holds with $r_{1n} = \lambda_n \sum_{k=1}^{s_0} p_k$, (3.2) holds with $r_{2n} = \lambda_n^2 \big(\sum_{k=1}^{s_0} p_k\big)^2 / p_{\min}$ and (3.3) holds with $r_{3n} = \lambda_n^2 \sum_{k=1}^{s_0} p_k$.*

3. *(Group Adaptive LASSO Projection-posterior) Let Assumptions 3.1, 3.2, 3.3, 3.4, and 3.5 hold. For some $c_1, c_2 > 0$, (3.1) holds with $r_{1n} = \lambda_n \sum_{k=1}^{s_0} p_k$, (3.2) holds with $r_{2n} = \lambda_n^2 \big(\sum_{k=1}^{s_0} p_k\big)^2 / p_{\min}$, and (3.3) holds with $r_{3n} = \lambda_n^2 \sum_{k=1}^{s_0} p_k$.*

3.3 Model Selection Consistency

Model selection consistency is a critical feature of sparse high-dimensional regression methods, especially when the number of predictors exceeds or is comparable to the sample size. A regression method is said to possess model selection consistency if, as the sample size increases, it accurately identifies the active groups of predictors, meaning that it selects the correct set of non-zero coefficients while exactly setting the irrelevant ones to zero. In high-dimensional settings, sparsity is key because the number of truly important variables is often much smaller than the total number of available predictors. However, one challenge of sparse regression is balancing bias and variance. For the sparse projection maps described in this paper, while aggressive penalties may shrink the components of the posterior samples too much, thus introducing bias, lighter penalties may not fully enforce sparsity. Model selection consistency ensures that, as the sample size grows, the chosen model correctly reflects the underlying sparse structure without underfitting or overfitting: $\Pi(\{\|\boldsymbol{\beta}_k^*\| > 0 \text{ for all } k \in S_0, \text{and } \boldsymbol{\beta}_k^* = 0 \text{ for all } k \in S_0^{\text{c}} | \boldsymbol{Y}) \to 1$ as $n \to \infty$ in probability under the true distribution.

The group LASSO penalized estimator is model selection consistent [Nardi and Rinaldo, 2008] under certain assumptions, including but not restricted to the Beta-min and the irrepresentability conditions. We show here that the group LASSO projection-posterior is also model selection consistent if the prior parameter a_n satisfies Assumption 3.3. The model selection consistency of the group SCAD penalized projection-posterior follows from the oracle property provided in Assumption **??**. According to the Oracle property, the minimizer of the optimization function using the group SCAD penalty results in the Oracle posterior $\boldsymbol{\beta}_{S_0}$ for large n under certain assumptions, including a restriction on the minimum signal

strength. The oracle projection is asymptotically a local minimizer under the group SCAD map, even in high-dimensional settings. Thus, for large n, there is a guarantee for the sparse projection $\boldsymbol{\beta}^*$ using the group SCAD map to select the correct model with high posterior probability. The sparse projection-posterior resulting from the group adaptive LASSO map, on the other hand, can be shown to be selection consistent under weaker conditions than the group LASSO map. We can avoid imposing the irrepresentability condition by weighting the penalty by the norm of the group LASSO estimator for each group. However, the Beta-min condition will still be required. Here, the assumption of the minimum signal strength differs from the requirements under the group LASSO and SCAD maps.

Theorem 2 (Model selection consistency)**.** *The following assertions hold:*

1. *(Group LASSO Projection-posterior) If* $\log(p - p_0)/n\lambda_n^2 \to 0$ as $n \to \infty$ *and Assumptions 3.1, 3.2, 3.3, 3.5, 3.6(A) and 3.7 hold, then the group LASSO projection-posterior is model selection consistent.*

2. *(Group SCAD Projection-posterior) If Assumptions 3.1,3.2,3.3 and 3.6(B) hold, then, for* $\lambda_n = o\big(n^{-(1-c_3+c_4)/2}\big)$ *and* $p/(\lambda_n n^{3/2}) \to 0$*, the group SCAD projection-posterior is model selection consistent.*

3. *(Group Adaptive LASSO Projection-posterior) If* $p_{\max}\log(K - s_0)/(\sqrt{n}\lambda_n r_n p_{\min}) \to 0$ *and Assumptions 3.1, 3.2, 3.3 and 3.6(C) hold, then the group adaptive LASSO projection-posterior is model selection consistent.*

3.4 Uncertainty Quantification

The sparse projection-posterior, although it concentrates around the true parameter at an optimal rate, consistently selects the correct signal groups nd automatically quantifies

the associated uncertainty by virtue of it being a Bayesian method; it does not give exact frequentist coverage. To achieve targeted coverage, we need to take a further step to adjust for the penalization bias. This process, known in the literature as 'debiasing,' has not previously been used in the Bayesian paradigm. Here, we debias each sparse projection-posterior sample to obtain the debiased projection-posterior samples, which we show have exact asymptotic coverage. It should be noted that no additional sampling is required at this stage. We only need to modify the sparse projection-posterior samples through a debiasing map. To define the debiased projection, we must first specify a reasonable representation of the inverse of the Gram matrix $\hat{\boldsymbol{\Sigma}}$. Following van de Geer et al. [2014], Honda [2021], define $\hat{\boldsymbol{\Theta}}$ that satisfies $\hat{\boldsymbol{\Theta}}\hat{\boldsymbol{\Sigma}} = \boldsymbol{I}_p$. Mimicking the construction of $\hat{\boldsymbol{\Theta}}$ from van de Geer et al. [2014] and Honda [2021], we define the group LASSO regression of the l-th variable in the g-th group on all the other groups, for $l = 1, 2, \ldots, p_g$, and $g = 1, 2, \ldots, K$, as follows: $\boldsymbol{\gamma}_g^l = \arg\min_{\gamma \in \mathbb{R}^{p-p_g}} \left\{ n^{-1} \| \boldsymbol{X}_{G_g}^{(l)} - \boldsymbol{X}_{-g}\boldsymbol{\gamma} \|^2 + \lambda_g^l \sum_{k \neq g} \sqrt{p_k} \|\boldsymbol{\gamma}_k\| \right\}$, where λ_g^l is the corresponding tuning parameter. Note that the vector $\boldsymbol{\gamma}_g^l$ can be rewritten as $\boldsymbol{\gamma}_g^l = ({\boldsymbol{\gamma}^l}_{g,1}^{\mathrm{T}}, \ldots, {\boldsymbol{\gamma}^l}_{g,g-1}^{\mathrm{T}}, {\boldsymbol{\gamma}^l}_{g,g+1}^{\mathrm{T}}, \ldots, {\boldsymbol{\gamma}^l}_{g,K}^{\mathrm{T}})^{\mathrm{T}}$, where each $\boldsymbol{\gamma}_{g,k}^l \in \mathbb{R}^{p_k}$ for $k \in \{1, 2, \ldots, K\} \setminus \{g\}$. We also define the matrix $\boldsymbol{\Gamma}_g \in \mathbb{R}^{p-p_g \times p_g}$ which is composed of the group LASSO estimators by regressing all the p_g variables of the g-th group on the rest of the $(K-1)$ groups, that is, $\boldsymbol{\Gamma}_g = (\boldsymbol{\gamma}_g^1, \boldsymbol{\gamma}_g^2, \ldots, \boldsymbol{\gamma}_g^{p_g})$. Next, let $\boldsymbol{R}_g$ denote the residual matrix obtained by regressing $\boldsymbol{X}_{G_g}$ on $\boldsymbol{X}_{-g}$ using the group LASSO penalty, that is, $\boldsymbol{R}_g = \boldsymbol{X}_{G_g} - \boldsymbol{X}_{-g}\boldsymbol{\Gamma}_g \in \mathbb{R}^{n \times p_g}$. The first-order KKT conditions for these column-wise group LASSO regressions are given by $-\frac{2}{n}\boldsymbol{X}_{-g}^{\mathrm{T}}(\boldsymbol{X}_{G_g}^{(l)} - \boldsymbol{X}_{-g}\boldsymbol{\gamma}_g^l) + \lambda_g^l \boldsymbol{\kappa}_g^l = \boldsymbol{0} \in \mathbb{R}^{p-p_g}$, where $\boldsymbol{\kappa}_g^l = ({\boldsymbol{\kappa}^l}_{g,1}^{\mathrm{T}}, \ldots, {\boldsymbol{\kappa}^l}_{g,g-1}^{\mathrm{T}}, {\boldsymbol{\kappa}^l}_{g,g+1}^{\mathrm{T}}, \ldots, {\boldsymbol{\kappa}^l}_{g,K}^{\mathrm{T}})^{\mathrm{T}} \in \mathbb{R}^{p-p_g}$ with $\boldsymbol{\kappa}_{g,k}^l \in \mathbb{R}^{p_k}$ for $k \neq g$ and $\|\boldsymbol{\kappa}_{g,k}^l\| \leq 1$ if $\|\boldsymbol{\gamma}_{g,k}^l\| = 0$ and $\boldsymbol{\kappa}_{g,k}^l = \boldsymbol{\gamma}_{g,k}^l / \|\boldsymbol{\gamma}_{g,k}^l\|$ if $\|\boldsymbol{\gamma}_{g,k}^l\| \neq 0$. Define the $(p - p_g) \times p_g$ matrix $\boldsymbol{K}_g = (\boldsymbol{\kappa}_g^1, \boldsymbol{\kappa}_g^2, \ldots, \boldsymbol{\kappa}_g^{p_g})$, and express the KKT condition in the matrix form $2n^{-1}\boldsymbol{X}_{-g}^{\mathrm{T}}\boldsymbol{R}_g = \boldsymbol{K}_g\boldsymbol{\Lambda}_g$, where $\boldsymbol{\Lambda}_g \in \mathbb{R}^{p_g \times p_g}$ is $\mathrm{diag}(\lambda_g^1, \ldots, \lambda_g^{p_g})$. Finally,

