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

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

March 3, 2020

Abstract

In Section A, we provide the complete block coordinate ascent algorithm described in Section 3, additional implementation details (i.e. tuning hyperparameters, updating the variance parameter, and initializing the parameters in the algorithm), and details of the nodewise regression approach for estimating $\widehat{\mathbf{\Theta}}$ in Section 4 of the main manuscript. In Section B, we provide additional simulation results on the performance of the SSGL approach in both sparse and dense settings, its performance in estimation of the residual variance σ^2, and timing comparisons. In Section C, we provide results for data analyses on benchmark data sets where $p < n$ and additional discussion and analysis of the two real data sets in Section 8 of the main manuscript. In Section D, we provide the proofs for the theoretical results in the main manuscript.

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

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

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

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

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

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

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

A Additional Computational Details

A.1 SSGL Block-Coordinate Ascent Algorithm

Algorithm 1 Spike-and-Slab Group Lasso

Input: grid of increasing λ_0 values $I = \{\lambda_0^1, \ldots, \lambda_0^L\}$, update frequency M

Initialize: $\boldsymbol{\beta}^* = \mathbf{0}_p$, $\theta^* = 0.5$, σ^{*2} as described in Section A.2, Δ^* according to (3.11) in the main manuscript

For $l = 1, \ldots, L$:

1. Set iteration counter $k_l = 0$
2. Initialize: $\widehat{\boldsymbol{\beta}}^{(k_l)} = \boldsymbol{\beta}^*$, $\theta^{(k_l)} = \theta^*$, $\sigma^{(k_l)2} = \sigma^{*2}$, $\Delta^U = \Delta^*$
3. While $\mathsf{diff} > \varepsilon$
   (a) Increment k_l
   (b) For $g = 1, \ldots, G$:
      i. Update
      $$\boldsymbol{\beta}_g^{(k_l)} \leftarrow \frac{1}{n}\left(1 - \frac{\sigma^{(k_l)2}\lambda^*(\boldsymbol{\beta}_g^{(k_l-1)}; \theta^{(k_l)})}{\|\boldsymbol{z}_g\|_2}\right)_+ \boldsymbol{z}_g \; \mathbb{I}(\|\boldsymbol{z}_g\|_2 > \Delta^U)$$
      ii. Update
      $$\widehat{Z}_g = \begin{cases} 1 & \text{if } \boldsymbol{\beta}_g^{(k_l)} \neq \mathbf{0}_{m_g} \\ 0 & \text{otherwise} \end{cases}$$
      iii. If $g \equiv 0 \mod M$:
         A. Update
         $$\theta^{(k_l)} \leftarrow \frac{a + \sum_{g=1}^G \widehat{Z}_g}{a + b + G}$$
         B. If $k_{l-1} < 100$:
         $$\text{Update } \sigma^{(k_l)2} \leftarrow \frac{\|\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta}^{(k_l)}\|_2^2}{n+2}$$
         C. Update
         $$\Delta^U \leftarrow \begin{cases} \sqrt{2n\sigma^{(k_l)2}\log[1/p^*(\mathbf{0}_{m_g}; \theta^{(k_l)})]} + \sigma^{(k_l)2}\lambda_1 & \text{if } h(\mathbf{0}_{m_g}; \theta^{(k_l)}) > 0 \\ \sigma^{(k_l)2}\lambda^*(\mathbf{0}_{m_g}; \theta^{(k_l)}) & \text{otherwise} \end{cases}$$
      iv. $\mathsf{diff} = \|\boldsymbol{\beta}^{(k_l)} - \boldsymbol{\beta}^{(k_l-1)}\|_2$
4. Assign $\boldsymbol{\beta}^* = \boldsymbol{\beta}^{(k_l)}$, $\theta^* = \theta^{(k_l)}$, $\sigma^{*2} = \sigma^{2(k_l)}$, $\Delta^* = \Delta^U$

A.2 Tuning Hyperparameters, Initializing Values, and Updating the Variance in Algorithm 1

We keep the slab hyperparameter λ_1 fixed at a small value. We have found that our results are not very sensitive to the choice of λ_1. This parameter controls the variance of the slab component of the prior, and the variance must simply be large enough to avoid overshrinkage of important covariates. For the default implementation, we recommend fixing $\lambda_1 = 1$. This applies minimal shrinkage to the significant groups of coefficients and affords these groups the ability to escape the pull of the spike.

Meanwhile, we choose the spike parameter λ_0 from an increasing ladder of values. We recommend selecting $\lambda_0 \in \{1, 2, ..., 100\}$, which represents a range from hardly any penalization to very strong penalization. Below, we describe precisely how to tune λ_0. To account for potentially different group sizes, we use the same λ_0 for all groups but multiply λ_0 by $\sqrt{m_g}$ for each gth group, $g = 1, \ldots, G$. As discussed in Huang et al. (2012), further scaling of the penalty by group size is necessary in order to ensure that the same degree of penalization is applied to potentially different sized groups. Otherwise, larger groups may be erroneously selected simply because they are larger (and thus have larger ℓ_2 norm), not because they contain significant entries.

When the spike parameter λ_0 is very large, the continuous spike density approximates the point-mass spike. Consequently, we face the computational challenge of navigating a highly multimodal posterior. To ameliorate this problem for the spike-and-slab lasso, Ročková and George (2018) recommend a "dynamic posterior exploration" strategy in which the slab parameter λ_1 is held fixed at a small value and λ_0 is gradually increased along a grid of values. Using the solution from a previous λ_0 as a "warm start" allows the procedure to more easily find optimal modes. In particular, when $(\lambda_1 - \lambda_0)^2 \leq 4$, the

posterior is convex.

Moran et al. (2019) modify this strategy for the unknown σ^2 case. This is because the posterior is always non-convex when σ^2 is unknown. Namely, when $p \gg n$ and $\lambda_0 \approx \lambda_1$, the model can become saturated, causing the residual variance to go to zero. To avoid this suboptimal mode at $\sigma^2 = 0$, Moran et al. (2019) recommend fixing σ^2 until the λ_0 value at which the algorithm starts to converge in less than 100 iterations. Then, $\boldsymbol{\beta}$ and σ^2 are simultaneously updated for the next largest λ_0 in the sequence. The intuition behind this strategy is we first find a solution to the convex problem (in which σ^2 is fixed) and then use this solution as a warm start for the non-convex problem (in which σ^2 can vary).

We pursue a similar "dynamic posterior exploration" strategy with the modification for the unknown variance case for the SSGL in Algorithm 1 of Section A.1. A key aspect of this algorithm is how to choose the maximum value of λ_0. Ročková and George (2018) recommend this maximum to be the λ_0 value at which the estimated coefficients stabilize. An alternative approach is to choose the maximum λ_0 using cross-validation, a strategy which is made computationally feasible by the speed of our block coordinate ascent algorithm. In our experience, the dynamic posterior exploration strategy favors more parsimonious models than cross-validation. In the simulation studies in Section 7, we utilize cross-validation to choose λ_0, as there, our primary goal is predictive accuracy rather than parsimony.

Following Moran et al. (2019), we initialize $\boldsymbol{\beta}^* = \mathbf{0}_p$ and $\theta^* = 0.5$. We also initialize σ^{*2} to be the mode of a scaled inverse chi-squared distribution with degrees of freedom $\nu = 3$ and scale parameter chosen such that the sample variance of $\boldsymbol{Y}$ corresponds to the 90th quantile of the prior. We have found this initialization to be quite effective in practice at ensuring that Algorithm 1 converges in less than 100 iterations for sufficiently large λ_0.

A.3 Additional Details for the Inference Procedure

Here, we describe the nodewise regression procedure for estimating $\widehat{\boldsymbol{\Theta}}$ in Section 4. This approach for estimating the inverse of the covariance matrix $\widehat{\boldsymbol{\Sigma}} = \boldsymbol{X}^T\boldsymbol{X}/n$ was originally proposed and studied theoretically in Meinshausen and Bühlmann (2006) and van de Geer et al. (2014).

For each $j = 1, \ldots, p$, let $\boldsymbol{X}_j$ denote the jth column of $\boldsymbol{X}$ and $\boldsymbol{X}_{-j}$ denote the submatrix of $\boldsymbol{X}$ with the jth column removed. Define $\widehat{\boldsymbol{\gamma}}_j$ as

$$\widehat{\boldsymbol{\gamma}}_j = \arg\min_{\gamma}(||\boldsymbol{X}_j - \boldsymbol{X}_{-j}\boldsymbol{\gamma}||_2^2/n + 2\lambda_j||\boldsymbol{\gamma}||_1).$$

Now we can define the components of $\widehat{\boldsymbol{\gamma}}_j$ as $\widehat{\boldsymbol{\gamma}}_{j,k}$ for $k = 1, \ldots, p$ and $k \neq p$, and create the following matrix:

$$\widehat{\boldsymbol{C}} = \begin{pmatrix} 1 & -\widehat{\boldsymbol{\gamma}}_{1,2} & \ldots & -\widehat{\boldsymbol{\gamma}}_{1,p} \\ -\widehat{\boldsymbol{\gamma}}_{2,1} & 1 & \ldots & -\widehat{\boldsymbol{\gamma}}_{2,p} \\ \vdots & \vdots & \ddots & \vdots \\ -\widehat{\boldsymbol{\gamma}}_{p,1} & -\widehat{\boldsymbol{\gamma}}_{p,2} & \ldots & 1 \end{pmatrix}.$$

Lastly, let $\widehat{\boldsymbol{T}}^2 = \text{diag}(\widehat{\tau}_1^2, \widehat{\tau}_2^2, \ldots, \widehat{\tau}_p^2)$, where

$$\widehat{\tau}_j = ||\boldsymbol{X}_j - \boldsymbol{X}_{-j}\widehat{\boldsymbol{\gamma}}_j||_2^2/n + \lambda_j||\widehat{\boldsymbol{\gamma}}_j||_1.$$

We can proceed with $\widehat{\boldsymbol{\Theta}} = \widehat{\boldsymbol{T}}^{-2}\widehat{\boldsymbol{C}}$. This choice is used because it puts an upper bound on $||\widehat{\boldsymbol{\Sigma}}\widehat{\boldsymbol{\Theta}}_j^T - \boldsymbol{e}_j||_\infty$. Other regression models such as the original spike-and-slab lasso (Ročková and George, 2018) could be used instead of the lasso (Tibshirani, 1996) regressions for each covariate. However, we will proceed with this choice, as it has already been studied theoretically and shown to have the required properties to be able to perform inference for $\boldsymbol{\beta}$.

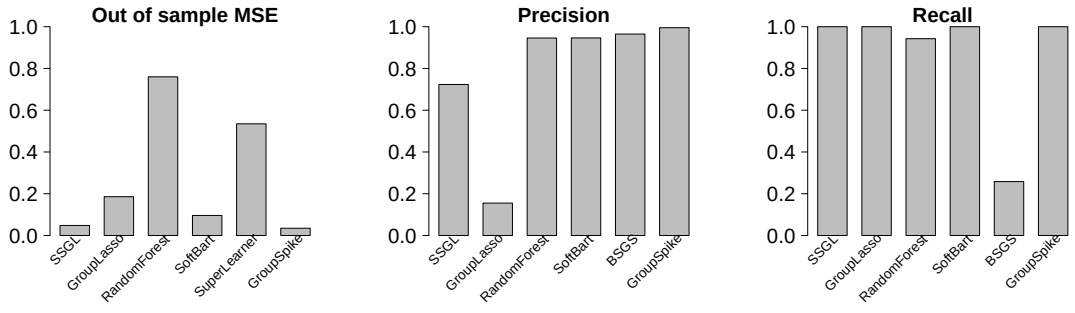

Figure 1: Simulation results from the sparse setting with $n = 300$. The left panel presents the out-of-sample mean squared error, the middle panel shows the precision score, and the right panel shows the recall score. The MSE for BSGS is not displayed as it lies outside of the plot area.

B Additional Simulation Results

Here, we present additional results which include different sample sizes than those seen in the manuscript, assessment of the SSGL procedure under dense settings, estimates of σ^2, timing comparisons, and additional figures.

B.1 Increased Sample Size for Sparse Simulation

Here, we present the same sparse simulation setup as that seen in Section 7.1, though we will increase n from 100 to 300. Figure 1 shows the results and we see that they are very similar to those from the manuscript, except that the mean squared error (MSE) for the SSGL approach is now nearly as low as the MSE for the GroupSpike approach, and the precision score has improved substantially.

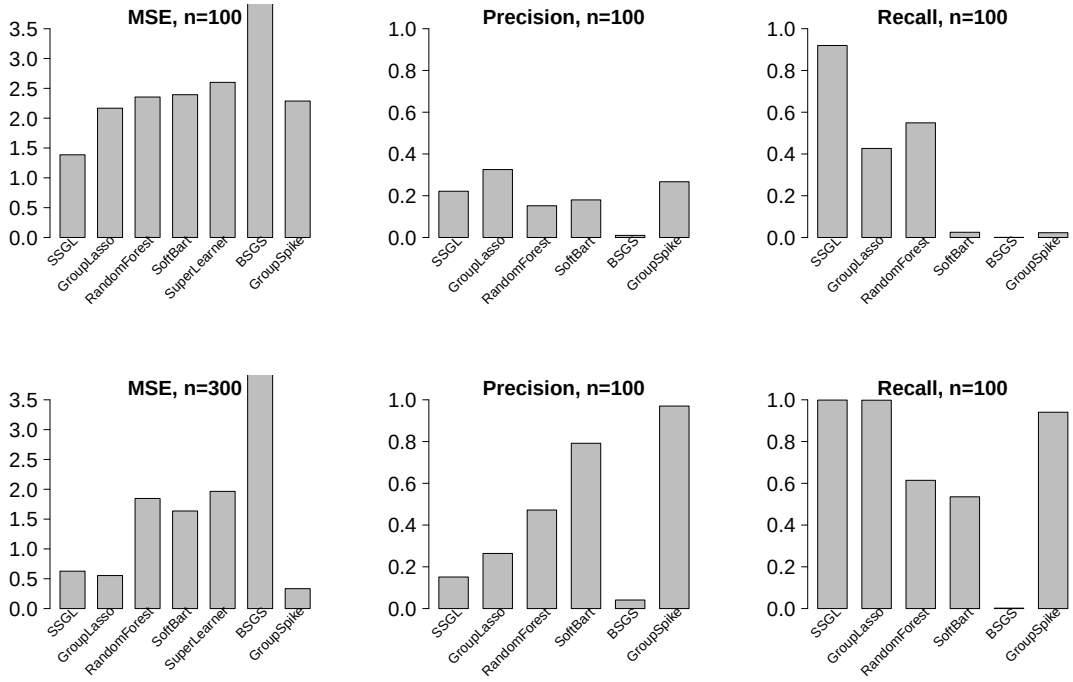

Figure 2: Simulation results from the less sparse setting with $n = 100$ and $n = 300$. The left column shows out-of-sample MSE, the middle panel shows the precision score, and the right column shows the recall score.

