entropy

MDPI

Article

Variable Selection Using Nonlocal Priors in High-Dimensional Generalized Linear Models With Application to fMRI Data Analysis

Xuan Cao [1] **and Kyoungjae Lee** [2,*]

[1] Department of Mathematical Sciences, University of Cincinnati, Cincinnati, OH 45221, USA; caox4@ucmail.uc.edu

[2] Department of Statistics, Inha University, Incheon 22212, Korea

* Correspondence: leekjstat@gmail.com

Received: 8 June 2020; Accepted: 21 July 2020; Published: 23 July 2020

Abstract: High-dimensional variable selection is an important research topic in modern statistics. While methods using nonlocal priors have been thoroughly studied for variable selection in linear regression, the crucial high-dimensional model selection properties for nonlocal priors in generalized linear models have not been investigated. In this paper, we consider a hierarchical generalized linear regression model with the product moment nonlocal prior over coefficients and examine its properties. Under standard regularity assumptions, we establish strong model selection consistency in a high-dimensional setting, where the number of covariates is allowed to increase at a sub-exponential rate with the sample size. The Laplace approximation is implemented for computing the posterior probabilities and the shotgun stochastic search procedure is suggested for exploring the posterior space. The proposed method is validated through simulation studies and illustrated by a real data example on functional activity analysis in fMRI study for predicting Parkinson's disease.

Keywords: high-dimensional; nonlocal prior; strong selection consistency

1. Introduction

With the increasing ability to collect and store data in large scales, we are facing the opportunities and challenges to analyze data with a large number of covariates per observation, the so-called high-dimensional problem. When this situation arises, variable selection is one of the most commonly used techniques, especially in radiological and genetic research, due to the nature of high-dimensional data extracted from imaging scans and gene sequencing. In the context of regression, when the number of covariates is greater than the sample size, the parameter estimation problem becomes ill posed, and variable selection is usually the first step for dimension reduction.

A good amount of work has recently been done for variable selection from both frequentist and Bayesian perspectives. On the frequentist side, extensive studies on variable selection have emerged ever since the appearance of least absolute shrinkage and selection operator (Lasso) [1]. Other penalization approaches for sparse model selection including smoothly clipped absolute deviation (SCAD) [2], minimum concave penalty (MCP) [3] and many variations have also been introduced. Most of these methods are first considered in the context of linear regression and then extended to generalized linear models. Because all the methods share the basic desire of shrinkage toward sparse models, it has been understood that most of these frequentist methods can be interpreted from a Bayesian perspective and many analogous Bayesian methods have also been proposed. See for example [4–6] that discuss the connection between penalized likelihood-based methods and Bayesian approaches. These Bayesian methods employed local priors, which still preserve positive values at null parameter values, to achieve desirable shrinkage.

In this paper, we are interested in nonlocal densities [7] that are identically zero whenever a model parameter is equal to its null value. Compared to local priors, nonlocal prior distributions have relatively appealing properties for Bayesian model selection. In particular, nonlocal priors discard spurious covariates faster as the sample size grows, while preserving exponential learning rates to detect nontrivial coefficients [7]. Johnson and Rossell [8] and Shin et al. [9] study the behavior of nonlocal densities for variable selection in a linear regression setting. When the number of covariates is much smaller than the sample size, [10] establish the posterior convergence rate for nonlocal priors in a logistic regression model and suggest a Metropolis–Hastings algorithm for computation.

To the best of our knowledge, a rigorous investigation of high-dimensional posterior consistency properties for nonlocal priors has not been undertaken in the context of generalized linear regression. Although [11] investigated the model selection consistency of nonlocal priors in generalized linear models, they assumed a fixed dimension p. Motivated by this gap, our first goal was to examine the model selection property for nonlocal priors, particularly, the product moment (pMOM) prior [8] in a high-dimensional generalized linear model. It is known that the computation problem can arise for Bayesian approaches due to the non-conjugate structure in generalized linear regression. Hence, our second goal was to develop efficient algorithms for exploring the massive posterior space. These were challenging goals of course, as the posterior distributions are not available in closed form for this type of nonlocal priors.