following Honda [2021], we have

$$\hat{\boldsymbol{\Theta}}^{\mathrm{T}} = \begin{bmatrix} \boldsymbol{I}_{p_1} & -\boldsymbol{\Gamma}_{2,1} & -\boldsymbol{\Gamma}_{3,1} & \dots & -\boldsymbol{\Gamma}_{K,1} \\ -\boldsymbol{\Gamma}_{1,2} & \boldsymbol{I}_{p_2} & -\boldsymbol{\Gamma}_{3,2} & \dots & -\boldsymbol{\Gamma}_{K,2} \\ -\boldsymbol{\Gamma}_{1,3} & -\boldsymbol{\Gamma}_{2,3} & \boldsymbol{I}_{p_3} & \dots & -\boldsymbol{\Gamma}_{K,3} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ -\boldsymbol{\Gamma}_{1,K} & -\boldsymbol{\Gamma}_{2,K} & -\boldsymbol{\Gamma}_{3,K} & \dots & \boldsymbol{I}_{p_K} \end{bmatrix} \begin{bmatrix} \boldsymbol{T}_1^2 & \mathbf{0} & \mathbf{0} & \dots & \mathbf{0} \\ \mathbf{0} & \boldsymbol{T}_2^2 & \mathbf{0} & \dots & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \boldsymbol{T}_3^2 & \dots & \mathbf{0} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \mathbf{0} & \mathbf{0} & \mathbf{0} & \dots & \boldsymbol{T}_K^2 \end{bmatrix}, \tag{3.5}$$

where $\boldsymbol{T}_g{}^2 = n^{-1}\boldsymbol{X}_{G_g}^{\mathrm{T}}\boldsymbol{R}_g \in \mathbb{R}^{p_g \times p_g}$ and $\boldsymbol{\Gamma}_{j,k}^{\mathrm{T}} = (\boldsymbol{\gamma}_{j,k}^1, \boldsymbol{\gamma}_{j,k}^2, \dots, \boldsymbol{\gamma}_{j,k}^{p_j})^{\mathrm{T}} \in \mathbb{R}^{p_j \times p_k}$. We now define the debiased Group LASSO projection as

$$\boldsymbol{\beta}^{**} = \boldsymbol{\beta}^* + \frac{1}{n}\hat{\boldsymbol{\Theta}}\boldsymbol{X}^{\mathrm{T}}(\boldsymbol{X}\boldsymbol{\beta} - \boldsymbol{X}\boldsymbol{\beta}^*) = \boldsymbol{\beta}^0 + \frac{1}{n}\hat{\boldsymbol{\Theta}}\boldsymbol{X}^{\mathrm{T}}\boldsymbol{\eta} - \boldsymbol{\Delta}, \tag{3.6}$$

where $\boldsymbol{\Delta} = (\hat{\boldsymbol{\Theta}}\hat{\boldsymbol{\Sigma}} - \boldsymbol{I}_p)(\boldsymbol{\beta}^* - \boldsymbol{\beta}^0) \in \mathbb{R}^p$. From (3.5), we have that the (j,k)-th block of $\hat{\boldsymbol{\Theta}}\hat{\boldsymbol{\Sigma}} - \boldsymbol{I}_p$ is given by the $(\boldsymbol{T}_j^{-2})^{\mathrm{T}}\boldsymbol{\Lambda}_j\boldsymbol{K}_{j,k}^{\mathrm{T}} \in \mathbb{R}^{p_j \times p_k}$ for $j \neq k$ and $\mathbf{0} \in \mathbb{R}^{p_j \times p_j}$ along the diagonal when $j = k$. Note that $\boldsymbol{K}_{j,k}^{\mathrm{T}} = (\boldsymbol{\kappa}_{j,k}^1, \boldsymbol{\kappa}_{j,k}^2, \dots, \boldsymbol{\kappa}_{j,k}^{p_j})^{\mathrm{T}} \in \mathbb{R}^{p_j \times p_k}$. A component-wise credible set defined based on the posterior distribution of this sparse projection $\boldsymbol{\beta}^{**}$ can be shown to achieve accurate asymptotic coverage corresponding to the debiased group-LASSO map motivated by the debiased group-LASSO estimator [Honda, 2021] defined by

$$\hat{\boldsymbol{\beta}}^{\mathrm{DGL}} = \boldsymbol{\beta}^0 + n^{-1}\hat{\boldsymbol{\Theta}}\boldsymbol{X}^{\mathrm{T}}\varepsilon - \boldsymbol{\Delta}^{\mathrm{DGL}}, \tag{3.7}$$

where $\boldsymbol{\Delta}^{\mathrm{DGL}} = (\hat{\boldsymbol{\Theta}}\hat{\boldsymbol{\Sigma}} - \boldsymbol{I}_p)(\hat{\boldsymbol{\beta}}^{\mathrm{GL}} - \boldsymbol{\beta}^0)$, and $\hat{\boldsymbol{\beta}}^{\mathrm{GL}}$ is the group LASSO estimator.