B.2 Dense Model

Here, we generate independent covariates from a standard normal distribution, and we let the true regression surface take the following form

$$\mathbb{E}(Y|\boldsymbol{X}) = \sum_{j=1}^{20} 0.2X_j + 0.2X_j^2,$$

with variance $\sigma^2 = 1$. In this model, there are no strong predictors of the outcome, but rather a large number of predictors which have small impacts on the outcome. Here, we

display results for both $n = 100$ and $p = 300$, as well as $n = 300$ and $p = 300$, as the qualitative results change across the different sample sizes. Our simulation results can be seen in Figure 2. When the sample size is 100, the SSGL procedure performs the best in terms of both MSE and recall score, while all approaches do poorly with the precision score. When the sample size increases to 300, the SSGL approach still performs quite well in terms of MSE and recall, though the GroupLasso and GroupSpike approaches are slightly better in terms of MSE. The SSGL approach still maintains a low precision score while the GroupSpike approach has a very high precision once the sample size is large enough.

B.3 Estimation of σ^2

To evaluate our ability to estimate σ^2 and confirm our theoretical results that the posterior of σ^2 contracts around the true parameter, we ran a simulation study using the following data generating model:

$$\mathbb{E}(Y|\boldsymbol{X}) = 0.5X_1 + 0.3X_2 + 0.6X_{10}^2 - 0.2X_{20},$$

with $\sigma^2 = 1$. We vary $n \in \{50, 100, 500, 1000, 2000\}$ and we set $G = n$ to confirm that the estimates are centering around the truth as both the sample size and covariate dimension grows. We use groups of size two that contain both the linear and quadratic term for each covariate. Note that in this setting, the total number of regression coefficients actually *exceeds* the sample size since each group has two terms, leading to a total of $p = 2G$ coefficients in the model.

Figure 3 shows box plots of the estimates for σ^2 across all simulations for each sample size and covariate dimension. We see that for small sample sizes there are some estimates well above 1 or far smaller than 1. This is because either some important variables are

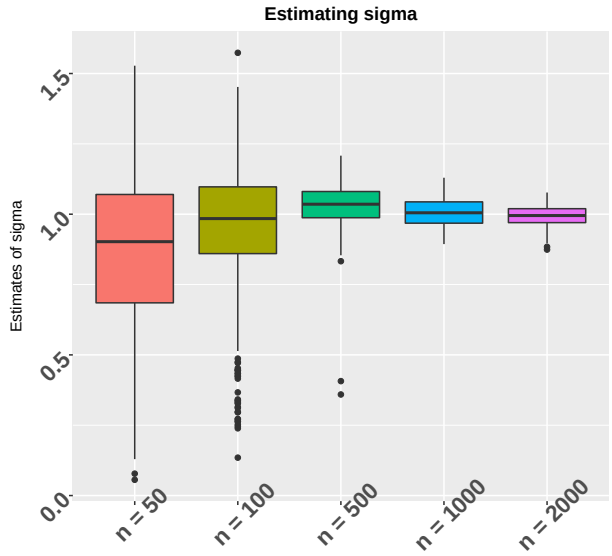

Figure 3: Boxplots of the estimates of σ^2 from the SSGL model for a range of sample sizes. Note that $n = G$ in each scenario.

excluded (so the sum of squared residuals gets inflated), or too many variables are included and the model is overfitted (leading to small $\widehat{\sigma}^2$). These problems disappear as the sample size grows to 500 or larger, where we observe that the estimates are closely centering around the true $\sigma^2 = 1$. Figure 3 confirms our theoretical results in Theorem 2 and Theorem 4 of the main manuscript, which state that as $n, G \to \infty$, the posterior $\pi(\sigma^2|\boldsymbol{Y})$ contracts around the true σ^2.

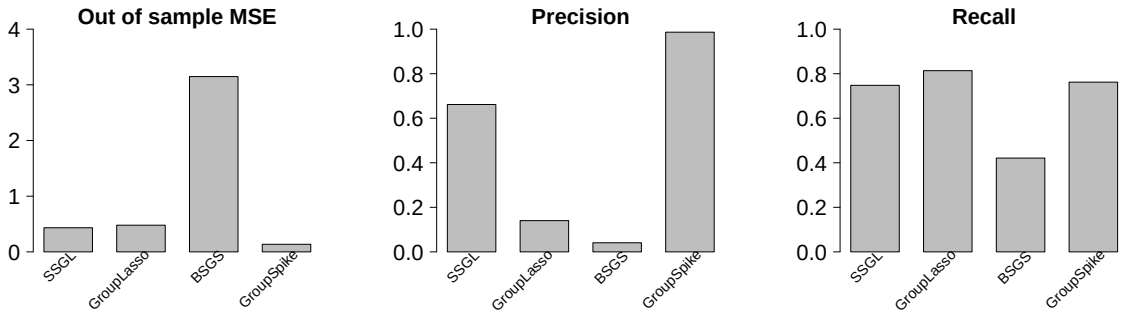

Figure 4: Simulation results from the many groups setting with $G = 2000$. The left panel presents the out-of-sample mean squared error, the middle panel shows the precision score, and the right panel shows the recall score.

B.4 Large Number of Groups

We now generate data with $n = 200$ and $G = 2000$, where each group contains three predictors. We generate data from the following model:

$$\mathbb{E}(\boldsymbol{Y}|\boldsymbol{X}) = \sum_{g=1}^{G} \boldsymbol{X}_g \boldsymbol{\beta}_g,$$

where we set $\boldsymbol{\beta}_g = \mathbf{0}$ for $g = 1, \ldots 1996$. For the final four groups, we draw individual coefficient values from independent normal distributions with mean 0 and standard deviation 0.4. These coefficients are redrawn for each data set in the simulation study, and therefore, the results are averaging over many possible combinations of magnitudes for the true nonzero coefficients. We see that the best performing approach in this scenario is the GroupSpike approach, followed by the SSGL approach. The SSGL approach outperforms group lasso in terms of MSE and precision, while group lasso has a slightly higher recall score.

B.5 Computation Time

In this study, we evaluate the computational speed of the SSGL procedure in comparison with the fully Bayesian GroupSpike approach that places point-mass spike-and-slab priors on groups of coefficients. We fix $n = 300$ and vary the number of groups $G \in \{100, 200, \ldots, 2000\}$, with two elements per group. For the SSGL approach, we keep track of the computation time for estimating the model for $\lambda_0 = 20$. For large values of λ_0, it typically takes 100 or fewer iterations for the SSGL method to converge. For the GroupSpike approach, we keep track of the computation time required to run 100 MCMC iterations. Both SSGL and GroupSpike models were run on an Intel E5-2698v3 processor.

In any given data set, the computation time will be higher than the numbers presented here because the SSGL procedure typically requires fitting the model for multiple values of λ_0, while the GroupSpike approach will likely take far more than 100 MCMC iterations to converge, especially in higher dimensions. Nonetheless, this should provide a comparison of the relative computation speeds for each approach.

The average CPU time in seconds can be found in Figure 5. We see that the SSGL approach is much faster as it is able to estimate all the model parameters for a chosen λ_0 in just a couple of seconds, even for $G = 2000$ (or $p = 4000$). When $p = 2000$, the SSGL returned a final solution in roughly three seconds on average, whereas GroupSpike required over two minutes to run 100 iterations (and would most likely require many more iterations to converge). This is to be expected as the GroupSpike approach relies on MCMC. Figure 5 shows the large computational gains that can be achieved using our MAP finding algorithm.

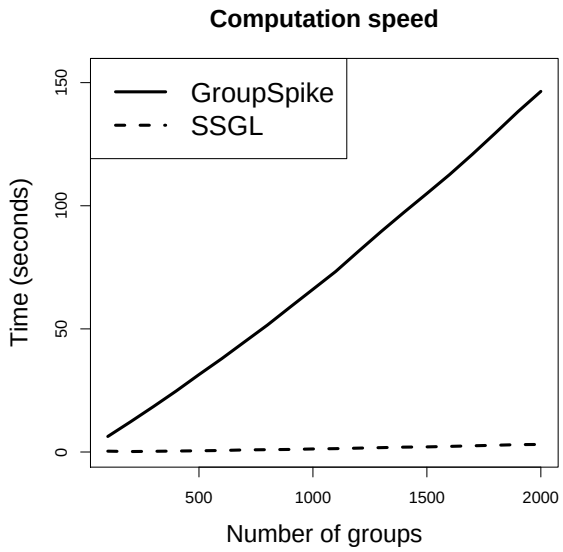

Figure 5: CPU time for the SSGL and GroupSpike approaches averaged across 1000 replications for fixed $n = 300$ and different group sizes G.

C Additional Results and Discussion for Real Data Examples

In this section, we perform additional data analysis of the SSGL method on benchmark datasets where $p < n$ to demonstrate that the SSGL model also works well in low-dimensional settings. We also provide additional analyses and discussion of the two real data examples analyzed in Section 8 of the main manuscript.

Data	SSGL	GroupLasso	RandomForest	SoftBart	SuperLearner	BSGS	GroupSpike
Tecator 1	1.41	1.57	2.75	1.93	1.00	5.16	1.67
Tecator 2	1.25	1.58	2.91	1.97	1.00	6.77	1.41
Tecator 3	1.14	1.38	1.94	1.81	1.10	3.31	1.00
BloodBrain	1.10	1.04	1.00	1.01	1.00	1.24	1.13
Wipp	1.44	1.30	1.46	1.00	1.17	4.68	1.30

Table 1: Standardized out-of-sample root mean squared prediction error averaged across 1000 replications for the data sets in Section C.1. An RMSE of 1 indicates the best performance within a data set.

C.1 Testing Predictive Performance of the SSGL on Datasets Where $p < n$

We first look at three data sets which have been analyzed in a number of manuscripts, most recently in Linero and Yang (2018). The tecator data set is available in the `caret` package in R (Kuhn, 2008) and has three different outcomes $\boldsymbol{Y}$ to analyze. Specifically, this data set looks at using 100 infrared absorbance spectra to predict three different features of chopped meat with a sample size of 215. The Blood-Brain data is also available in the `caret` package and aims to predict levels of a particular compound in the brain given 134 molecular descriptors with a sample size of 208. Lastly, the Wipp data set contains 300 samples with 30 features from a computer model used to estimate two-phase fluid flow (Storlie et al., 2011). For each of these data sets, we hold out 20 of the subjects in the data as a validation sample and see how well the model predicts the outcome in the held-out data. We repeat this 1000 times and compute the root mean squared error (RMSE) for prediction.

Table 1 shows the results for each of the methods considered in the simulation study. The results are standardized so that for each data set, the RMSE is divided by the minimum RMSE for that data set. This means that the model with an RMSE of 1 had the best predictive performance, and all others should be greater than 1, with the magnitude indicating how poor the performance was. We see that the top performer across the data sets was SuperLearner, which is not surprising given that SuperLearner is quite flexible and averages over many different prediction models. Our simulation studies showed that SuperLearner may not work as well when $p > n$. However, the data sets considered here all have $p < n$, which could explain its improved performance here. Among the other approaches, SSGL performs quite well as it has RMSE's relatively close to 1 for all the data sets considered.

C.2 Additional Details for Bardet-Biedl Analysis

Here we present additional results for the Bardet-Biedl Syndrome gene expression analysis conducted in Section 8.1 of the main manuscript. Table 2 displays the 12 probes found by SSGL. Table 3 displays the terms for which SSGL was enriched in a gene ontology enrichment analysis.

Probe ID	Gene Symbol	SSGL Norm	Group Lasso Norm
1374131_at		0.034	
1383749_at	Phospho1	0.067	0.088
1393735_at		0.033	0.002
1379029_at	Zfp62	0.074	
1383110_at	Klhl24	0.246	
1384470_at	Maneal	0.087	0.005
1395284_at		0.014	
1383673_at	Nap1l2	0.045	
1379971_at	Zc3h6	0.162	
1384860_at	Zfp84	0.008	
1376747_at		0.489	0.002
1379094_at		0.220	

Table 2: Probes found by SSGL on the Bardet-Biedl syndrome gene expression data set. The probes which were also found by the Group Lasso have nonzero group norm values.

	SSGL: enriched terms in gene ontology enrichment analysis
1	alpha-mannosidase activity
2	RNA polymerase II intronic transcription regulatory region sequence-specific DNA binding
3	mannosidase activity
4	intronic transcription regulatory region sequence-specific DNA binding
5	intronic transcription regulatory region DNA binding

Table 3: This table displays the terms for which SSGL was found to be enriched in a gene ontology enrichment analysis, ordered by statistical significance.

C.3 Additional Details for Analysis of NHANES Data

Here we will present additional results from the NHANES data analysis in Section 8.2 of the main manuscript. Here, the aim is to identify environmental exposures that are associated with leukocyte telomere length. In the NHANES data, we have measurements from 18 persistent organic pollutants. Persistent organic pollutants are toxic chemicals that have potential to adversely affect health. They are known to remain in the environment for long periods of time and can travel through wind, water, or even the food chain. Our data set consists of 11 polychlorinated biphenyls (PCBs), three Dioxins, and four Furans. We want to understand the impact that these can have on telomere length, and to understand if any of these pollutants interact in their effect on humans.

The data also contains additional covariates that we will adjust for such as age, a squared term for age, gender, BMI, education status, race, lymphocyte count, monocyte count, cotinine level, basophil count, eosinophil count, and neutrophil count. To better understand the data set, we have shown the correlation matrix between all organic pollutants and covariates in Figure 6. We can see that the environmental exposures are all fairly positively correlated with each other. In particular, the PCBs are highly correlated among themselves. The correlation across chemical types, such as the correlation between PCBs and Dioxins or Furans are lower, though still positively correlated. The correlation between the covariates that we place into our model and the exposures is generally extremely low, and the correlation among the individual covariates is also low, with the exception of a few blood cell types as seen in the upper right of Figure 6.

As discussed in Section 8.2 of the main manuscript, when we fit the SSGL model to this data set, we identified four main effects (plotted in Figure 3 of the main manuscript). Our model also identified six interactions as having nonzero parameter estimates. The identified

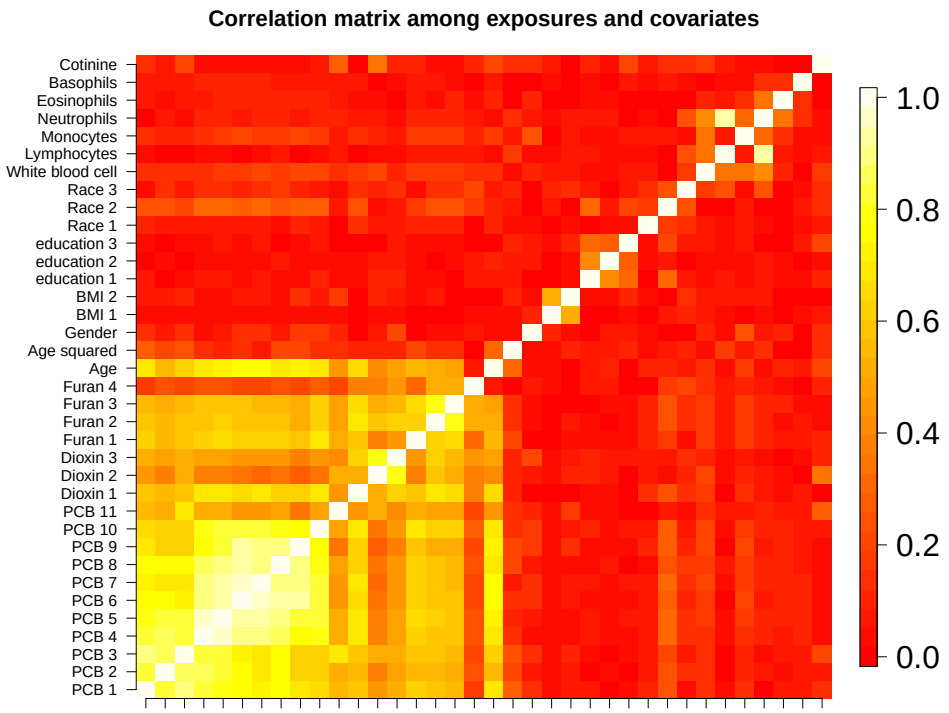

Figure 6: Correlation matrix among the 18 exposures and 18 demographic variables used in the final analysis for the NHANES study.

interactions are PCB 10 - PCB 7, Dioxin 1 - PCB 11, Dioxin 2 - PCB 2, Dioxin 2 - Dioxin 1, Furan 1 - PCB 10, and Furan 4 - Furan 3. We see that there are interactions both within a certain type of pollutant (Dioxin and Dioxin, etc.) and across pollutant types (Furan and PCB).

Lastly, looking at Figure 3 of the main manuscript, we can see that the exposure response curves for the four identified main effects are relatively linear, particularly for PCB 11 and Furan 1. With this in mind, we also ran our SSGL model with one degree of freedom splines for each main effect. Note that this does not require a model that

handles grouped covariate structures as the main effects and interactions in this case are both single parameters. Cross-validated error from the model with one degree of freedom is nearly identical to the model with two degrees of freedom, though the linear model selects far more terms. The linear model selects six main effect terms and 20 interaction terms. As the two models provide similar predictive performance but the model with two degrees of freedom is far more parsimonious, we elect to focus on the model with two degrees of freedom.

D Proofs of Main Results

D.1 Preliminary Lemmas

Before proving the main results in the paper, we first prove the following lemmas.

Lemma D.1. *Suppose that* $\boldsymbol{\beta}_g \in \mathbb{R}^{m_g}$ *follows a group lasso density indexed by* λ*, i.e.* $\boldsymbol{\beta}_g \sim \boldsymbol{\Psi}(\boldsymbol{\beta}_g|\lambda)$*. Then*

$$\mathbb{E}(\|\boldsymbol{\beta}_g\|_2^2) = \frac{m_g(m_g+1)}{\lambda^2}.$$

Proof. The group lasso density, $\boldsymbol{\Psi}(\boldsymbol{\beta}_g|\lambda)$, is the marginal density of a scale mixture,

$$\boldsymbol{\beta}_g \sim \mathcal{N}_{m_g}(\mathbf{0}, \tau \boldsymbol{I}_{m_g}), \quad \tau \sim \mathcal{G}\left(\frac{m_g+1}{2}, \frac{\lambda^2}{2}\right).$$

Therefore, using iterated expectations, we have

$$\begin{aligned}\mathbb{E}(\|\boldsymbol{\beta}_g\|_2^2) &= \mathbb{E}\left[\mathbb{E}(\|\boldsymbol{\beta}_g\|_2^2 \mid \tau)\right] \\ &= m_g \mathbb{E}(\tau) \\ &= \frac{m_g(m_g+1)}{\lambda^2}.\end{aligned}$$

□

Lemma D.2. *Suppose $\sigma^2 > 0, \sigma_0^2 > 0$. Then for any $\epsilon_n \in (0,1)$ such that $\epsilon_n \to 0$ as $n \to \infty$, we have for sufficiently large n,*

$$\left\{|\sigma^2 - \sigma_0^2| \geq 4\sigma_0^2\epsilon_n\right\} \subseteq \left\{\frac{\sigma^2}{\sigma_0^2} > \frac{1-\epsilon_n}{1-\epsilon_n} \text{ or } \frac{\sigma^2}{\sigma_0^2} < \frac{1-\epsilon_n}{1+\epsilon_n}\right\}.$$

Proof. For large n, $\epsilon_n < 1/2$, so $2\epsilon_n/(1-\epsilon_n) < 4\epsilon_n, -2\epsilon_n/(1+\epsilon_n) > -4\epsilon_n$, and thus,

$$\begin{aligned}
|\sigma^2 - \sigma_0^2| \geq 4\sigma_0^2\epsilon_n &\Rightarrow (\sigma^2 - \sigma_0^2)/\sigma_0^2 \geq 4\epsilon_n \text{ or } (\sigma^2 - \sigma_0^2)/\sigma_0^2 \leq -4\epsilon_n \\
&\Rightarrow \frac{\sigma^2}{\sigma_0^2} - 1 > \frac{2\epsilon_n}{1-\epsilon_n} \text{ or } \frac{\sigma^2}{\sigma_0^2} - 1 < -\frac{2\epsilon_n}{1+\epsilon_n} \\
&\Rightarrow \frac{\sigma^2}{\sigma_0^2} > \frac{1+\epsilon_n}{1-\epsilon_n} \text{ or } \frac{\sigma^2}{\sigma_0^2} < \frac{1-\epsilon_n}{1+\epsilon_n},
\end{aligned}$$

and hence,

$$|\sigma^2 - \sigma_0^2| \geq 4\sigma_0^2\epsilon_n \;\Rightarrow\; \frac{\sigma^2}{\sigma_0^2} > \frac{1+\epsilon_n}{1-\epsilon_n} \text{ or } \frac{\sigma^2}{\sigma_0^2} < \frac{1-\epsilon_n}{1+\epsilon_n}.$$

. □

Lemma D.3. *Suppose that a vector $\boldsymbol{z} \in \mathbb{R}^m$ can be decomposed into subvectors, $\boldsymbol{z} = [\boldsymbol{z}_1', \ldots, \boldsymbol{z}_d']$, where $\sum_{i=1}^d |\boldsymbol{z}_i| = m$ and $|\boldsymbol{z}_i|$ denotes the length of $\boldsymbol{z}_i$. Then $\|\boldsymbol{z}\|_2 \leq \sum_{i=1}^d \|\boldsymbol{z}_i\|_2$.*

Proof. We have

$$\begin{aligned}
\|\boldsymbol{z}\|_2 &= \sqrt{z_{11}^2 + \ldots + z_{1|z_1|}^2 + \ldots + z_{d1}^2 + \ldots + z_{d|z_d|}} \\
&\leq \sqrt{z_{11}^2 + \ldots + z_{1|z_1|}^2} + \ldots + \sqrt{z_{d1}^2 + \ldots + z_{d|z_d|}} \\
&= \|\boldsymbol{z}_1\|_2 + \ldots + \|\boldsymbol{z}_d\|_2.
\end{aligned}$$

□

D.2 Proofs for Section 3

Proof of Proposition 2. This result follows from an adaptation of the arguments of Zhang and Zhang (2012). The group-specific optimization problem is:

$$\widehat{\boldsymbol{\beta}}_g = \arg\max_{\boldsymbol{\beta}_g} \left\{ -\frac{1}{2}\|\boldsymbol{z}_g - \boldsymbol{\beta}_g\|_2^2 + \sigma^2 pen_S(\boldsymbol{\beta}|\theta) \right\}. \tag{D.1}$$

We first note that the optimization problem (D.1) is equivalent to maximizing the objective

$$L(\boldsymbol{\beta}_g) = -\frac{1}{2}\|\boldsymbol{z}_g - \boldsymbol{\beta}_g\|_2^2 + \sigma^2 pen_S(\boldsymbol{\beta}|\theta) + \frac{1}{2}\|\boldsymbol{z}_g\|_2^2 \tag{D.2}$$

$$= \|\boldsymbol{\beta}_g\|_2 \left[\frac{\boldsymbol{\beta}_g^T \boldsymbol{z}_g}{\|\boldsymbol{\beta}\|_2} - \left(\frac{\|\boldsymbol{\beta}_g\|_2}{2} - \frac{\sigma^2 pen_S(\boldsymbol{\beta}|\theta)}{\|\boldsymbol{\beta}_g\|_2} \right) \right] \tag{D.3}$$

$$= \|\boldsymbol{\beta}_g\|_2 \left[\|\boldsymbol{z}_g\|_2 \cos\varphi - \left(\frac{\|\boldsymbol{\beta}_g\|_2}{2} - \frac{\sigma^2 pen_S(\boldsymbol{\beta}|\theta)}{\|\boldsymbol{\beta}_g\|_2} \right) \right] \tag{D.4}$$

where φ is the angle between $\boldsymbol{z}_g$ and $\boldsymbol{\beta}_g$. Then, when $\|\boldsymbol{z}_g\|_2 < \Delta$, the second factorized term of (D.4) is always less than zero, and so $\widehat{\boldsymbol{\beta}}_g = \mathbf{0}_{m_g}$ must be the global maximizer of L. On the other hand, when the global maximizer $\widehat{\boldsymbol{\beta}}_g = \mathbf{0}_{m_g}$, then the second factorized term must always be less than zero, otherwise $\widehat{\boldsymbol{\beta}}_g = \mathbf{0}_{m_g}$ would no longer be the global maximizer and so $\|\boldsymbol{z}_g\|_2 < \Delta$. □

Proof of Lemma 3. We have

$$\mathbb{E}[\theta|\widehat{\boldsymbol{\beta}}] = \frac{\int_0^1 \theta^a (1-\theta)^{b-1} (1-\theta z)^{G-\widehat{q}} \prod_{g=1}^{\widehat{q}} (1-\theta x_g) d\theta}{\int_0^1 \theta^{a-1} (1-\theta)^{b-1} (1-\theta z)^{G-\widehat{q}} \prod_{g=1}^{\widehat{q}} (1-\theta x_g) d\theta}. \tag{D.5}$$

When $\lambda_0 \to \infty$, we have $z \to 1$ and $x_g \to -\infty$ for all $g = 1, \ldots, \widehat{q}$. Hence,

$$\lim_{\lambda_0 \to \infty} \mathbb{E}[\theta|\widehat{\boldsymbol{\beta}}] = \lim_{z \to 1} \lim_{x_g \to -\infty} \frac{\int_0^1 \theta^a (1-\theta)^{b+G-\widehat{q}-1} \prod_{g=1}^{\widehat{q}} (1-\theta x_g)}{\int_0^1 \theta^{a-1} (1-\theta)^{b-1} (1-\theta z)^{G-\widehat{q}} \prod_{g=1}^{\widehat{q}} (1-\theta x_g)} \tag{D.6}$$

$$= \frac{\int_0^1 \theta^{a+\widehat{q}} (1-\theta)^{b+G-\widehat{q}-1} d\theta}{\int_0^1 \theta^{a+\widehat{q}-1} (1-\theta)^{b+G-\widehat{q}-1} d\theta} \tag{D.7}$$

$$= \frac{a + \widehat{q}}{a + b + G}. \tag{D.8}$$

□

D.3 Proofs for Section 6

In this section, we use proof techniques from Ning et al. (2019), Song and Liang (2017), and Wei et al. (2020) rather than the ones in Ročková and George (2018). However, none of these other papers considers *both* continuous spike-and-slab priors for groups of regression coefficients *and* an independent prior on the unknown variance.

Proof of Theorem 2. Our proof is based on first principles of verifying Kullback-Leibler (KL) and testing conditions (see e.g., Ghosal et al. (2000)). We first prove (6.3) and (6.5).