Assumption 3.8. *For all $j = 1, 2 \dots, p$, $\lambda_{\max}(\hat{\Sigma}_{j,j}) \leq C_1$ for some $C_1 > 0$, and for all $l = 1, \dots p_g$ and $g = 1, \dots, K$, $\|\boldsymbol{\gamma}_l^g\| \leq C_2$ for some $C_2 > 0$.*

Theorem 3 (Bernstein-von Mises Theorem)**.** *If Assumptions 3.1, 3.2, 3.3 and 3.8 hold, we have $\|\boldsymbol{\Delta}_k\| = o_P(n^{-1/2})$ uniformly in $k = 1, 2, \dots, K$, and*

$$\sup_B \left| \Pi\big(\sigma^{-1}\sqrt{n}(\boldsymbol{\beta}^{**} - \boldsymbol{\beta}^0) \in B \big| \boldsymbol{Y}\big) - \mathcal{N}_p(B; \boldsymbol{m}, \hat{\boldsymbol{V}}) \right| \to 0, \text{ as } n \to \infty,$$

where $\boldsymbol{m} = n^{-1/2}\hat{\boldsymbol{\Theta}}\boldsymbol{X}^{\mathrm{T}}\boldsymbol{\mu}$ and $\boldsymbol{V} = n^{-1}\hat{\boldsymbol{\Theta}}\boldsymbol{X}^{\mathrm{T}}\boldsymbol{H}(a_n)\boldsymbol{X}\hat{\boldsymbol{\Theta}}^{\mathrm{T}}$.

Let $\mathcal{C}_j = [\tilde{\beta}_j - q_{j,\alpha},\ \tilde{\beta}_j + q_{j,\alpha}]$ denote the symmetric component-wise $(1-\alpha)$-credible interval for β_j, for $j = 1, \ldots, p$, where $\tilde{\beta}_j$ is the median of the posterior distribution of β_j^{**}, and $q_{j,\alpha}$ is the $(1-\alpha/2)$-quantile of the posterior distribution of $|\beta_j^{**} - \tilde{\beta}_j|$. The proposed credible sets will provide accurate frequentist coverage in the limit if the normal limit of the centered debiased projection-posterior matches the asymptotic normal distribution of the debiased group-LASSO estimator.

Corollary 1 (Coverage of credible region)**.** *If Assumptions 3.1, 3.2, 3.3 and 3.8 hold, then for all* $j = 1, 2, \ldots, p$, $\lim_{n\to\infty} \sup_{\beta_j^0 \in \mathbb{R}} |\mathbb{P}_{\beta_j^0}(\beta_j^0 \in \mathcal{C}_j) - (1-\alpha)| = 0$.

The debiased projection described above only used two ingredients: a sparse approximate inverse $\hat{\boldsymbol{\Theta}}$ of the Gram matrix satisfying $\|\hat{\boldsymbol{\Theta}}\,\hat{\boldsymbol{\Sigma}} - \boldsymbol{I}_p\|_\infty = o_P(1)$ (constructed by the nodewise group LASSO) and a penalized estimator $\hat{\boldsymbol{\beta}}^{\text{init}}$ for which the remainder $\boldsymbol{\Delta}^{\text{init}} := (\hat{\boldsymbol{\Theta}}\hat{\boldsymbol{\Sigma}} - \boldsymbol{I}_p)(\hat{\boldsymbol{\beta}}^{\text{init}} - \boldsymbol{\beta}^0)$ is $o_P(n^{-1/2})$. Hence the same debiasing formula applies verbatim when $\hat{\boldsymbol{\beta}}^{\text{init}}$ is the group SCAD or adaptive group LASSO (AGL) estimator. We state the resulting Bernstein-von Mises limits in Supplement D; the proofs follow exactly the argument of Theorem 3, and are therefore omitted.

4. Nonparametric Additive Regression

High-dimensional nonparametric additive regression under sparsity can effectively use the group LASSO formulation. The COSSO proposed by Lin and Zhang [2006] is a regularization technique that uses the sum of component norms as the penalty function. They also studied the convergence rate of the COSSO estimator. Meier et al. [2009] proposed a combined sparsity-smoothness penalty for high-dimensional generalized additive models, penalizing

both the sparsity and the roughness. Estimation and variable selection consistency results of group LASSO and adaptive group LASSO methods under the high-dimensional additive model were established by Huang et al. [2010].

Let the response variable be $\boldsymbol{Y} \in \mathbb{R}^n$ and the K-dimensional covariate vector $(X_1, \ldots, X_K)$ be collected for the n subjects and stored in the matrix $\boldsymbol{X} \in \mathbb{R}^{n \times K}$. Consider the model

$$Y_i = \sum_{k=1}^{K} f_k^0(X_{k,i}) + \varepsilon_i, \quad i = 1, 2, \ldots, n, \tag{4.1}$$

where $X_{k,i}$ denotes the covariate information corresponding to i-th subject for the k-th variable. Here, $f_k^0(\cdot)$ is unknown for all $j = 1, 2, \ldots, K$ and ε_i is the mean zero random noise with variance σ^2 for some $\sigma > 0$. We suppose that not all of these variables are important to the response. Thus, some f_k^0's are zero. The goal is to accurately and consistently identify these zero-additive components and estimate the non-zero components.

Define B_n as the number of B-spline basis functions used to expand each $f_k^0(x)$, $k = 1, 2, \ldots, K$. We approximate $f_k^0(X_{k,i})$ using the following B-spline expansion

$$f_k(X_{k,i}) = \sum_{j=1}^{B_n} \mathcal{B}_{k,j}(X_{k,i})\beta_{k,j}. \tag{4.2}$$

The group selection problem, as discussed earlier, aptly fits the goal here: to identify the set of spline coefficients that are significant in the model and discard the remaining variables. The model now becomes $\boldsymbol{Y} = \mathcal{B}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$, where $\boldsymbol{\beta} = (\boldsymbol{\beta}_1, \boldsymbol{\beta}_2, \ldots, \boldsymbol{\beta}_K)^{\mathrm{T}} \in \mathbb{R}^{B_n K}$, such that $\boldsymbol{\beta}_k \in \mathbb{R}^{B_n}$ for all $k = 1, 2, \ldots, K$. In this model, the dimension of the coefficient vector is $B_n K = p$, say. Then, we may have $p \gg n$ and p increasing with n. The i-th row of the matrix $\mathcal{B} \in \mathbb{R}^{n \times p}$ is $(\mathcal{B}_1^{\mathrm{T}}(X_{1,i}), \ldots, \mathcal{B}_K^{\mathrm{T}}(X_{K,i}))$, where $\mathcal{B}_k(X_{k,i}) \in \mathbb{R}^B$ for each $k = 1, 2, \ldots, K$. We write $\boldsymbol{\beta}^0$ as the true coefficient vector. Now, we propose to use the sparse projection-posterior method described in Section 2 to get the induced posterior of $\boldsymbol{\beta}^*$ and define $f_k^*(X_{k,i}) = \sum_{j=1}^{B_n} \mathcal{B}_{k,j}(X_{k,i})\beta_{k,j}^*$.