Part I: Kullback-Leibler conditions. Let $f \sim \mathcal{N}_n(\boldsymbol{X}\boldsymbol{\beta}, \sigma^2 \boldsymbol{I}_n)$ and $f_0 \sim \mathcal{N}_n(\boldsymbol{X}\boldsymbol{\beta}_0, \sigma_0^2 \boldsymbol{I}_n)$, and let $\Pi(\cdot)$ denote the prior (6.2). We first show that for our choice of $\epsilon_n = \sqrt{s_0 \log G/n}$,

$$\Pi\left(K(f_0, f) \leq n\epsilon_n^2, V(f_0, f) \leq n\epsilon_n^2\right) \geq \exp(-C_1 n\epsilon_n^2), \tag{D.9}$$

for some constant $C_1 > 0$, where $K(\cdot, \cdot)$ denotes the KL divergence and $V(\cdot, \cdot)$ denotes the KL variation. The KL divergence between f_0 and f is

$$K(f_0, f) = \frac{1}{2}\left[n\left(\frac{\sigma_0^2}{\sigma^2}\right) - n - n\log\left(\frac{\sigma_0^2}{\sigma^2}\right) + \frac{\|\boldsymbol{X}(\boldsymbol{\beta} - \boldsymbol{\beta}_0)\|_2^2}{\sigma^2}\right], \tag{D.10}$$

and the KL variation between f_0 and f is

$$V(f_0, f) = \frac{1}{2}\left[n\left(\frac{\sigma_0^2}{\sigma^2}\right)^2 - 2n\left(\frac{\sigma_0^2}{\sigma^2}\right) + n\right] + \frac{\sigma_0^2}{(\sigma^2)^2}\|\boldsymbol{X}(\boldsymbol{\beta} - \boldsymbol{\beta}_0)\|_2^2. \tag{D.11}$$

Define the two events $\mathcal{A}_1$ and $\mathcal{A}_2$ as follows:

$$\mathcal{A}_1 = \left\{ \sigma^2 : n\left(\frac{\sigma_0^2}{\sigma^2}\right) - n - n\log\left(\frac{\sigma_0^2}{\sigma^2}\right) \leq n\epsilon_n^2, \;\; n\left(\frac{\sigma_0^2}{\sigma^2}\right)^2 - 2n\left(\frac{\sigma_0^2}{\sigma^2}\right) + n \leq n\epsilon_n^2 \right\} \quad \text{(D.12)}$$

and

$$\mathcal{A}_2 = \left\{ (\boldsymbol{\beta}, \sigma^2) : \frac{\|\boldsymbol{X}(\boldsymbol{\beta} - \boldsymbol{\beta}_0)\|_2^2}{\sigma^2} \leq n\epsilon_n^2, \;\; \frac{\sigma_0^2}{(\sigma^2)^2}\|\boldsymbol{X}(\boldsymbol{\beta} - \boldsymbol{\beta}_0)\|_2^2 \leq n\epsilon_n^2/2 \right\}. \quad \text{(D.13)}$$

Following from (D.9)-(D.13), we may write $\Pi(K(f_0, f) \leq \epsilon_n^2, V(f_0, f) \leq \epsilon_n^2) = \Pi(\mathcal{A}_2|\mathcal{A}_1)\Pi(\mathcal{A}_1)$. We derive lower bounds for $\Pi(\mathcal{A}_1)$ and $\Pi(\mathcal{A}_2|\mathcal{A}_1)$ separately. Noting that we may rewrite $\mathcal{A}_1$ as

$$\mathcal{A}_1 = \left\{ \sigma^2 : \frac{\sigma_0^2}{\sigma^2} - 1 - \log\left(\frac{\sigma_0^2}{\sigma^2}\right) \leq \epsilon_n^2, \;\; \left(\frac{\sigma_0^2}{\sigma^2} - 1\right)^2 \leq \epsilon_n^2 \right\},$$

and expanding $\log(\sigma_0^2/\sigma^2)$ in the powers of $1 - \sigma_0^2/\sigma^2$ to get $\sigma_0^2/\sigma^2 - 1 - \log(\sigma_0^2/\sigma^2) \sim (1 - \sigma_0^2/\sigma^2)^2/2$, it is clear that $\mathcal{A}_1 \supset \mathcal{A}_1^\star$, where $\mathcal{A}_1^\star = \{\sigma^2 : \sigma_0^2/(\epsilon_n + 1) \leq \sigma^2 \leq \sigma_0^2\}$. Thus, since $\sigma^2 \sim \mathcal{IG}(c_0, d_0)$, we have for sufficiently large n,

$$\begin{aligned} \Pi(\mathcal{A}_1) \geq \Pi(\mathcal{A}_1^\star) &\asymp \int_{\sigma_0^2/(\epsilon_n+1)}^{\sigma_0^2} (\sigma^2)^{-c_0-1} e^{-d_0/\sigma^2} d\sigma^2 \\ &\geq (\sigma_0^2)^{-c_0-1} e^{-d_0(\epsilon_n+1)/\sigma_0^2}. \end{aligned} \quad \text{(D.14)}$$

Thus, from (D.14), we have

$$-\log\Pi(\mathcal{A}_1) \lesssim \epsilon_n + 1 \lesssim n\epsilon_n^2, \quad \text{(D.15)}$$

since $n\epsilon_n^2 \to \infty$. Next, we consider $\Pi(\mathcal{A}_2|\mathcal{A}_1)$. We have

$$\frac{\sigma_0^2}{(\sigma^2)^2}\|\boldsymbol{X}(\boldsymbol{\beta} - \boldsymbol{\beta}_0)\|_2^2 = \left\|\frac{\boldsymbol{X}(\boldsymbol{\beta} - \boldsymbol{\beta}_0)}{\sigma}\right\|_2^2 \left(\frac{\sigma_0^2}{\sigma^2} - 1\right) + \left\|\frac{\boldsymbol{X}(\boldsymbol{\beta} - \boldsymbol{\beta}_0)}{\sigma}\right\|_2^2,$$

and conditional on $\mathcal{A}_1$, we have that the previous display is bounded above by

$$\left|\left|\frac{\boldsymbol{X}(\boldsymbol{\beta}-\boldsymbol{\beta}_0)}{\sigma}\right|\right|_2^2(\epsilon_n+1) < \frac{2}{\sigma^2}\|\boldsymbol{X}(\boldsymbol{\beta}-\boldsymbol{\beta}_0)\|_2^2,$$

for large n (since $\epsilon_n < 1$ when n is large). Since $\mathcal{A}_1 \supset \mathcal{A}_1^\star$, where $\mathcal{A}_1^\star$ was defined earlier, the left-hand side of both expressions in (D.13) can be bounded above by a constant multiple of $\|\boldsymbol{X}(\boldsymbol{\beta}-\boldsymbol{\beta}_0)\|_2^2$, conditional on $\mathcal{A}_1$. Therefore, for some constant $b_1 > 0$, $\Pi(\mathcal{A}_2|\mathcal{A}_1)$ is bounded below by

$$\begin{aligned}
\Pi(\mathcal{A}_2|\mathcal{A}_1) &\geq \Pi\left(\|\boldsymbol{X}(\boldsymbol{\beta}-\boldsymbol{\beta}_0)\|_2^2 \leq \frac{b_1^2 n\epsilon_n^2}{2}\right) \\
&\geq \Pi\left(\|\boldsymbol{\beta}-\boldsymbol{\beta}_0\|_2^2 \leq \frac{b_1^2\epsilon_n^2}{2kn^{\alpha-1}}\right) \\
&\geq \int_0^1 \left\{\Pi_{S_0}\left(\|\boldsymbol{\beta}_{S_0}-\boldsymbol{\beta}_{0S_0}\|_2^2 \leq \frac{b_1^2\epsilon_n^2}{4kn^\alpha}\middle|\theta\right)\right\}\left\{\Pi_{S_0^c}\left(\|\boldsymbol{\beta}_{S_0^c}\|_2^2 \leq \frac{b_1^2\epsilon_n^2}{4kn^\alpha}\middle|\theta\right)\right\} d\pi(\theta), \qquad \text{(D.16)}
\end{aligned}$$

where we used Assumption (A3) in the second inequality, and in the third inequality, we used the fact that conditional on θ, the SSGL prior is separable, so $\pi(\boldsymbol{\beta}|\theta) = \pi_{S_0}(\boldsymbol{\beta}|\theta)\pi_{S_0^c}(\boldsymbol{\beta}|\theta)$. We proceed to lower-bound each bracketed integrand term in (D.16) separately. Changing the variable $\boldsymbol{\beta}-\boldsymbol{\beta}_0 \to \boldsymbol{b}$ and using the fact that $\pi_{S_0}(\boldsymbol{\beta}|\theta) > \theta^{s_0}\prod_{g\in S_0}\left[C_g\lambda_1^{m_g}e^{-\lambda_1\|\boldsymbol{\beta}_g\|_2}\right]$ and $\|\boldsymbol{z}\|_2 \leq \|\boldsymbol{z}\|_1$ for any vector $\boldsymbol{z}$, we have as a lower bound for the first term in (D.16),

$$\theta^{s_0}e^{-\lambda_1\|\boldsymbol{\beta}_{S_0}\|_2}\prod_{g\in S_0}C_g\left\{\int_{\|\boldsymbol{b}_g\|_1\leq\frac{b_1\epsilon_n}{2s_0\sqrt{kn^\alpha}}}\lambda_1^{m_g}e^{-\lambda_1\|\boldsymbol{b}_g\|_1}d\boldsymbol{b}_g\right\}. \qquad \text{(D.17)}$$

Each of the integral terms in (D.17) is the probability of the first m_g events of a Poisson process happening before time $b_1\epsilon_n/2s_0\sqrt{kn^\alpha}$. Using similar arguments as those in the proof of Lemma 5.1 of Ning et al. (2019), we obtain as a lower bound for the product of integrals in (D.17),

$$\prod_{g\in S_0}C_g\left\{\int_{\|\boldsymbol{b}_g\|_1\leq\frac{b_1\epsilon_n}{2s_0\sqrt{kn^\alpha}}}\lambda_1^{m_g}e^{-\lambda_1\|\boldsymbol{b}_g\|_1}d\boldsymbol{b}_g\right\} \geq \prod_{g\in S_0}C_g e^{-\lambda_1 b_1\epsilon_n/2s_0\sqrt{kn^\alpha}}\frac{1}{m_g!}\left(\frac{\lambda_1 b_1\epsilon_n}{s_0\sqrt{kn^\alpha}}\right)^{m_g}$$

$$= e^{-\lambda_1 b_1 \epsilon_n / 2\sqrt{kn^\alpha}} \prod_{g \in S_0} \frac{C_g}{m_g!} \left(\frac{\lambda_1 b_1 \epsilon_n}{s_0 \sqrt{kn^\alpha}} \right)^{m_g}. \quad \text{(D.18)}$$

Combining (D.17)-(D.18), we have the following lower bound for the first bracketed term in (D.16):

$$\theta^{s_0} e^{-\lambda_1 \|\boldsymbol{\beta}_{S_0}\|_2} e^{-\lambda_1 b_1 \epsilon_n / 2\sqrt{kn^\alpha}} \prod_{g \in S_0} \frac{C_g}{m_g!} \left(\frac{\lambda_1 b_1 \epsilon_n}{s_0 \sqrt{kn^\alpha}} \right)^{m_g}. \quad \text{(D.19)}$$

Now, noting that $\pi_{S_0^c}(\boldsymbol{\beta}|\theta) > (1-\theta)^{G-s_0} \prod_{g \in S_0^c} \left[C_g \lambda_0^{m_g} e^{-\lambda_0 \|\boldsymbol{\beta}_g\|_2} \right]$, we further bound the second bracketed term in (D.16) from below. Let $\check{\pi}(\cdot)$ denote the density, $\check{\pi}(\boldsymbol{\beta}_g) = C_g \lambda_0^{m_g} e^{-\lambda_0 \|\boldsymbol{\beta}_g\|_2}$. We have

$$\begin{aligned}
\Pi_{S_0^c} \left(\|\boldsymbol{\beta}_{S_0^c}\|_2^2 \leq \frac{b_1^2 \epsilon_n^2}{4kn^\alpha} \middle| \theta \right) &> (1-\theta)^{G-s_0} \prod_{g \in S_0^c} \check{\Pi} \left(\|\boldsymbol{\beta}_g\|_2^2 \leq \frac{b_1^2 \epsilon_n^2}{4kn^\alpha (G-s_0)} \right) \\
&\geq (1-\theta)^{G-s_0} \prod_{g \in S_0^c} \left[1 - \frac{4kn^\alpha (G-s_0) \mathbb{E}_{\check{\pi}} \left(\|\boldsymbol{\beta}_g\|_2^2 \right)}{b_1^2 \epsilon_n^2} \right] \\
&= (1-\theta)^{G-s_0} \prod_{g \in S_0^c} \left[1 - \frac{4kn^\alpha (G-s_0) m_g (m_g+1)}{\lambda_0^2 b_1^2 \epsilon_n^2} \right] \\
&\geq (1-\theta)^{G-s_0} \left[1 - \frac{4kn^\alpha G m_{\max} (m_{\max}+1)}{\lambda_0^2 b_1 \epsilon_n^2} \right]^{G-s_0}, \quad \text{(D.20)}
\end{aligned}$$

where we used an application of the Markov inequality and Lemma D.1 in the second line. Combining (D.19)-(D.20) gives as a lower-bound for (D.16),

$$\begin{aligned}
\Pi(\mathcal{A}_2|\mathcal{A}_1) \geq &\left\{ e^{-\lambda_1 \|\boldsymbol{\beta}_{S_0}\|_2} e^{-\lambda_1 b_1 \epsilon_n / 2\sqrt{kn^\alpha}} \prod_{g \in S_0} \frac{C_g}{m_g!} \left(\frac{\lambda_1 b_1 \epsilon_n}{s_0 \sqrt{kn^\alpha}} \right)^{m_g} \right\} \\
&\times \left\{ \int_0^1 \theta^{s_0} (1-\theta)^{G-s_0} \left[1 - \frac{4kn^\alpha G m_{\max} (m_{\max}+1)}{\lambda_0^2 b_1 \epsilon_n^2} \right]^{G-s_0} d\pi(\theta) \right\} \quad \text{(D.21)}
\end{aligned}$$

Let us consider the second bracketed term in (D.21) first. By assumption, $\lambda_0 = (1-\theta)/\theta$. Further, $\lambda_0^2 = (1-\theta)^2/\theta^2$ is monotonically decreasing in θ for $\theta \in (0,1)$. Hence, for constant

$c > 2$ in the $\mathcal{B}(1, G^c)$ prior on θ, a lower bound for the second bracketed term in (D.21) is

$$\begin{aligned}
&\int_{1/(2G^c+1)}^{1/(G^c+1)} \theta^{s_0}(1-\theta)^{G-s_0}\left[1 - \frac{4kn^\alpha G m_{\max}(m_{\max}+1)}{\lambda_0^2 b_1 \epsilon_n^2}\right]^{G-s_0} d\pi(\theta) \\
&\geq (2G^c+1)^{-s_0}\left[1 - \frac{4kn^\alpha G m_{\max}(m_{\max}+1)}{G^{2c} b_1 \epsilon_n^2}\right]^{G-s_0} \int_{1/(2G^c+1)}^{1/(G^c+1)} (1-\theta)^{G-s_0} d\pi(\theta) \\
&\gtrsim (2G^c+1)^{-s_0}\left[1 - \frac{1}{G-s_0}\right]^{G-s_0} \int_{1/(2G^c+1)}^{1/(G^c+1)} (1-\theta)^{G-s_0} d\pi(\theta) \\
&\asymp (2G^c+1)^{-s_0} G^{-c} \int_{1/(2G^c+1)}^{1/(G^c+1)} (1-\theta)^{G^c+G-s_0-1} d\theta \\
&= (2G^c+1)^{-s_0} G^{-c}(G^c+G-s_0)^{-1}\left[\left(1 - \frac{1}{2G^c+1}\right)^{G^c+G-s_0} - \left(1 - \frac{1}{G^c+1}\right)^{G^c+G-s_0}\right] \\
&\gtrsim (2G^c+1)^{-s_0} G^{-c}(G^c+G-s_0)^{-1}, \qquad \text{(D.22)}
\end{aligned}$$

where in the third line, we used our assumptions about the growth rates for $m_{\max}$, G, and s_0 in Assumptions (A1)-(A2) and the fact that $c > 2$. In the fourth line, we used the fact that $(1 - 1/x)^x \to e^{-1}$ as $x \to \infty$ and $\theta \sim \mathcal{B}(1, G^c)$. In the sixth line, we used the fact that the bracketed term in the fifth line can be bounded below by $e^{-1} - e^{-2}$ for sufficiently large n.

Combining (D.21)-(D.22), we obtain for sufficiently large n,

$$\begin{aligned}
-\log \Pi(\mathcal{A}_2|\mathcal{A}_1) \lesssim\ & \lambda_1 \|\boldsymbol{\beta}_{0S_0}\|_2 + \frac{\lambda_1 b_1 \epsilon_n}{2\sqrt{kn^\alpha}} + \sum_{g \in S_0} \log(m_g!) - \sum_{g \in S_0} \log C_g \\
&+ \sum_{g \in S_0} m_g \log\left(\frac{s_0\sqrt{kn^\alpha}}{\lambda_1 b_1 \epsilon_n}\right) + s_0 \log(2G^c+1) + c \log G \\
&+ \log(G^c + G - s_0) \qquad \text{(D.23)}
\end{aligned}$$

We examine each of the terms in (D.23) separately. By Assumptions (A1) and (A5) and

the fact that $\lambda_1 \asymp 1/n$, we have

$$\lambda_1 \|\boldsymbol{\beta}_{0S_0}\|_2 \leq \lambda_1 \sqrt{s_0 m_{\max}} \|\boldsymbol{\beta}_{0S_0}\|_\infty \lesssim s_0 \log G \lesssim n\epsilon_n^2,$$

and

$$\frac{\lambda_1 b_1 \epsilon_n}{2\sqrt{kn^\alpha}} \lesssim \epsilon_n \lesssim n\epsilon_n^2.$$

Next, using the facts that $x! \leq x^x$ for $x \in \mathbb{N}$ and Assumption (A1), we have

$$\sum_{g \in S_0} \log(m_g!) \leq s_0 \log(m_{\max}!) \leq s_0 m_{\max} \log(m_{\max}) \leq s_0 m_{\max} \log n \lesssim n\epsilon_n^2.$$

Using the fact that the normalizing constant, $C_g = 2^{-m_g} \pi^{-(m_g-1)/2} [\Gamma((m_g+1)/2)]^{-1}$, we also have

$$\begin{aligned}
\sum_{g \in S_0} -\log C_g &= \sum_{g \in S_0} \left\{ m_g \log 2 + \left(\frac{m_g - 1}{2} \right) \log \pi + \log \left[\Gamma \left(\frac{m_g + 1}{2} \right) \right] \right\} \\
&\leq s_0 m_{\max} (\log 2 + \log \pi) + \sum_{g \in S_0} \log(m_g!) \\
&\lesssim s_0 m_{\max} (\log 2 + \log \pi) + s_0 m_{\max} \log n \\
&\lesssim s_0 \log G \\
&\lesssim n\epsilon_n^2,
\end{aligned}$$

where we used the fact that $\Gamma((m_g+1)/2) \leq \Gamma(m_g + 1) = m_g!$. Finally, since $\lambda_1 \asymp 1/n$ and using Assumption (A1) that $m_{\max} = O(\log G / \log n)$, we have

$$\begin{aligned}
\sum_{g \in S_0} m_g \log \left(\frac{s_0 \sqrt{kn^\alpha}}{\lambda_1 b_1 \epsilon_n} \right) &\lesssim s_0 m_{\max} \log \left(\frac{s_0 n^{\alpha/2+1} \sqrt{k}}{b_1 \epsilon_n^2} \right) \\
&= s_0 m_{\max} \log \left(\frac{n^{\alpha/2+2} \sqrt{k}}{b_1 \log G} \right)
\end{aligned}$$

$$\lesssim s_0 m_{\max} \log n$$
$$\lesssim s_0 \log G$$
$$\lesssim n\epsilon_n^2.$$

Finally, it is clear by the definition of $n\epsilon_n^2$ and the fact that $c > 2$ is a constant that

$$s_0 \log(2G^c + 1) + c \log G + \log(G^c + G - s_0) \asymp s_0 \log G = n\epsilon_n^2.$$

Combining all of the above, together with (D.23), we have

$$-\log \Pi(\mathcal{A}_2|\mathcal{A}_1) \lesssim n\epsilon_n^2. \tag{D.24}$$

By (D.15) and (D.24), we may choose a large constant $C_1 > 0$, so that

$$\Pi(\mathcal{A}_2|\mathcal{A}_1)\Pi(\mathcal{A}_1) \gtrsim \exp(-C_1 n\epsilon_n^2/2)\exp(-C_1 n\epsilon_n^2/2) = \exp(-C_1 n\epsilon_n^2),$$

so the Kullback-Leibler condition (D.9) holds.

Part II: Testing conditions. To complete the proof, we show the existence of a sieve $\mathcal{F}_n$ such that

$$\Pi(\mathcal{F}_n^c) \leq \exp(-C_2 n\epsilon_n^2), \tag{D.25}$$

for positive constant $C_2 > C_1 + 2$, where C_1 is the constant from (D.9), and a sequence of test functions $\phi_n \in [0, 1]$ such that

$$\mathbb{E}_{f_0}\phi_n \leq e^{-C_4 n\epsilon_n^2}, \tag{D.26}$$

and

$$\sup_{f \in \mathcal{F}_n :\ \|\boldsymbol{\beta} - \boldsymbol{\beta}_0\|_2 \geq (3+\sqrt{\nu_1})\sigma_0\epsilon_n,\ \text{or } |\sigma^2 - \sigma_0^2| \geq 4\sigma_0^2\epsilon_n} \mathbb{E}_f(1 - \phi_n) \leq e^{-C_4 n\epsilon_n^2}, \tag{D.27}$$

for some $C_4 > 0$, where ν is from Assumption (A4). Recall that $\omega_g \equiv \omega_g(\lambda_0, \lambda_1, \theta) = \frac{1}{\lambda_0 - \lambda_1} \log\left[\frac{1-\theta}{\theta} \frac{\lambda_0^{m_g}}{\lambda_1^{m_g}}\right]$. Choose $C_3 \geq C_1 + 2 + \log 3$, and consider the sieve,

$$\mathcal{F}_n = \left\{ f : |\boldsymbol{\gamma}(\boldsymbol{\beta})| \leq C_3 s_0, 0 < \sigma^2 \leq G^{C_3 s_0 / c_0} \right\}, \tag{D.28}$$

where c_0 is from $\mathcal{IG}(c_0, d_0)$ prior on σ^2 and $|\boldsymbol{\gamma}(\boldsymbol{\beta})|$ denotes the generalized dimensionality (6.6).

We first verify (D.25). We have

$$\Pi(\mathcal{F}_n^c) \leq \Pi\left(|\boldsymbol{\gamma}(\boldsymbol{\beta})| > C_3 s_0\right) + \Pi\left(\sigma^2 > G^{C_3 s_0 / c_0}\right). \tag{D.29}$$

We focus on bounding each of the terms in (D.29) separately. First, let $\theta_0 = C_3 s_0 \log G / G^c$, where $c > 2$ is the constant in the $\mathcal{B}(1, G^c)$ prior on θ. Similarly as in the proof of Theorem 6.3 in Ročková and George (2018), we have $\pi(\boldsymbol{\beta}_g | \theta) < 2\theta C_g \lambda_1^{m_g} e^{-\lambda_1 \|\boldsymbol{\beta}_g\|_2}$ for all $\|\boldsymbol{\beta}_g\|_2 > \omega_g$. We have for any $\theta \leq \theta_0$ that

$$\begin{aligned}
\Pi(|\boldsymbol{\gamma}(\boldsymbol{\beta})| > C_3 s_0 | \theta) &\leq \sum_{S : |S| > C_3 s_0} 2^{|S|} \theta_0^{|S|} \int_{\|\boldsymbol{\beta}_g\|_2 > \omega_g ; g \in S} C_g \lambda_1^{m_g} e^{-\lambda_1 \|\boldsymbol{\beta}_g\|_2} d\boldsymbol{\beta}_S \\
&\quad \times \int_{\|\boldsymbol{\beta}_g\|_2 \leq \omega_g ; g \in S^c} \Pi_{S^c}(\boldsymbol{\beta}) d\boldsymbol{\beta}_{S^c} \\
&\lesssim \sum_{S : |S| > C_3 s_0} \theta_0^{|S|},
\end{aligned} \tag{D.30}$$

where we used the assumption that $\lambda_1 \asymp 1/n$, the definition of ω_g, and the fact that $\theta \leq \theta_0$ to bound the first integral term from above by $\prod_{g \in S} (1/n)^{m_g} \leq n^{-|S|}$, and we bounded the second integral term above by one. We then have

$$\begin{aligned}
\Pi(|\boldsymbol{\gamma}(\boldsymbol{\beta})| > C_3 s_0) &= \int_0^1 \Pi(|\boldsymbol{\gamma}(\boldsymbol{\beta})| > C_3 s_0 | \theta) d\pi(\theta) \\
&\leq \int_0^{\theta_0} \Pi(|\boldsymbol{\gamma}(\boldsymbol{\beta})| > C_3 s_0 | \theta) d\pi(\theta) + \Pi(\theta > \theta_0).
\end{aligned} \tag{D.31}$$

Note that since $s_0 = o(n/\log G)$ by Assumption (A1), $G \gg n$, and $c > 2$, we have $\theta_0 \leq C_3 n/G^c < 1/G^2$ for sufficiently large n. Following from (D.30), we thus have that for sufficiently large n,

$$\begin{aligned}
\int_0^{\theta_0} \Pi(|\boldsymbol{\gamma}(\boldsymbol{\beta})| > C_3 s_0 | \theta) d\pi(\theta) &\leq \sum_{S:|S|>C_3 s_0} \theta_0^{|S|} \\
&\leq \sum_{k=\lfloor C_3 s_0 \rfloor + 1}^{G} \binom{G}{k} \left(\frac{1}{G^2}\right)^k \\
&\leq \sum_{k=\lfloor C_3 s_0 \rfloor + 1}^{G} \left(\frac{e}{kG}\right)^k \\
&< \sum_{k=\lfloor C_3 s_0 \rfloor + 1}^{G} \left(\frac{e}{G(\lfloor C_3 s_0 \rfloor + 1)}\right)^k \\
&= \frac{\left(\frac{e}{G(\lfloor C_3 s_0 \rfloor + 1)}\right)^{\lfloor C_3 s_0 \rfloor + 1} - \left(\frac{e}{G(\lfloor C_3 s_0 \rfloor + 1)}\right)^{G+1}}{1 - \frac{e}{G(\lfloor C_3 s_0 \rfloor + 1)}} \\
&\lesssim G^{-(\lfloor C_3 s_0 \rfloor + 1)} \\
&\lesssim \exp\left(-C_3 n \epsilon_n^2\right). \qquad \text{(D.32)}
\end{aligned}$$

where we used the inequality $\binom{G}{k} \leq (eG/k)^k$ in the third line of the display.

Next, since $\theta \sim \mathcal{B}(1, G^c)$, we have

$$\begin{aligned}
\Pi(\theta > \theta_0) &= (1 - \theta_0)^{G^c} \\
&= \left(1 - \frac{C_3 s_0 \log G}{G^c}\right)^{G^c} \\
&\leq e^{-C_3 s_0 \log G} \\
&= e^{-C_3 n \epsilon_n^2}. \qquad \text{(D.33)}
\end{aligned}$$

Combining (D.31)-(D.33), we have that

$$\Pi(|\boldsymbol{\gamma}(\boldsymbol{\beta})| > C_3 s_0) \leq 2e^{-C_3 n\epsilon_n^2}. \tag{D.34}$$

Finally, we have as a bound for the second term in (D.29),

$$\begin{aligned}
\Pi\left(\sigma^2 > G^{C_3 s_0/c_0}\right) &= \int_{G^{C_3 s_0/c_0}}^{\infty} \frac{d_0^{c_0}}{\Gamma(c_0)} (\sigma^2)^{-c_0-1} e^{-d_0/\sigma^2} d\sigma^2 \\
&\lesssim \int_{G^{C_3 s_0/c_0}}^{\infty} (\sigma^2)^{-c_0-1} \\
&\asymp G^{-C_3 s_0} \\
&\lesssim \exp(-C_3 n\epsilon_n^2). \qquad \text{(D.35)}
\end{aligned}$$

Combining (D.29)-(D.35), we have

$$\Pi(\mathcal{F}_n^c) \leq 3\exp\left(-C_3 n\epsilon_n^2\right) = \exp\left(-C_3 n\epsilon_n^2 + \log 3\right),$$

and so given our choice of C_3, (D.29) is asymptotically bounded from above by $\exp(-C_2 n\epsilon_n^2)$ for some $C_2 \geq C_1 + 2$. This proves (D.25).

We now proceed to prove (D.26). Our proof is based on the technique used in Song and Liang (2017) with suitable modifications. For $\xi \subset \{1, \ldots, G\}$, let $\boldsymbol{X}_\xi$ denote the submatrix of $\boldsymbol{X}$ with submatrices indexed by ξ, where $|\xi| \leq \bar{p}$ and $\bar{p}$ is from Assumption (A4). Let $\widehat{\boldsymbol{\beta}}_\xi = (\boldsymbol{X}_\xi^T \boldsymbol{X}_\xi)^{-1} \boldsymbol{X}_\xi^T \boldsymbol{Y}$ and $\boldsymbol{\beta}_{0\xi}$ denote the subvector of $\boldsymbol{\beta}_0$ with groups indexed by ξ. Let $m_\xi = \sum_{g\in\xi} m_g$, and let $\widehat{\sigma}_\xi^2 = \|\boldsymbol{Y} - \boldsymbol{X}_\xi \widehat{\boldsymbol{\beta}}_\xi\|_2^2/(n - m_\xi)$. Note that $\widehat{\boldsymbol{\beta}}_\xi$ and $\widehat{\sigma}_\xi^2$ both exist and are unique because of Assumptions (A1), (A2), and (A4) (which combined, gives us that $m_\xi = o(n)$).