Theorem 4 (Consistency of Projection-Posteriors for Additive Model)**.** *Under the additive model, the group LASSO projection-posterior satisfies the following:*

1. *(Recovery rate) If* $\lambda_n \asymp \sqrt{\log(KB_n)/n}$ *and* $B_n \asymp n^{1/(1+2\alpha)}$*, then, under the same Assumptions as in Theorem 1, the projection-posteriors obtained by any of the three mappings satisfy* $\Pi(\|f^* - f^0\|^2 \geq c_5 s_0 n^{-2\alpha/(2\alpha+1)} \log p \big| \boldsymbol{Y}) \to 0$ *for* $c_5 > 0$ *as* $n \to \infty$*, in probability under the true distribution.*

2. *(Model selection consistency) Under the same Assumptions as in Theorem 2 and for* $B_n \asymp n^{1/(1+2\alpha)}$*, all the three proposed projection-posteriors are selection consistent for the respective choices of* λ_n*: for the gruop LASSO* $\lambda_n \asymp \sqrt{\log(KB_n)/n}$*, for the group SCAD* $\lambda_n = o\big(n^{-(1-c_3+c_4)/2}\big)$ *and* $KB_n/(\lambda_n n^{3/2}) \to 0$ *and for the adaptive group LASSO* $\log(K - s_0)/(\sqrt{n}\lambda_n r_n) \to 0$*.*

Theorem 4 shows that the posterior probability that the distance between the induced posterior of the coefficients in the spline representation of the nonparametric functions and the true model converges to zero in probability. The convergence rate is influenced by three primary factors: stochastic error in estimating the nonparametric components, spline approximation error, and bias introduced by penalization. Next, we show that the posterior probability that the three projection-posteriors select all nonzero additive components converges to 1 in probability under the true distribution. In the additive model, the need to estimate smooth functions introduces an additional error source resulting from the spline approximation; that is, the modified error $\boldsymbol{\varepsilon}^{(2)} = \boldsymbol{\varepsilon} + \boldsymbol{\delta}$, where $\boldsymbol{\delta} = (\delta_1, \ldots, \delta_n)^{\mathrm{T}}$, with $\delta_i = \sum_{k=1}^{K} \big(f_k^0(X_{k,i}) - f_k(X_{k,i})\big)$. Since for the j-th component we have $\|f_j^0 - f_j\| = \mathcal{O}\big(B_n^{-\alpha}\big) = \mathcal{O}\big(n^{-\alpha/(2\alpha+1)}\big)$, with $B_n = \mathcal{O}\big(n^{1/(2\alpha+1)}\big)$, it follows that $\|\delta_i\| = \|\sum_{j=1}^{s_0}(f_j^0 - f_j)\| \leq s_0\|f_j^0 - f_j\| = \mathcal{O}\big(s_0 n^{-\alpha/(2\alpha+1)}\big)$. Consequently, $\|\boldsymbol{\delta}\| = \sqrt{\sum_{i=1}^{n}\|\delta_i\|^2} =$

$\mathcal{O}\big(\sqrt{ns_0^2 n^{-2\alpha/(2\alpha+1)}}\big) = \mathcal{O}\big(s_0 n^{1/(4\alpha+2)}\big)$. Thus, to prove the model selection consistency result for the additive model, we need to adjust for this error and then proceed as in Theorem 2.

5. Numerical Results

5.1 Simulation Study 1: Grouped Linear Regression

For the first simulation study, we evaluated the performance of competing methods in high-dimensional grouped regression under different sparsity and dimensionality scenarios. Specifically, we considered four combinations of feature dimensions or the number of groups (K) and sample sizes (n): $(K, n) \in \{(50, 100), (100, 100), (50, 500), (100, 500)\}$. The number of active groups, indicating nonzero coefficients, was varied across $s_0 \in \{10, 20\}$, reflecting varying sparsity levels. Keeping an average of 10 variables per group, we get $p = 500$ when $K = 50$ and $p = 1000$ when $K = 100$, respectively, for the three (K, n) pairs. Eight competing methods were compared over 100 replications for each setting. The methods with which we compared our three proposed maps, namely group LASSO projection-posterior (GL-P), group SCAD projection-posterior (GS-P), and adaptive group LASSO projection-posterior (AGL-P), include group LASSO (GL), group SCAD (GS), group adaptive LASSO (AGL), debiased group LASSO, Bayesian group LASSO (BGL), and the spike-and-slab (SS). For Bayesian sparse projection-posterior methods, 5000 initial posterior samples were drawn, projected using the three proposed methods, and the median probability model (MPM) was selected as the chosen model. The Bayes estimator was selected as the posterior mean corresponding to the MPM, and the rest were set to 0. When computing the induced posterior samples using the three proposed maps, we use the penalty parameter obtained by cross-validation when regressing $\boldsymbol{Y}$ on $\boldsymbol{X}$ under each regularization approach. For other

Bayesian methods used for the purpose of comparison, we have drawn 10000 initial samples and allowed a burn in of 5000. For each replication, we recorded the Mean Squared Error (MSE) for all methods and presented the results as box plots in Figure 1, corresponding to $s = 10$. The results for $s = 20$ are similar and are displayed in **??** in the supplement. Additionally, we recorded the True Positives (TP) for each method. We measured the F1-score defined by $F_1 = (2 \times \text{Precision} \times \text{Recall})/(\text{Precision} + \text{Recall})$, where $\text{Precision} = \text{TP}/\hat{s}$ and $\text{Recall} = \text{TP}/s_0$ to assess the recovery of the true sparsity pattern. Here, $\hat{s}$ is the symbol used for the selected set of groups of each method. A value of $F_1 = 1$ indicates perfect precision and recall, meaning all detected signals were correct, and all actual signals were identified, whereas 0 indicates the worst performance, meaning the model did not correctly identify any signal instances. The F1-scores are visualized using forestplots (Figures 2 and **??** for $s = 10$ and 20, respectively) to demonstrate each method's ability to identify active groups across different scenarios accurately. This comprehensive setup allows us to evaluate the method's consistency and robustness under varying group structures and sparsity levels.