Let $\widetilde{p}$ be an integer satisfying $\widetilde{p} \asymp s_0$ and $\widetilde{p} \leq \bar{p} - s_0$, where $\bar{p}$ is from Assumption (A4), and the specific choice of $\widetilde{p}$ will be given below. Recall that S_0 is the set of true nonzero

groups with cardinality $s_0 = |S_0|$. Similar to Song and Liang (2017), we consider the test function $\phi_n = \max\{\phi'_n, \tilde{\phi}_n\}$, where

$$\begin{aligned} \phi'_n &= \max_{\xi \supset S_0, |\xi| \le \widetilde{p}+s_0} 1\left\{|\widehat{\sigma}^2_\xi - \sigma^2_0| \ge \sigma^2_0 \epsilon_n\right\}, \qquad \text{and} \\ \tilde{\phi}_n &= \max_{\xi \supset S_0, |\xi| \le \widetilde{p}+s_0} 1\left\{\|\widehat{\boldsymbol{\beta}}_\xi - \boldsymbol{\beta}_{0\xi}\|_2 \ge \sigma_0 \epsilon_n\right\}. \end{aligned} \tag{D.36}$$

Because of Assumption (A4), we have $\widetilde{p} \prec n$ and $\widetilde{p} \prec n\epsilon_n^2$. Additionally, since $\epsilon_n = o(1)$, we can use almost identical arguments as those used to establish (A.5)-(A.6) in the proof of Theorem A.1 of Song and Liang (2017) to show that for any ξ satisfying $\xi \supset S_0, |\xi| \le \widetilde{p}$,

$$\mathbb{E}_{(\boldsymbol{\beta}_0, \sigma_0^2)} 1\left\{|\widehat{\sigma}^2_\xi - \sigma^2_0| \ge \sigma^2_0 \epsilon_n\right\} \le \exp(-c'_4 n \epsilon_n^2),$$

for some constant $\hat{c}_4 > 0$, and for any ξ satisfying $\xi \supset S_0, |\xi| \le \widetilde{p}$,

$$\mathbb{E}_{(\boldsymbol{\beta}_0, \sigma_0^2)} 1\left\{\|\widehat{\boldsymbol{\beta}}_\xi - \boldsymbol{\beta}_{0\xi}\|_2 \ge \sigma_0 \epsilon_n\right\} \le \exp(-\tilde{c}_4 n \epsilon_n^2),$$

for some $\tilde{c}_4 > 0$. Using the proof of Theorem A.1 in Song and Liang (2017), we may then choose $\widetilde{p} = \lfloor \min\{c'_4, \tilde{c}_4\} n\epsilon_n^2 / (2 \log G) \rfloor$, and then

$$\mathbb{E}_{f_0} \phi_n \le \exp(-\check{c}_4 n \epsilon_n^2), \tag{D.37}$$

for some $\check{c}_4 > 0$. Next, define the set,

$$\begin{aligned} \mathcal{C} \;= \big\{ &\|\boldsymbol{\beta} - \boldsymbol{\beta}_0\|_2 \ge (3 + \sqrt{\nu_1}) \sigma_0 \epsilon_n \text{ or } \sigma^2/\sigma_0^2 > (1+\epsilon_n)/(1-\epsilon_n) \\ &\text{or } \sigma^2/\sigma_0^2 < (1-\epsilon_n)/(1+\epsilon_n) \big\}. \end{aligned}$$

By Lemma D.2, we have

$$\sup_{\substack{f \in \mathcal{F}_n: \; \|\boldsymbol{\beta} - \boldsymbol{\beta}_0\|_2 \ge (3+\sqrt{\nu_1})\sigma_0\epsilon_n, \\ \text{or } |\sigma^2 - \sigma_0^2| \ge 4\sigma_0^2 \epsilon_n}} \mathbb{E}_f(1 - \phi_n) \le \sup_{f \in \mathcal{F}_n : (\boldsymbol{\beta}, \sigma^2) \in \mathcal{C}} \mathbb{E}_f(1 - \phi_n). \tag{D.38}$$

Similar to Song and Liang (2017), we consider $\mathcal{C} \subset \widehat{\mathcal{C}} \cup \widetilde{\mathcal{C}}$, where

$$\widehat{\mathcal{C}} = \{\sigma^2/\sigma_0^2 > (1+\epsilon_n)/(1-\epsilon_n) \text{ or } \sigma^2/\sigma_0^2 < (1-\epsilon_n)/(1+\epsilon_n)\},$$
$$\tilde{\mathcal{C}} = \{\|\boldsymbol{\beta} - \boldsymbol{\beta}_0\| \geq (3+\sqrt{\nu_1})\sigma_0\epsilon_n \text{ and } \sigma^2 = \sigma_0^2\},$$

and so an upper bound for (D.38) is

$$\begin{aligned}\sup_{f \in \mathcal{F}_n : (\boldsymbol{\beta},\sigma^2) \in \mathcal{C}} \mathbb{E}_f(1-\phi_n) &= \sup_{f \in \mathcal{F}_n : (\boldsymbol{\beta},\sigma^2) \in \mathcal{C}} \mathbb{E}_f \min\{1-\phi_n', 1-\tilde{\phi}_n\} \\ &\leq \max\left\{ \sup_{f \in \mathcal{F}_n : (\boldsymbol{\beta},\sigma^2) \in \hat{\mathcal{C}}} \mathbb{E}_f(1-\phi_n'), \sup_{f \in \mathcal{F}_n : (\boldsymbol{\beta},\sigma^2) \in \tilde{\mathcal{C}}} \mathbb{E}_f(1-\tilde{\phi}_n) \right\}. \end{aligned} \tag{D.39}$$

Let $\tilde{\xi} = \{g : \|\boldsymbol{\beta}_g\|_2 > \omega_g\} \cup S_0$, $m_{\tilde{\xi}} = \sum_{g \in \tilde{\xi}} m_g$, and $\tilde{\xi}^c = \{1, \ldots, G\} \setminus \tilde{\xi}$. For any $f \in \mathcal{F}_n$ such that $(\boldsymbol{\beta}, \sigma^2) \in \hat{\mathcal{C}} \cup \tilde{\mathcal{C}}$, we must have then that $|\tilde{\xi}| \leq C_3 s_0 + s_0 \leq \bar{p}$, by Assumption (A4). By (D.33), the prior puts exponentially vanishing probability on values of $\theta > \theta_0$ where $\theta_0 = C_3 s_0 \log G / G^c < 1/(G^2+1)$ for large G. Since $\lambda_0 = (1-\theta)/\theta$ is monotonic decreasing in θ, we have that with probability greater than $1 - e^{-C_3 n \epsilon_n^2}$, $\lambda_0 \geq G^2$. Combining this fact with Assumption (A3) and using $\mathcal{F}_n$ in (D.28), we have that for any $f \in \mathcal{F}_n$, $(\boldsymbol{\beta}, \sigma^2) \in \hat{C} \cup \tilde{C}$ and sufficiently large n,

$$\begin{aligned}\|\boldsymbol{X}_{\tilde{\xi}^c}\boldsymbol{\beta}_{\tilde{\xi}^c}\|_2 &\leq \sqrt{kn^\alpha}\|\boldsymbol{\beta}_{\tilde{\xi}^c}\|_2 \\ &\leq \sqrt{kn^\alpha}\left[(G - |\tilde{\xi}|)\max_{g \in \tilde{\xi}^c} \omega_g\right] \\ &\leq \sqrt{kn^\alpha}\left\{\frac{G}{\lambda_0 - \lambda_1}\log\left[\frac{1-\theta}{\theta}\left(\frac{\lambda_0}{\lambda_1}\right)^{m_{\max}}\right]\right\} \\ &\lesssim \min\{\sqrt{k}, 1\} \times \sqrt{\nu_1}\sqrt{n}\sigma_0\epsilon_n, \end{aligned} \tag{D.40}$$

where ν is from Assumption (A4). In the above display, we used Lemma D.3 in the second inequality, while the last inequality follows from our assumptions on $(\theta, \lambda_0, \lambda_1)$ and $m_{\max}$,

so one can show that the bracketed term in the third line is asymptotically bounded above by $D\sqrt{\nu_1}\sqrt{n^{1-\alpha}}\sigma_0\epsilon_n$ for large n and any constant $D > 0$. Thus, using nearly identical arguments as those used to prove Part I of Theorem A.1 in Song and Liang (2017), we have

$$\begin{aligned}
&\sup_{f\in\mathcal{F}_n:(\boldsymbol{\beta},\sigma^2)\in\hat{\mathcal{C}}} \mathbb{E}_f(1-\phi_n') \\
&\quad\leq \sup_{f\in\mathcal{F}_n:(\boldsymbol{\beta},\sigma^2)\in\hat{\mathcal{C}}} \Pr\left(|\chi^2_{n-m_{\tilde{\xi}}}(\zeta)-(n-m_{\tilde{\xi}})| \geq (n-m_{\tilde{\xi}})\epsilon_n\right) \\
&\quad\leq \exp(-\hat{c}_4 n\epsilon_n^2), \qquad\qquad (D.41)
\end{aligned}$$

where the noncentrality parameter ζ satisfies $\zeta \leq n\epsilon_n^2\nu_1\sigma_0^2/16\sigma^2$, and the last inequality follows from the fact that the noncentral χ^2 distribution is subexponential and Bernstein's inequality (see Lemmas A.1 and A.2 in Song and Liang (2017)).

Using the arguments in Part I of the proof of Theorem A.1 in Song and Liang (2017), we also have that for large n,

$$\begin{aligned}
&\sup_{f\in\mathcal{F}_n:(\boldsymbol{\beta},\sigma^2)\in\tilde{\mathcal{C}}} \mathbb{E}_f(1-\tilde{\phi}_n) \\
&\quad\leq \sup_{f\in\mathcal{F}_n:(\boldsymbol{\beta},\sigma^2)\in\hat{\mathcal{C}}} \Pr\Big(\|(\boldsymbol{X}_{\tilde{\xi}}^T\boldsymbol{X}_{\tilde{\xi}})^{-1}\boldsymbol{X}_{\tilde{\xi}}^T\boldsymbol{\varepsilon}\|_2 \geq \big[\|\boldsymbol{\beta}_{\tilde{\xi}}-\boldsymbol{\beta}_{0\tilde{\xi}}\|_2 - \sigma_0\epsilon_n - \\
&\qquad\qquad \|(\boldsymbol{X}_{\tilde{\xi}}^T\boldsymbol{X}_{\tilde{\xi}})^{-1}\boldsymbol{X}_{\tilde{\xi}}^T\boldsymbol{X}_{\tilde{\xi}^c}\boldsymbol{\beta}_{\tilde{\xi}^c}\|_2\big]/\sigma\Big) \\
&\quad\leq \sup_{f\in\mathcal{F}_n:(\boldsymbol{\beta},\sigma^2)\in\hat{\mathcal{C}}} \Pr\left(\|(\boldsymbol{X}_{\tilde{\xi}}^T\boldsymbol{X}_{\tilde{\xi}})^{-1}\boldsymbol{X}_{\tilde{\xi}}^T\boldsymbol{\varepsilon}\|_2 \geq \epsilon_n\right) \\
&\quad\leq \sup_{f\in\mathcal{F}_n:(\boldsymbol{\beta},\sigma^2)\in\hat{\mathcal{C}}} \Pr(\chi^2_{|\tilde{\xi}|} \geq n\nu_1\epsilon_n^2) \\
&\quad\leq \exp(-\tilde{c}_4 n\epsilon_n^2), \qquad\qquad (D.42)
\end{aligned}$$

where the second inequality in the above display holds since $\|\boldsymbol{\beta}_{\tilde{\xi}}-\boldsymbol{\beta}_{0\tilde{\xi}}\|_2 \geq \|\boldsymbol{\beta}-\boldsymbol{\beta}_0\|_2 - \|\boldsymbol{\beta}_{\tilde{\xi}^c}\|_2$, and since (D.40) can be further bounded from above by $\sqrt{kn^\alpha}\sqrt{\nu_1}\sigma_0\epsilon_n$ and thus

$\|\boldsymbol{\beta}_{\tilde{\xi}^c}\| \leq \sqrt{\nu_1}\sigma_0\epsilon_n$. Therefore, we have for $f \in \mathcal{F}_n, (\boldsymbol{\beta}, \sigma^2) \in \tilde{C}$,

$$\|\boldsymbol{\beta}_{\tilde{\xi}} - \boldsymbol{\beta}_{0\tilde{\xi}}\|_2 \geq (3 + \sqrt{\nu_1})\sigma_0\epsilon_n - \sqrt{\nu_1}\sigma_0\epsilon_n = 3\sigma_0\epsilon_n,$$

while by Assumption (A4) and (D.40), we also have

$$\begin{aligned}\|(\boldsymbol{X}_{\tilde{\xi}}^T\boldsymbol{X}_{\tilde{\xi}})^{-1}\boldsymbol{X}_{\tilde{\xi}}^T\boldsymbol{X}_{\tilde{\xi}^c}\boldsymbol{\beta}_{\tilde{\xi}^c}\|_2 &\leq \sqrt{\lambda_{\max}\left((\boldsymbol{X}_{\tilde{\xi}}^T\boldsymbol{X}_{\tilde{\xi}})^{-1}\right)}\|\boldsymbol{X}_{\tilde{\xi}^c}\boldsymbol{\beta}_{\tilde{\xi}^c}\|_2 \\ &\leq \left(\sqrt{1/n\nu_1}\right)(\sqrt{n\nu_1}\sigma_0\epsilon_n) = \sigma_0\epsilon_n,\end{aligned}$$

and then we used the fact that on the set $\tilde{C}$, $\sigma = \sigma_0$. The last three inequalities in (D.42) follow from Assumption (A4), the fact that $|\widetilde{\xi}| \leq \bar{p} \asymp n\epsilon_n^2$, and the fact that for all $m > 0$, $\Pr(\chi_m^2 \geq x) \leq \exp(-x/4)$ whenever $x \geq 8m$. Altogether, combining (D.38)-(D.42), we have that

$$\begin{aligned}&\sup && \mathbb{E}_f(1 - \phi_n) \leq \exp\left(-\min\{\hat{c}_4, \tilde{c}_4\}n\epsilon_n^2\right), \\ f \in \mathcal{F}_n : \;& \|\boldsymbol{\beta} - \boldsymbol{\beta}_0\|_2 \geq (3 + \sqrt{\nu})\sigma_0\epsilon_n, \\ & \text{or } |\sigma^2 - \sigma_0^2| \geq 4\sigma_0^2\epsilon_n\end{aligned} \tag{D.43}$$

where $\hat{c}_4 > 0$ and $\tilde{c}_4 > 0$ are the constants from (D.41) and (D.42).

Now set $C_4 = \min\{\hat{c}_4, \tilde{c}_4, \check{c}_4\}$, where $\check{c}_4$ is the constant from (D.37). By and (D.37) and (D.43), this choice of C_4 will satisfy both testing conditions (D.26) and (D.27).

Since we have verified (D.9) and (D.25)-(D.27) for $\epsilon_n = \sqrt{s_0 \log G/n}$, we have

$$\Pi\left(\boldsymbol{\beta} : \|\boldsymbol{\beta} - \boldsymbol{\beta}_0\|_2 \geq (3 + \sqrt{\nu})\sigma_0\epsilon_n \middle| \boldsymbol{Y}\right) \to 0 \text{ a.s. } \mathbb{P}_0 \text{ as } n, G \to \infty,$$

and

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

i.e. we have proven (6.3) and (6.5).

Part III. Posterior contraction under prediction error loss. The proof is very similar to the proof of (6.3). The only difference is the testing conditions. We use the same sieve $\mathcal{F}_n$ as that in (D.28) so that (D.25) holds, but now, we need to show the existence of a different sequence of test functions $\tau_n \in [0,1]$ such that

$$\mathbb{E}_{f_0}\tau_n \leq e^{-C_4 n\epsilon_n^2}, \tag{D.44}$$

and

$$\sup_{\substack{f\in\mathcal{F}_n: \ \|\boldsymbol{X}\boldsymbol{\beta}-\boldsymbol{X}\boldsymbol{\beta}_0\|_2 \geq M_2\sigma_0\sqrt{n}\epsilon_n, \\ \text{or } |\sigma^2-\sigma_0^2|\geq 4\sigma_0^2\epsilon_n}} \mathbb{E}_f(1-\tau_n) \leq e^{-C_4 n\epsilon_n^2}. \tag{D.45}$$

Let $\widetilde{p}$ be the same integer from (D.36) and consider the test function $\tau_n = \max\{\tau_n', \tilde{\tau}_n\}$, where

$$\begin{aligned} \tau_n' &= \max_{\xi\supset S_0, |\xi|\leq \widetilde{p}+s_0} 1\left\{|\widehat{\sigma}_\xi^2-\sigma_0^2|\geq \sigma_0^2\epsilon_n\right\}, \quad \text{and} \\ \tilde{\tau}_n &= \max_{\xi\supset S_0, |\xi|\leq \widetilde{p}+s_0} 1\left\{\|\boldsymbol{X}_\xi\widehat{\boldsymbol{\beta}}_\xi-\boldsymbol{X}_\xi\boldsymbol{\beta}_{0\xi}\|_2\geq \sigma_0\sqrt{n}\epsilon_n\right\}. \end{aligned} \tag{D.46}$$

Using Assumption (A4) that for any $\xi\subset\{1,\ldots,G\}$ such that $|\xi|\leq\bar{p}$, $\lambda_{\max}(\boldsymbol{X}_\xi^T\boldsymbol{X}_\xi)\leq n\nu_2$ for some $\nu_2>0$, we have that

$$\|\boldsymbol{X}_\xi\widehat{\boldsymbol{\beta}}_\xi-\boldsymbol{X}_\xi\boldsymbol{\beta}_{0\xi}\|_2\leq \sqrt{n\nu_2}\|\widehat{\boldsymbol{\beta}}_\xi-\boldsymbol{\beta}_{0\xi}\|_2,$$

and so

$$\Pr\left(\|\boldsymbol{X}_\xi\widehat{\boldsymbol{\beta}}_\xi-\boldsymbol{X}_\xi\boldsymbol{\beta}_{0\xi}\|_2\geq \sigma_0\sqrt{n}\epsilon_n\right)\leq \Pr\left(\|\widehat{\boldsymbol{\beta}}_\xi-\boldsymbol{\beta}_{0\xi}\|_2\geq \nu_2^{-1/2}\sigma_0\epsilon_n\right).$$

Therefore, using similar steps as those in Part II of the proof, we can show that our chosen sequence of tests τ_n satisfies (D.44) and (D.45). We thus arrive at

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

i.e. we have proven (6.4). □

Proof of Theorem 3. According to Part I of the proof of Theorem 2, we have that for $\epsilon_n = \sqrt{s_0 \log G/n}$,

$$\Pi\left(K(f_0, f) \leq n\epsilon_n^2, V(f_0, f) \leq n\epsilon_n^2\right) \geq \exp\left(-Cn\epsilon_n^2\right)$$

for some $C > 0$. Thus, by Lemma 8.10 of Ghosal and van der Vaart (2017), there exist positive constants C_1 and C_2 such that the event,

$$E_n = \left\{ \int \int \frac{f(\boldsymbol{Y})}{f_0(\boldsymbol{Y})} d\Pi(\boldsymbol{\beta}) d\Pi(\sigma^2) \geq e^{-C_1 n\epsilon_n^2} \right\}, \tag{D.47}$$

satisfies

$$\mathbb{P}_0(E_n^c) \leq e^{-(1+C_2)n\epsilon_n^2}. \tag{D.48}$$

Define the set $\mathcal{T} = \{\boldsymbol{\beta} : |\boldsymbol{\gamma}(\boldsymbol{\beta})| \leq C_3 s_0\}$, where we choose $C_3 > 1 + C_2$. We must show that $\mathbb{E}_0 \Pi(\mathcal{T}^c|\boldsymbol{Y}) \to 0$ as $n \to \infty$. The posterior probability $\Pi(\mathcal{T}^c|\boldsymbol{Y})$ is given by

$$\Pi(\mathcal{T}^c|\boldsymbol{Y}) = \frac{\int \int_{\mathcal{T}^c} \frac{f(\boldsymbol{Y})}{f_0(\boldsymbol{Y})} d\Pi(\boldsymbol{\beta}) d\Pi(\sigma^2)}{\int \int \frac{f(\boldsymbol{Y})}{f_0(\boldsymbol{Y})} d\Pi(\boldsymbol{\beta}) d\Pi(\sigma^2)}. \tag{D.49}$$

By (D.48), the denominator of (D.49) is bounded below by $e^{-(1+C_2)n\epsilon_n^2}$. For the numerator of (D.49), we have as an upper bound,

$$\mathbb{E}_0 \left(\int \int_{\mathcal{T}^c} \frac{f(\boldsymbol{Y})}{f_0(\boldsymbol{Y})} d\Pi(\boldsymbol{\beta}) \Pi(\sigma^2) \right) \leq \int_{\mathcal{T}^c} d\Pi(\boldsymbol{\beta}) = \Pi\left(|\boldsymbol{\gamma}(\boldsymbol{\beta})| > C_3 s_0\right). \tag{D.50}$$

Using the same arguments as (D.30)-(D.34) in the proof of Theorem 2, we can show that

$$\Pi\left(|\boldsymbol{\gamma}(\boldsymbol{\beta})| > C_3 s_0\right) \prec e^{-C_3 n\epsilon_n^2}. \tag{D.51}$$

Combining (D.47)-(D.50), we have that

$$\begin{aligned}\mathbb{E}_0\Pi\left(\mathcal{T}^c|\boldsymbol{Y}\right) &\leq \mathbb{E}_0\Pi(\mathcal{T}^c|\boldsymbol{Y})1_{E_n} + \mathbb{P}_0(E_n^c) \\ &< \exp\left((1+C_2)n\epsilon_n^2 - C_3 n\epsilon_n^2\right) + o(1) \\ &\to 0 \text{ as } n, G \to \infty,\end{aligned}$$

since $C_3 > 1 + C_2$. This proves (6.8). □

Proof of Theorem 4. Let $f_{0j}(\boldsymbol{X}_j)$ be an $n \times 1$ vector with ith entry equal to $f_{0j}(X_{ij})$. Note that proving posterior contraction with respect to the empirical norm (6.11) is equivalent to proving that

$$\Pi\left(\boldsymbol{\beta} : \|\widetilde{\boldsymbol{X}}\boldsymbol{\beta} - \sum_{j=1}^{p} f_{0j}(\boldsymbol{X}_j)\|_2 \geq \widetilde{M}_1\sqrt{n}\epsilon_n \middle| \boldsymbol{Y}\right) \to 0 \text{ a.s. } \widetilde{\mathbb{P}}_0 \text{ as } n, p \to \infty, \tag{D.52}$$

so to prove the theorem, it suffices to prove (D.52). Let $f \sim \mathcal{N}_n(\widetilde{\boldsymbol{X}}\boldsymbol{\beta}, \sigma^2\boldsymbol{I}_n)$ and $f_0 \sim \mathcal{N}_n(\widetilde{\boldsymbol{X}}\boldsymbol{\beta}_0 + \boldsymbol{\delta}, \sigma_0^2\boldsymbol{I}_n)$, and let $\Pi(\cdot)$ denote the prior (6.2). Similar to the proof for Theorem 2, we show that for our choice of $\epsilon_n^2 = s_0 \log p/n + s_0 n^{-2\kappa/(2\kappa+1)}$ and some constant $C_1 > 0$,

$$\Pi\left(K(f_0, f) \leq n\epsilon_n^2, V(f_0, f) \leq n\epsilon_n^2\right) \geq \exp(-C_1 n\epsilon_n^2), \tag{D.53}$$

and the existence of a sieve $\mathcal{F}_n$ such that

$$\Pi(\mathcal{F}_n^c) \leq \exp(-C_2 n\epsilon_n^2), \tag{D.54}$$

for positive constant $C_2 > C_1 + 2$, and a sequence of test functions $\phi_n \in [0, 1]$ such that

$$\mathbb{E}_{f_0}\phi_n \leq e^{-C_4 n\epsilon_n^2}, \tag{D.55}$$

and

$$\sup_{f \in \mathcal{F}_n : \ \begin{array}{l} \|\widetilde{\boldsymbol{X}}\boldsymbol{\beta} - \sum_{j=1}^p f_{0j}(\boldsymbol{X}_j)\|_2 \geq \tilde{c}_0 \sigma_0 \sqrt{n} \epsilon_n, \\ \text{or } |\sigma^2 - \sigma_0^2| \geq 4\sigma_0^2 \epsilon_n \end{array}} \mathbb{E}_f(1 - \phi_n) \leq e^{-C_4 n \epsilon_n^2}, \tag{D.56}$$

for some $C_4 > 0$ and $\tilde{c}_0 > 0$.

We first verify (D.53). The KL divergence between f_0 and f is

$$K(f_0, f) = \frac{1}{2}\left[n\left(\frac{\sigma_0^2}{\sigma^2}\right) - n - n\log\left(\frac{\sigma_0^2}{\sigma^2}\right) + \frac{\|\widetilde{\boldsymbol{X}}(\boldsymbol{\beta} - \boldsymbol{\beta}_0) - \boldsymbol{\delta}\|_2^2}{\sigma^2} \right], \tag{D.57}$$

and the KL variation between f_0 and f is

$$V(f_0, f) = \frac{1}{2}\left[n\left(\frac{\sigma_0^2}{\sigma^2}\right)^2 - 2n\left(\frac{\sigma_0^2}{\sigma^2}\right) + n \right] + \frac{\sigma_0^2}{(\sigma^2)^2}\|\widetilde{\boldsymbol{X}}(\boldsymbol{\beta} - \boldsymbol{\beta}_0) - \boldsymbol{\delta}\|_2^2. \tag{D.58}$$

Define the two events $\widetilde{\mathcal{A}}_1$ and $\widetilde{\mathcal{A}}_2$ as follows:

$$\widetilde{\mathcal{A}}_1 = \left\{ \sigma^2 : n\left(\frac{\sigma_0^2}{\sigma^2}\right) - n - n\log\left(\frac{\sigma_0^2}{\sigma^2}\right) \leq n\epsilon_n^2, \ \ n\left(\frac{\sigma_0^2}{\sigma^2}\right)^2 - 2n\left(\frac{\sigma_0^2}{\sigma^2}\right) + n \leq n\epsilon_n^2 \right\} \tag{D.59}$$

and

$$\widetilde{\mathcal{A}}_2 = \left\{ (\boldsymbol{\beta}, \sigma^2) : \frac{\|\widetilde{\boldsymbol{X}}(\boldsymbol{\beta} - \boldsymbol{\beta}_0) - \boldsymbol{\delta}\|_2^2}{\sigma^2} \leq n\epsilon_n^2, \ \ \frac{\sigma_0^2}{(\sigma^2)^2}\|\widetilde{\boldsymbol{X}}(\boldsymbol{\beta} - \boldsymbol{\beta}_0) - \boldsymbol{\delta}\|_2^2 \leq n\epsilon_n^2/2 \right\}. \tag{D.60}$$

Following from (D.57)-(D.60), we have $\Pi(K(f_0, f) \leq n\epsilon_n^2, V(f_0, f) \leq n\epsilon_n^2) = \Pi(\widetilde{\mathcal{A}}_2 | \widetilde{\mathcal{A}}_1)\Pi(\widetilde{\mathcal{A}}_1)$. Using the steps we used to prove (D.15) in part I of the proof of Theorem 2, we have

$$\Pi(\widetilde{\mathcal{A}}_1) \gtrsim \exp(-C_1 n\epsilon_n^2/2), \tag{D.61}$$

for some sufficiently large $C_1 > 0$. Following similar reasoning as in the proof of Theorem 2, we also have for some $b_2 > 0$,

$$\Pi\left(\widetilde{A}_2 | \widetilde{A}_1\right) \geq \Pi\left(\|\widetilde{\boldsymbol{X}}(\boldsymbol{\beta} - \boldsymbol{\beta}_0) - \boldsymbol{\delta}\|_2^2 \leq \frac{b_2^2 n\epsilon_n^2}{2} \right). \tag{D.62}$$