We observe from the boxplots presented in Figures 1 and **??** that the MSEs corresponding to all the methods reduce when the sample size increases from $n = 100$ (upper panels) to $n = 500$ (lower panels). We also notice a pattern among the penalization methods and their Bayesian projection counterparts. The MSEs for the maps using any of the three penalty functions are similar to those of the corresponding penalization methods in some cases. However, as n grows, the mean squared error of BGL is maximum. GL, AGL, GL-P, and AGL-P have more or less similar characteristics. Both GS and GS-P perform better in terms of estimation error than the other two penalties and maps. The estimation accuracies of GS and GS-P are similar to those of the spike-and-slab method. Although the MSE of

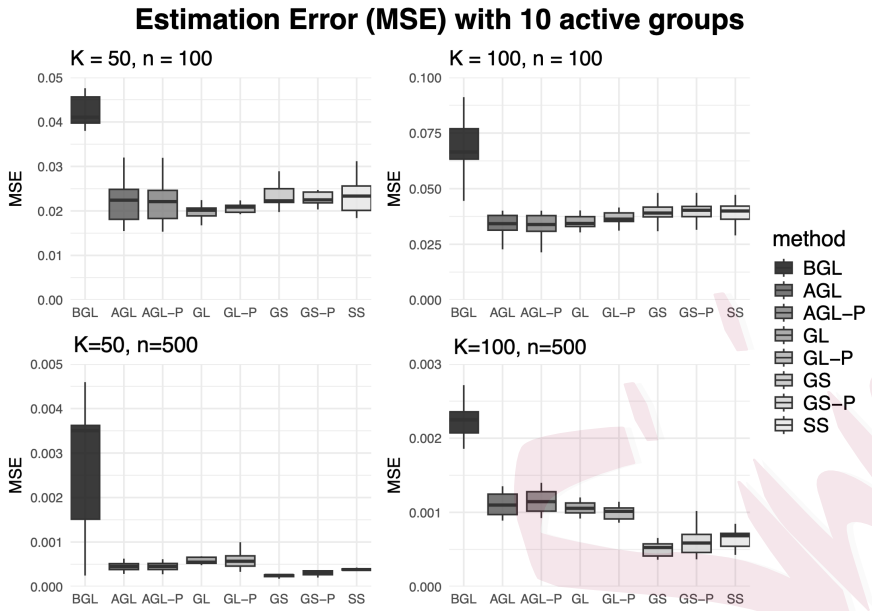

Figure 1: MSE comparisons for the four pairs of (K, n), namely (50,100),(50,500),(100,100) and (100,500) when $s_0 = 10$ groups are active, replicated 100 times

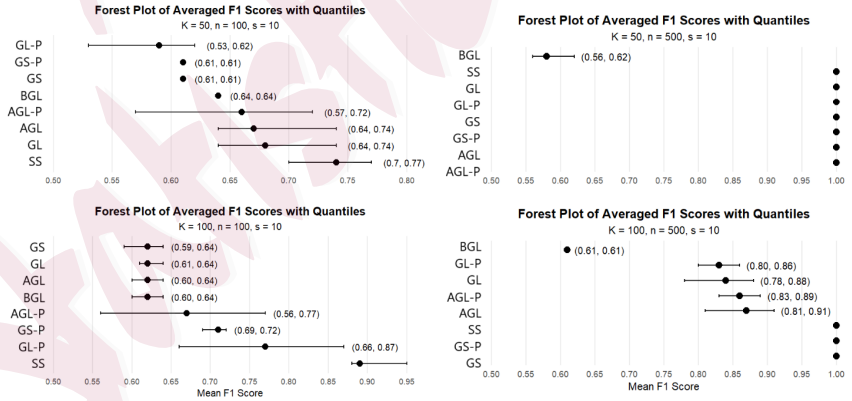

Figure 2: F1-score comparisons for the four pairs of (K, n), namely (50,100),(50,500),(100,100) and (100,500) when $s_0 = 10$ groups are active, replicated 100 times.

SS is lower, its running time is considerably higher. The run times of all these methods, averaged over 100 replications, are provided in Table 1.

Setting	GL	GS	AGL	BGL	SS	GL-P	GS-P	AGL-P
$K = 100$	2.7 s	2.5 s	3.1 s	1.4 m	5.8 m	2.3 m	3.4 m	2.9 m
	(0.5 s)	(0.4 s)	(0.8 s)	(0.5 m)	(2.7 m)	(0.7 m)	(1.1 m)	(0.9 m)
$K = 500$	4.3 s	5.2 s	4.2 s	4.1 m	18.8 m	6.8 m	8.1 m	7.2 m
	(1.5 s)	(1.9 s)	(2.0 s)	(2.8 m)	(5.6 m)	(2.3 m)	(3.0 m)	(2.2 m)

Table 1: Run-times averaged over 100 replicates with standard error for $n = 100$ and $s = 10$. Here s stands for seconds and m for minutes.

In Figures 2 and **??**, we plotted the average F1 scores for each of the 8 competing methods with their corresponding standard error bars for the for $s_0 = 10$ and $s_0 = 20$ respectively. Each of the plots contains four forestplots pertaining to the four settings, $(K, n) \in \{(50, 100), (50, 500), (100, 100), (100, 500)\}$. The F1 score, a harmonic mean of precision and recall, reflects the balance between identifying relevant groups and avoiding irrelevant ones. We observe that the F1 scores stabilize around 1 for most methods when n increases from 100 to 500 for both $s_0 = 10$ and $s_0 = 20$. We also see that GS and GS-P exhibit model selection properties similar to those of the spike-and-slab method. Although the performances of GL-P and AGL-P are similar to BGL when $n = 100$, their selection accuracy exceeds that of BGL when n increases to 500. The results corresponding to $s = 20$ are displayed in the supplement.

Next, we provide the average coverage of the Bayesian methods along with Debiased Group LASSO and Debiased Group LASSO Projection in Tables 2 and **??** respectively for the $s_0 = 10$ and $s_0 = 20$ setups. We used 95% quantile-based credible regions for each Bayesian method. We averaged across all signal and noise variables to get an idea of the coverage and length patterns for each separately. Among the three sparse projections we

implemented, GS-P yields accurate coverage for large samples ($n = 500$). GL-P and AGL-P fall slightly short of exact coverage. However, the debiased group LASSO map, denoted by the Debiased Projection method, achieves exact 95% coverage with $n = 500$. The same is observed for the (non-Bayesian) debiased group LASSO confidence interval. Since the discrepancy between the results for $s_0 = 10$ and $s_0 = 20$ is minimal, we have moved **??** to the supplement.

Setting		Debiased GL	BGL	SS	GL-P	GS-P	AGL-P	Debiased Projection
$K = 50$, $n = 100$	Signal Coverage	0.91(0.11)	0.53(0.03)	0.81(0.02)	0.72(0.16)	0.89(0.13)	0.79(0.17)	0.91(0.05)
	Signal Length	5.55(0.70)	1.41(0.21)	5.48(0.11)	5.33(0.12)	5.19(0.98)	5.19(0.89)	5.62(0.34)
	Noise Coverage	1.00	1.00	1.00	0.98(0.64)	0.95(0.12)	1.00	1.00
	Noise Length	2.18(0.12)	1.33(0.19)	3.05(0.07)	0.44(0.03)	0.41(0.04)	0.33(0.04)	2.21(0.21)
$K = 50$, $n = 500$	Signal Coverage	0.95(0.02)	0.87	0.91(0.03)	0.93(0.09)	0.97(0.10)	0.93(0.09)	0.96(0.01)
	Signal Length	1.72(0.01)	0.60(0.01)	0.60(0.01)	1.72(0.01)	0.82(0.10)	1.43(0.01)	1.60(0.08)
	Noise Coverage	1.00	1.00	1.00(0.00)	1.00	1.00	1.00	1.00
	Noise Length	1.62(0.01)	0.49(0.00)	0.11(0.00)	1.70(0.01)	0.71(0.33)	1.38(0.01)	1.57(0.01)
$K = 100$, $n = 100$	Signal Coverage	0.91(0.08)	0.40(0.01)	0.69(0.01)	0.65(0.01)	0.73(0.01)	0.72(0.06)	0.91(0.08)
	Signal Length	1.31(0.02)	0.61(0.01)	0.85(0.02)	0.82(0.00)	0.76(0.01)	1.02(0.01)	1.43(0.03)
	Noise Coverage	1.00	1.00	1.00	0.96(0.01)	1.00(0.00)	0.96(0.01)	1.00
	Noise Length	1.84(0.02)	1.58(0.09)	2.34(0.07)	0.31(0.00)	0.35(0.00)	0.46(0.00)	1.77(0.01)
$K = 100$, $n = 500$	Signal Coverage	0.95(0.00)	0.80(0.01)	0.88(0.02)	0.93(0.00)	0.95(0.00)	0.94(0.00)	0.95(0.01)
	Signal Length	1.65(0.01)	0.79(0.02)	0.88(0.03)	1.30(0.04)	1.26(0.03)	1.62(0.03)	1.56(0.02)
	Noise Coverage	1.00	1.00	1.00	1.00	1.00	1.00	1.00
	Noise Length	1.95(0.02)	0.88(0.08)	2.81(0.05)	0.59(0.01)	0.53(0.00)	0.55(0.00)	1.97(0.02)

Table 2: Average signal and noise coverage and interval length (standard errors in parentheses) over 100 replicates with $s_0 = 10$.

5.2 Simulation Study 2: Additive Regression

In this simulation study, we consider an additive nonparametric regression model, as described in (4.1). We simulate data with $n = 100$ observations and $k = 50$ predictor variables, where only $s_0 = 5$ variables are relevant to the response. These five important variables follow nonlinear relationships with the response. That is, $f_1^0(X_1) = 3\sin(X_1)$ is a sinusoidal

term, $f_2^0(X_2) = 2X_2^2$ is a quadratic term, $f_3^0(X_3) = -1.5X_3$ is a linear term with a negative effect, $f_4^0(X_4) = \exp(X_4)$ is an exponential term, and finally $f_5^0(X_5) = \log(|X_5|+1)$ is a logarithmic transformation, with Gaussian noise added to the response. To model the nonlinear effects across all predictors, we use a cubic B-spline basis expansion with $B_n = 8$ for each variable, yielding a transformed feature matrix in which a set of basis functions represents each variable. This new matrix, now of order 100×400, contains the expanded set of predictors, with the dimensions reflecting the number of observations and the total number of transformed features. Hence, we have $p = K \times B_n = 400$ in the model. Clearly, $\boldsymbol{\beta}_k \in \mathbb{R}^8$ for each $k = 1, 2, \ldots, 50$. The setup allows for applying group regularization techniques to select the true active groups of variables while shrinking the coefficients of irrelevant variables to zero. Here, each of the 50 variables can be treated as a group, with the B-spline expansions of the variables forming a set of related features. Under sparsity, we aim to identify the true active variables among the 50 and recover the few truly important nonlinear relationships that affect the response, while keeping the number of selected groups minimal. Although the additive model's estimation and selection consistency results have been proven only for the group LASSO projection-posterior, we demonstrate the performance of all three proposed projection maps and five other competing methods.

In Figure 3, the boxplots of the prediction MSEs of the different methods are presented in the left panel. We see that GS-P has the minimum MSE. The performances of other projection maps are again similar to those of their corresponding penalized regression techniques. On the right panel of Figure 3, we show the forestplot of the F1 scores for the eight competing methods. Again, we observe that the F1 score of the GS-P is as good as that of spikes and slabs. Thus, the projection methods demonstrate overall successful effectuation

of the asymptotic results proved in Section 4 on both estimation and selection fronts.

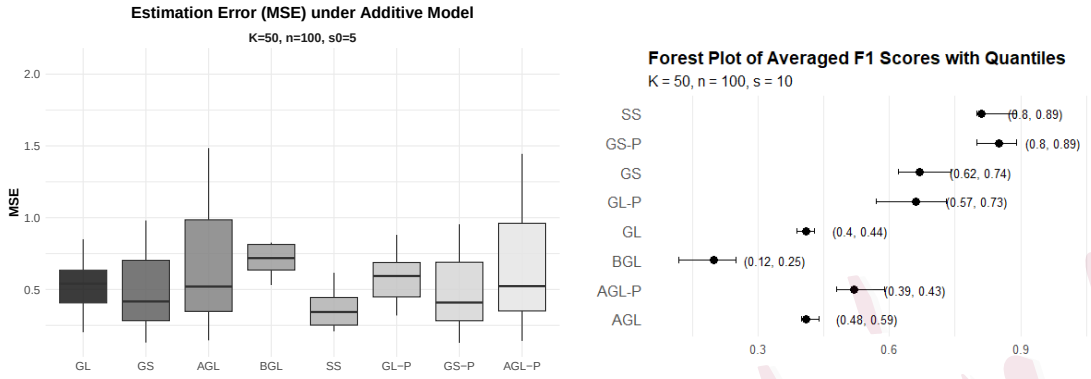

Figure 3: Boxplot of MSEs and Forestplot of F1 scores of all competing methods under the additive model with $K = 50, B_n = 8, n = 100$ and $s_0 = 5$.

5.3 Real Data Analysis

MRI-based volumetric measurements from Alzheimer's Disease Neuroimaging Initiative (ADNI), specifically those assessing regional brain volumes, are widely regarded as biomarkers for neurodegeneration and cognitive decline. In Alzheimer's disease (AD) research, structural changes in specific brain regions, such as the hippocampus and entorhinal cortex, are known to correlate with the progression from cognitively normal (CN) status to mild cognitive impairment (MCI) and eventually AD. By examining brain volume data, researchers can quantify these structural changes and better understand the associations between brain atrophy patterns and clinical outcomes. Using an additive model in this setting is highly appropriate, as the relationship between brain volumes and AD severity will likely be nonlinear. This approach allows for flexible modeling of complex relationships between volumetric features and cognitive outcomes, accommodating potential nonlinear effects of brain regions on disease risk. We then use the B-spline basis expansion and the group selection procedures proposed

in this paper, along with existing methods, to identify the groups of features (e.g., brain regions) that are collectively relevant for predicting AD severity. In this study, we have used the Mini-Mental State Examination (MMSE) score as the indicator of Alzheimer's disease and, consequently, as the response variable of interest. This analytical framework aligns well with clinical research goals in AD, aiming to identify which brain regions and patterns of atrophy are most predictive of disease severity. By leveraging the power of grouped regularization, this study can provide insights into the neuroanatomical correlates of AD, potentially aiding in early diagnosis and targeted interventions. In this study, a total of $K = 70$ brain regions were considered as predictors for $n = 458$ subjects, each predictor corresponding to a distinct anatomical region of interest (ROI) such as hippocampal and parahippocampal subfields, entorhinal and parahippocampal cortices, and neighboring temporal regions. Each region was modeled nonparametrically using a cubic B-spline basis with $B_n = 8$ basis functions, resulting in a total of $p = K \times B_n = 560$ predictors. Thus, each brain region formed a group of eight correlated spline-expanded variables, collectively representing the smooth volumetric effect of that region on the response. We note that the real data analysis uses only baseline neuroimaging-derived phenotypes (IDPs) from the ADNI database, where volumetric measures of cortical and subcortical regions serve as predictors and the MMSE score as the response. Only baseline data were analyzed to focus on cross-sectional associations between brain structure and cognitive function. Age, sex, education, and intracranial volume (ICV) were included as covariates to adjust for subject-specific confounding effects.