Using Assumptions (B3) and (B6), we then have

$$\begin{aligned}\|\widetilde{\boldsymbol{X}}(\boldsymbol{\beta}-\boldsymbol{\beta}_0)-\boldsymbol{\delta}\|_2^2 &\leq \left(\|\widetilde{\boldsymbol{X}}(\boldsymbol{\beta}-\boldsymbol{\beta}_0)\|_2+\|\boldsymbol{\delta}\|_2\right)^2\\ &\leq 2\|\widetilde{\boldsymbol{X}}(\boldsymbol{\beta}-\boldsymbol{\beta}_0)\|_2^2+2\|\boldsymbol{\delta}\|_2^2\\ &\lesssim 2\left(nk_1\|\boldsymbol{\beta}-\boldsymbol{\beta}_0\|_2^2+\frac{k_1b_2^2ns_0d^{-2\kappa}}{4}\right)\\ &\asymp 2n\left(\|\boldsymbol{\beta}-\boldsymbol{\beta}_0\|_2^2+\frac{b_2^2s_0d^{-2\kappa}}{4}\right),\end{aligned}$$

and so (D.62) can be asymptotically lower bounded by

$$\begin{aligned}&\Pi\left(\|\boldsymbol{\beta}-\boldsymbol{\beta}_0\|_2^2+\frac{b_2^2s_0d^{-2\kappa}}{4}\leq\frac{b_2^2\epsilon_n^2}{4}\right)\\ &=\Pi\left(\|\boldsymbol{\beta}-\boldsymbol{\beta}_0\|_2^2\leq\frac{b_2^2}{4}\left(\epsilon_n^2-s_0n^{-2\kappa/(2\kappa+1)}\right)\right),\end{aligned}$$

where we used Assumption (B1) that $d\asymp n^{1/(2\kappa+1)}$. Using very similar arguments as those used to prove (D.24), this term can also be lower bounded by $\exp(-C_1n\epsilon_n^2/2)$. Altogether, we have

$$\Pi(\widetilde{A}_2|\widetilde{A}_1)\gtrsim\exp(-C_1\epsilon_n^2/2). \tag{D.63}$$

Combining (D.61) and (D.63), we have that (D.53) holds. To verify (D.54), we choose $C_3\geq C_1+2+\log 3$ and use the same sieve $\mathcal{F}_n$ as the one we employed in the proof of Theorem 2 (eq. (D.28)), and then (D.54) holds for our choice of $\mathcal{F}_n$.

Finally, we follow the recipe of Wei et al. (2020) and Song and Liang (2017) to construct our test function ϕ_n which will satisfy both (D.55) and (D.56). For $\xi\subset\{1,\ldots,p\}$, let $\widetilde{\boldsymbol{X}}_\xi$ denote the submatrix of $\widetilde{\boldsymbol{X}}$ with submatrices indexed by ξ, where $|\xi|\leq\bar{p}$ and $\bar{p}$ is from Assumption (B4). Let $\widehat{\boldsymbol{\beta}}_\xi=(\widetilde{\boldsymbol{X}}_\xi^T\widetilde{\boldsymbol{X}}_\xi)^{-1}\widetilde{\boldsymbol{X}}_\xi^T\boldsymbol{Y}$ and $\boldsymbol{\beta}_{0\xi}$ denote the subvector of $\boldsymbol{\beta}_0$ with basis coefficients appearing in ξ. Then the total number of elements in $\widehat{\boldsymbol{\beta}}_\xi$ is $d|\xi|$. Finally,

let $\widehat{\sigma}_\xi^2 = \boldsymbol{Y}^T(\boldsymbol{I}_n - \boldsymbol{H}_\xi)\boldsymbol{Y}/(n - d|\xi|)$, where $\boldsymbol{H}_\xi = \widetilde{\boldsymbol{X}}_\xi(\widetilde{\boldsymbol{X}}_\xi^T\widetilde{\boldsymbol{X}}_\xi)^{-1}\widetilde{\boldsymbol{X}}_\xi^T$ is the hat matrix for the subgroup ξ.

Let $\widetilde{p}$ be an integer satisfying $\widetilde{p} \asymp s_0$ and $\widetilde{p} \leq \bar{p} - s_0$, where $\bar{p}$ is from Assumption (B4) and the specific choice for $\widetilde{p}$ will be given later. Recall that S_0 is the set of true nonzero groups with cardinality $s_0 = |S_0|$. Similar to Wei et al. (2020), we consider the test function, $\phi_n = \max\{\phi_n', \tilde{\phi}_n\}$, where

$$\begin{aligned} \phi_n' &= \max_{\xi \supset S_0, |\xi| \leq \widetilde{p}+s_0} 1\left\{|\widehat{\sigma}_\xi^2 - \sigma_0^2| \geq c_0'\sigma_0^2\epsilon_n\right\}, \qquad \text{and} \\ \tilde{\phi}_n &= \max_{\xi \supset S_0, |\xi| \leq \widetilde{p}+s_0} 1\left\{\left|\left|\widetilde{\boldsymbol{X}}\widehat{\boldsymbol{\beta}}_\xi - \sum_{j\in\xi} f_{0j}(\boldsymbol{X}_j)\right|\right|_2 \geq \tilde{c}_0\sigma_0\sqrt{n}\epsilon_n\right\}, \end{aligned} \tag{D.64}$$

for some positive constants c_0' and $\tilde{c}_0$. Using Assumptions (B1) and (B4), we have that for any ξ in our test ϕ_n, $d|\xi| \leq d(\widetilde{p} + s_0) \leq d\bar{p} \prec n\epsilon_n^2$. Using essentially the same arguments as those in the proof for Theorem 4.1 in Wei et al. (2020), we have that for any ξ which satisfies $\xi \supset S_0$ so that $|\xi| \leq \widetilde{p} + s_0$,

$$\mathbb{E}_{(\boldsymbol{\beta}_0,\sigma_0^2)} 1\left\{|\widehat{\sigma}_\xi^2 - \sigma_0^2| \geq c_0'\epsilon_n\right\} \leq \exp(-c_4'n\epsilon_n^2), \tag{D.65}$$

for some $c_0'' > 0$. By Assumption (B3), we also have

$$\begin{aligned} \left|\left|\widetilde{\boldsymbol{X}}\widehat{\boldsymbol{\beta}} - \sum_{j=1}^p f_{0j}(\boldsymbol{X}_j)\right|\right|_2 &= \|\widetilde{\boldsymbol{X}}(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0) - \boldsymbol{\delta}\|_2 \\ &\leq \sqrt{nk_1}\|\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0\|_2 + \|\boldsymbol{\delta}\|_2, \end{aligned}$$

and using the fact that $\|\boldsymbol{\delta}\|_2 \lesssim \sqrt{ns_0}d^{-\kappa} \lesssim \tilde{c}_0\sigma_0\sqrt{n}\epsilon_n/2$ (by Assumptions (B1) and (B6)), we have that for any ξ such that $\xi \supset S_0, |\xi| \leq \widetilde{p} + s_0$,

$$\mathbb{E}_{(\boldsymbol{\beta}_0,\sigma_0^2)} 1\left\{\left|\left|\widetilde{\boldsymbol{X}}\widehat{\boldsymbol{\beta}} - \sum_{j=1}^p f_{0j}(\boldsymbol{X}_j)\right|\right|_2 \geq \tilde{c}_0\sigma_0\sqrt{n}\epsilon_n\right\}$$

$$\leq \mathbb{E}_{(\boldsymbol{\beta}_0,\sigma_0^2)}\left\{\|\widehat{\boldsymbol{\beta}}-\boldsymbol{\beta}_0\|_2 \geq \tilde{c}_0\sigma_0\epsilon_n/2\sqrt{k_1}\right\}$$
$$\leq \exp(-\tilde{c}_4 n\epsilon_n^2),$$

for some $\tilde{c}_4 > 0$, where we used the proof of Theorem A.1 in Song and Liang (2017) to arrive at the final inequality. Again, as in the proof of Theorem A.1 of Song and Liang (2017), we choose $\widetilde{p} = \lfloor \min\{c_4', \tilde{c}_4\} n\epsilon_n^2/(2\log p)\rfloor$, and then

$$\mathbb{E}_{f_0}\phi_n \leq \exp(-\check{c}_4 n\epsilon_n^2), \tag{D.66}$$

for some $\check{c}_4 > 0$. Next, we define the set,

$$\mathcal{C} = \left\{\|\widetilde{\boldsymbol{X}}\boldsymbol{\beta} - \textstyle\sum_{j=1}^p f_{0j}(\boldsymbol{X}_j)\|_2 \geq \tilde{c}_0\sigma_0\sqrt{n}\epsilon_n \text{ or } \sigma^2/\sigma_0^2 > (1+\epsilon_n)/(1-\epsilon_n) \right.$$
$$\left. \text{or } \sigma^2/\sigma_0^2 < (1-\epsilon_n)/(1+\epsilon_n)\right\}.$$

By Lemma D.2, we have

$$\sup_{f\in\mathcal{F}_n:\ \|\widetilde{\boldsymbol{X}}\boldsymbol{\beta} - \sum_{j=1}^p f_{0j}(\boldsymbol{X}_j)\|_2 \geq \tilde{c}_0\sigma_0\sqrt{n}\epsilon_n, \text{ or } |\sigma^2-\sigma_0^2| \geq 4\sigma_0^2\epsilon_n} \mathbb{E}_f(1-\phi_n)$$
$$\leq \sup_{f\in\mathcal{F}_n:(\boldsymbol{\beta},\sigma^2)\in\mathcal{C}} \mathbb{E}_f(1-\phi_n). \tag{D.67}$$

Similar to Song and Liang (2017), we consider $\mathcal{C} \subset \widehat{\mathcal{C}} \cup \widetilde{\mathcal{C}}$, where

$$\widehat{\mathcal{C}} = \{\sigma^2/\sigma_0^2 > (1+\epsilon_n)/(1-\epsilon_n) \text{ or } \sigma^2/\sigma_0^2 < (1-\epsilon_n)/(1+\epsilon_n)\},$$
$$\tilde{\mathcal{C}} = \{\|\widetilde{\boldsymbol{X}}\boldsymbol{\beta} - \sum_{j=1}^p f_{0j}(\boldsymbol{X}_j)\|_2 \geq \tilde{c}_0\sigma_0\epsilon_n \text{ and } \sigma^2 = \sigma_0^2\},$$

and so an upper bound for (D.67) is

$$\sup_{f\in\mathcal{F}_n:(\boldsymbol{\beta},\sigma^2)\in\mathcal{C}} \mathbb{E}_f(1-\phi_n) = \sup_{f\in\mathcal{F}_n:(\boldsymbol{\beta},\sigma^2)\in\mathcal{C}} \mathbb{E}_f \min\{1-\phi_n', 1-\tilde{\phi}_n\}$$

$$\leq \max\left\{ \sup_{f\in\mathcal{F}_n:(\boldsymbol{\beta},\sigma^2)\in\hat{\mathcal{C}}} \mathbb{E}_f(1-\phi_n'), \sup_{f\in\mathcal{F}_n:(\boldsymbol{\beta},\sigma^2)\in\tilde{\mathcal{C}}} \mathbb{E}_f(1-\tilde{\phi}_n) \right\}. \tag{D.68}$$

Using very similar arguments as those used to prove (D.43) in Theorem 2 and using Assumptions (B1) and (B6), so that the bias $\|\boldsymbol{\delta}\|_2^2 \lesssim n s_0 d^{-2\kappa} \lesssim n\epsilon_n^2$, we can show that (D.68) can be further bounded from above as

$$\begin{aligned} &\sup_{f\in\mathcal{F}_n:\ \|\widetilde{\boldsymbol{X}}\boldsymbol{\beta}-\sum_{j=1}^p f_{0j}(\boldsymbol{X}_j)\|_2 \geq \tilde{c}_0\sigma_0\sqrt{n}\epsilon_n,\ \text{or } |\sigma^2-\sigma_0^2|\geq 4\sigma_0^2\epsilon_n} \mathbb{E}_f(1-\phi_n) \\ &\leq \exp\left(-\min\{\hat{c}_4,\tilde{c}_4\}n\epsilon_n^2\right), \end{aligned} \tag{D.69}$$

where $\hat{c}_4 > 0$ and $\tilde{c}_4 > 0$ are the constants from (D.65) and (D.66).

Choose $C_4 = \min\{\check{c}_4, \hat{c}_4, \tilde{c}_4\}$, and we have from (D.66) and (D.69) that (D.55) and (D.56) both hold.

Since we have verified (D.53) and (D.54)-(D.56) for our choice of $\epsilon_n^2 = s_0 \log p/n + s_0 n^{-2\kappa/(2\kappa+1)}$, it follows that

$$\Pi\left(\boldsymbol{\beta}: \left|\left|\widetilde{\boldsymbol{X}}\boldsymbol{\beta}-\sum_{j=1}^p f_{0j}(\boldsymbol{X}_j)\right|\right|_2 \geq \tilde{c}_0\sigma_0\sqrt{n}\epsilon_n|\boldsymbol{Y}\right) \to 0 \text{ a.s. } \widetilde{\mathbb{P}}_0 \text{ as } n,p\to\infty,$$

and

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

i.e. we have proven (D.52), or equivalently, (6.11) and (6.12).

□

Proof of Theorem 5. The proof is very similar to the proof of Theorem 3 and is thus omitted. □

References

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

Ghosal, S. and A. van der Vaart (2017). *Fundamentals of Nonparametric Bayesian Inference*. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press.

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

Kuhn, M. (2008). Building predictive models in R using the caret package. *Journal of Statistical Software 28*(5), 1–26.

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

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

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

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

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

Song, Q. and F. Liang (2017). Nearly optimal Bayesian shrinkage for high dimensional regression. *arXiv pre-print arXiv: 1712.08964*.

Storlie, C. B., H. D. Bondell, B. J. Reich, and H. H. Zhang (2011). Surface estimation, variable selection, and the nonparametric oracle property. *Statistica Sinica 21*(2), 679.

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

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

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

Zhang, C.-H. and T. Zhang (2012). A general theory of concave regularization for high-dimensional sparse estimation problems. *Statistical Science 27*(4), 576–593.