Table 3 presents the average test prediction errors (with standard deviations) for different regularization and Bayesian approaches in a high-dimensional regression scenario. The dataset was split into training (70%) and test (30%) sets, and the reported values were

Methods	GL	GS	AGL	GL-P	GS-P	AGL-P	BGL	SS
Test Prediction	1.80	1.14	2.23	1.33	1.25	2.45	9.70	1.27
Error	(0.36)	(0.44)	(0.03)	(0.56)	(0.47)	(0.16)	(2.22)	(0.03)

Table 3: Average Test Prediction Error for the different methods applied to the ADNI MRI volume data.

averaged across 10 random splits. GS-P and GS methods achieved the lowest test errors, indicating their effectiveness in reducing prediction error under group-sparsity constraints. AGL-P and GL-P performed moderately well, balancing prediction error with model complexity. BGL had the highest test prediction error, suggesting it may not be as effective at achieving high prediction accuracy in this high-dimensional setup. Next, we discuss the brain regions (treated here as groups) selected by different methods as highly associated with the response variable (MMSE score).

The study examined a set of hippocampal and parahippocampal brain regions to identify those most associated with Alzheimer's disease severity. Specifically, the following regions were considered: Cornu Ammonis 1 (CA1), CA2, CA3, Dentate Gyrus (DG), Miscellaneous Regions (MISC), Subiculum (SUB), Entorhinal Cortex (ERC), Brodmann Area 35 (BA35), BA36, Parahippocampal Cortex (PHC), Sulcus, and a combined CA (CA) region. These regions, bilaterally present in the brain, play critical roles in memory formation, spatial navigation, and early neurodegenerative changes, particularly CA1 and ERC, which are heavily implicated in Alzheimer's pathology [Masurkar, 2018, Rao et al., 2022, Igarashi, 2023]. For simplicity, each region was abbreviated, such as Left CA1 for the left hemisphere CA1 region and similarly for other regions.

Different sparse regression methods selected varying subsets of these regions, illustrat-

ing distinct selection patterns. Group LASSO selected Left CA2, Left CA3, Left MISC, Left PHC, Left Sulcus, Right SUB, Right ERC, Right BA35, Right BA36, Right PHC, Right Sulcus, and Right CA, indicating these areas as particularly relevant to Alzheimer's severity. AGL largely aligned with GL but included Left BA36, highlighting its potential significance. GS, on the other hand, captured a distinct subset, including Left CA2, Left CA3, Left BA35, Left PHC, Left Sulcus, Right CA1, Right SUB, Right BA35, Right PHC, and Right CA, reflecting some divergence in selection preferences for certain regions on the right hemisphere. Bayesian-inspired methods showed similar patterns. GL-P and AGL-P selected the same regions as GL, demonstrating consistency in variable selection across posterior models. However, GS-P identified Left CA2, Left CA3, Left MISC, Left BA36, Left PHC, Left Sulcus, Right SUB, Right ERC, Right PHC, and Right Sulcus, diverging slightly from GS by emphasizing a few different regions. Both SS and BGL selected the complete set of regions, suggesting a tendency towards less selectivity in high-dimensional settings. In summary, CA2, CA3, SUB, ERC, BA35, BA36, and PHC consistently emerged across models, underscoring their potential role in Alzheimer's-related neurodegeneration. This analysis reveals that a sparse projection-posterior performs effective variable selection, comparable to traditional penalization methods, and also provides quantifiable uncertainty about the selected variables. The fact that we get the entire posterior distribution for the three sparse projection-posteriors is depicted in Figure 4. This plot shows that the in-sample prediction errors remain essentially constant for different sparse projection-posterior methods. Moreover, the density of the lines for each selected model reflects the number of models that show up with the given sparsity, thereby giving us an idea about which models appeared more frequently in the posterior.

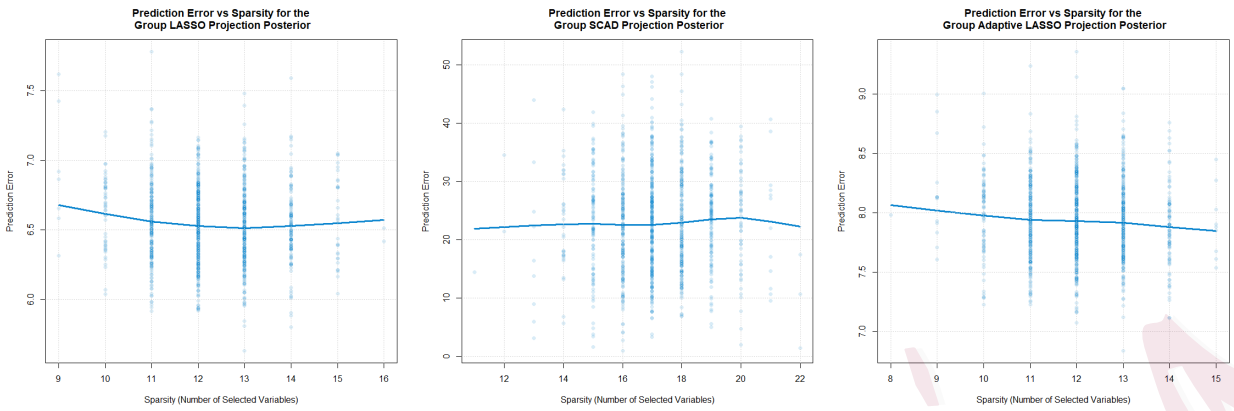

Figure 4: Plot of the distributions of the average prediction error versus the model sparsity for the three Bayesian projection-posterior methods.

6. Discussion

The sparse projection-posterior method is a relatively new Bayesian inference procedure with immense potential for high-dimensional models under sparsity. In this paper, we discussed several sparse projection-posterior methods for high-dimensional grouped regression. We demonstrate how different penalty functions can be used to map non-sparse posterior samples to group-sparse samples from the induced posterior. We also showed that the theoretical properties of the induced posterior depend on the chosen map. Thus, the sparse-projection-posterior method can reproduce the results of the corresponding regularization technique, with the added advantage of automatically quantifying uncertainty. The coverage properties of this approach have also been discussed. The extensive simulation studies suggest no loss in estimation and variable selection when the Bayesian regularized maps are used instead of their classical counterparts. Moreover, the proposed methods allow automatic quantification of the associated uncertainty. Compared to commonly used Bayesian methods in the grouped regression setup, we showed that all three proposed projection maps outperform the Bayesian

group LASSO in terms of estimation and selection accuracy, as well as coverage properties. Although almost on par with the proposed methods, the spike-and-slab Bayesian method is computationally costly in high dimensions. In conclusion, our findings highlight the sparse projection-posterior method as a powerful alternative to standard regularization techniques and traditional Bayesian approaches. This work paves the way for a broader exploration of Bayesian approaches to sparse high-dimensional models, laying the foundation for theoretical advances and practical applications.

7. Acknowledgement

Data collection and sharing for the Alzheimer's Disease Neuroimaging Initiative (ADNI) is funded by the National Institute on Aging (National Institutes of Health Grant U19 AG024904). The grantee organization is the Northern California Institute for Research and Education.

8. Funding

This research is partially supported by ARO grant number 76643MA 2020-0945. The research of the second author was partially supported by NSF grant number DMS-2609931.

References

S Derin Babacan, Shinichi Nakajima, and Minh N Do. Bayesian group-sparse modeling and variational inference. *IEEE Transactions on Signal Processing*, 62(11):2906–2921, 2014.

Prithwish Bhaumik and Subhashis Ghosal. Bayesian two-step estimation in differential equation models. *Electronic Journal of Statistics*, 9(2):3124–3154, 2015.

Peter J Bickel, Ya'acov Ritov, and Alexandre B Tsybakov. Simultaneous analysis of Lasso and Dantzig selector. *The Annals of Statistics*, 37(4):1705–1732, 2009.

Peter Bühlmann and Sara van de Geer. *Statistics for High-Dimensional Data: Methods, Theory and Applications*. Springer Science & Business Media, 2011.

George Casella, Malay Ghosh, Jeff Gill, and Minjung Kyung. Penalized regression, standard errors, and Bayesian lassos. *Bayesian Analysis*, 5(2):369–411, 2010.

Moumita Chakraborty and Subhashis Ghosal. Coverage of credible intervals in nonparametric monotone regression. *The Annals of Statistics*, 49(2):1011–1028, 2021.

Jianqing Fan and Runze Li. Variable selection via nonconcave penalized likelihood and its oracle properties. *Journal of the American Statistical Association*, 96(456):1348–1360, 2001.

Alessio Farcomeni. Bayesian constrained variable selection. *Statistica Sinica*, 20:1043–1062, 2010.

Edward I George and Robert E McCulloch. Stochastic search variable selection. *Markov chain Monte Carlo in Practice*, 68(1):203–214, 1995.

Xiao Guo, Hai Zhang, Yao Wang, and Jiang-Lun Wu. Model selection and estimation in high dimensional regression models with group scad. *Statistics & Probability Letters*, 103: 86–92, 2015.

Toshio Honda. The de-biased group lasso estimation for varying coefficient models. *Annals of the Institute of Statistical Mathematics*, 73:3–29, 2021.

Jian Huang, Shuange Ma, Huiliang Xie, and Cun-Hui Zhang. A group bridge approach for variable selection. *Biometrika*, 96(2):339–355, 2009.

Jian Huang, Joel L Horowitz, and Fengrong Wei. Variable selection in nonparametric additive models. *The Annals of Statistics*, 38(4):2282–2313, 2010.

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

Kei M Igarashi. Entorhinal cortex dysfunction in alzheimer's disease. *Trends in neurosciences*, 46(2):124–136, 2023.

Adel Javanmard and Andrea Montanari. Model selection for high-dimensional regression under the generalized irrepresentability condition. *Advances in Neural Information Processing Systems*, 26, 2013.

Michael Komodromos, Marina Evangelou, Sarah Filippi, and Kolyan Ray. Group spike and slab variational Bayes. *arXiv preprint arXiv:2309.10378*, 2023.

Wei-Ting Lai and Ray-Bing Chen. A review of Bayesian group selection approaches for linear regression models. *Wiley Interdisciplinary Reviews: Computational Statistics*, 13 (4):e1513, 2021.

Yi Lin and Hao Helen Zhang. Component selection and smoothing in smoothing in multivariate nonparametric regression. *The Annals of Statistics*, 38(5):2272–2297, 2006.

Arjun V Masurkar. Towards a circuit-level understanding of hippocampal ca1 dysfunction

in alzheimer's disease across anatomical axes. *Journal of Alzheimer's disease & Parkinsonism*, 8(1):412, 2018.

Lukas Meier, Sara van de Geer, and Peter Bühlmann. High-dimensional additive modeling. *The Annals of Statistics*, 37(6(B)):3779–3821, 2009.

Yuval Nardi and Alessandro Rinaldo. On the asymptotic properties of the group lasso estimator for linear models. *Electronic Journal of Statistics*, 9(2):605–633, 2008.

Samhita Pal and Subhashis Ghoshal. Bayesian high-dimensional linear regression with sparse projection-posterior. *arXiv preprint arXiv:2410.16577*, 2024.

Trevor Park and George Casella. The Bayesian lasso. *Journal of the American Statistical Association*, 103(482):681–686, 2008.

Sudhir Raman, Thomas J Fuchs, Peter J Wild, Edgar Dahl, and Volker Roth. The Bayesian group-lasso for analyzing contingency tables. In *Proceedings of the 26th Annual International Conference on Machine Learning*, pages 881–888, 2009.

Y Lakshmisha Rao, B Ganaraja, BV Murlimanju, Teresa Joy, Ashwin Krishnamurthy, and Amit Agrawal. Hippocampus and its involvement in alzheimer's disease: a review. *3 Biotech*, 12(2):55, 2022.

Pradeep Ravikumar, John Lafferty, Han Liu, and Larry Wasserman. Sparse additive models. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 71(5):1009–1030, 2009.

Niranjan Subrahmanya and Yung C Shin. A variational Bayesian framework for group feature selection. *International Journal of Machine Learning and Cybernetics*, 4:609–619, 2013.

Tingni Sun and Cun-Hui Zhang. Scaled sparse linear regression. *Biometrika*, 99(4):879–898, 2012.

Sara van de Geer, Peter Bühlmann, Ya'acov Ritov, and Ruben Dezeure. On asymptotically optimal confidence regions and tests for high-dimensional models. *The Annals of Statistics*, 42(3):1166–1202, 2014.

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

Kang Wang and Subhashis Ghosal. Coverage of credible intervals in Bayesian multivariate isotonic regression. *The Annals of Statistics*, 51(3):1376–1400, 2023.

Fengrong Wei and Jian Huang. Consistent group selection in high-dimensional linear regression. *Bernoulli*, 16(4):1369, 2010.

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

Zemei Xu, Daniel F Schmidt, Enes Makalic, Guoqi Qian, and John L Hopper. Bayesian grouped horseshoe regression with application to additive models. In *AI 2016: Advances in Artificial Intelligence: 29th Australasian Joint Conference, Hobart, TAS, Australia, December 5-8, 2016, Proceedings 29*, pages 229–240. Springer, 2016.

Xinming Yang and Naveen N Narisetty. Consistent group selection with Bayesian high dimensional modeling. *Bayesian Analysis*, 15(3):909–935, 2020.

Ming Yuan and Yi Lin. Model selection and estimation in regression with grouped variables.

Journal of the Royal Statistical Society Series B: Statistical Methodology, 68(1):49–67, 2006.

Lingmin Zeng and Jun Xie. Group variable selection via scad-l 2. *Statistics*, 48(1):49–66, 2014.

Cun-Hui Zhang. Nearly unbiased variable selection under minimax concave penalty. *Annals of Statistics*, 38(2):894–942, 2010.

Cun-Hui Zhang and Jian Huang. The sparsity and bias of the lasso selection in high-dimensional linear regression. *The Annals of Statistics*, 36(4):1567–1594, 2008.

Peng Zhao and Bin Yu. On model selection consistency of lasso. *The Journal of Machine Learning Research*, 7:2541–2563, 2006.

Peng Zhao, Guilherme Rocha, and Bin Yu. The composite absolute penalties family for grouped and hierarchical variable selection. *The Annals of Statistics*, 37(6A):3468–3497, 2009.