As the main contributions of this paper, we first establish model selection consistency for generalized linear models with pMOM prior on regression coefficients (Theorems 1–3) when the number of covariates grows at a sub-exponential rate of the sample size. Next, n terms of computation, we first obtain the posteriors via Laplace approximation and then implement an efficient shotgun stochastic search (SSS) algorithm for exploring the sparsity pattern of the regression coefficients. In particular, the SSS-based methods have been shown to significantly reduce the computational time compared with standard Markov chain Monte Carlo (MCMC) algorithms in various settings [9,12,13]. We demonstrate that our model can outperform existing state-of-the-art methods including both penalized likelihood and Bayesian approaches in different settings. Finally, the proposed method is applied to a functional Magnetic Resonance Imaging (fMRI) data set for identifying alternative brain activities and for predicting Parkinson's disease.

The rest of paper is organized as follows. Section 2 provides background material regarding generalized linear models and revisits the pMOM distribution. We detail strong selection consistency results in Section 3, and proofs are provided in the Appendix A. The posterior computation algorithm is described in Section 4, and we show the performance of the proposed method and compare it with other competitors through simulation studies in Section 5. In Section 6, we conduct a real data analysis for predicting Parkinson's disease and show our method yields better prediction performance compared with other contenders. To conclude our paper, a discussion is given in Section 7.

2. Preliminaries

2.1. Model Specification for Logistic Regression

We first describe the framework for Bayesian variable selection in logistic regression followed by our hierarchical model specification. Let $\boldsymbol{y} \in \{0,1\}^n$ be the binary response vector and $\boldsymbol{X} = (x_{ij}) \in \mathbb{R}^{n\times p}$ be the design matrix. Without loss of generality, we assume that the columns of $\boldsymbol{X}$ are standardized to have zero mean and unit variance. Let $\boldsymbol{x}_i \in \mathbb{R}^p$ denote the ith row vector of $\boldsymbol{X}$ that contains the covariates for the ith subject. Let $\boldsymbol{\beta}$ be the $p \times 1$ vector of regression coefficients. We first consider the following standard logistic regression model:

$$\mathbb{P}\left(y_i = 1 \mid \boldsymbol{x}_i, \boldsymbol{\beta}\right) = \frac{\exp\left(\boldsymbol{x}_i^\top \boldsymbol{\beta}\right)}{1+\exp\left(\boldsymbol{x}_i^\top \boldsymbol{\beta}\right)}, \quad i = 1, 2, \ldots, n, \tag{1}$$

We will work in a scenario where the dimension of predictors, p grows with the sample size n. Thus, we consider the number of predictors is function of n, i.e., $p = p_n$, but we denote it as p for notational simplicity.

Our goal is variable selection, i.e., the correct identification of all non-zero regression coefficients. In light of that, we denote a model by $\boldsymbol{k} = \left\{k_1, k_2, \ldots, k_{|\boldsymbol{k}|}\right\} \subseteq [p] =: \{1, 2, \ldots, p\}$ if and only if all the nonzero elements of $\boldsymbol{\beta}$ are $\beta_{k_1}, \beta_{k_2}, \ldots, \beta_{k_{|\boldsymbol{k}|}}$ and denote $\boldsymbol{\beta}_{\boldsymbol{k}} = \left(\beta_{k_1}, \beta_{k_2}, \ldots, \beta_{k_{|\boldsymbol{k}|}}\right)^\top$, where $|\boldsymbol{k}|$ is the cardinality of $\boldsymbol{k}$. For any $m \times p$ matrix A, let $\boldsymbol{A}_{\boldsymbol{k}} \in \mathbb{R}^{m \times |\boldsymbol{k}|}$ denote the submatrix of $\boldsymbol{A}$ containing the columns of $\boldsymbol{A}$ indexed by model $\boldsymbol{k}$. In particular, for $1 \leq i \leq n$, we denote $\boldsymbol{x}_{i\boldsymbol{k}}$ as the subvector of $\boldsymbol{x}_i$ containing the entries of $\boldsymbol{x}_i$ corresponding to model $\boldsymbol{k}$.

The class of pMOM densities [8] can be used for model selection through the following hierarchical model

$$\pi\left(\boldsymbol{\beta}_{\boldsymbol{k}} \mid \tau, \boldsymbol{k}\right) = d_{\boldsymbol{k}} (2\pi)^{-\frac{|\boldsymbol{k}|}{2}} (\tau)^{-r|\boldsymbol{k}| - \frac{|\boldsymbol{k}|}{2}} |\boldsymbol{U}_{\boldsymbol{k}}|^{\frac{1}{2}} \exp\left(-\frac{\boldsymbol{\beta}_{\boldsymbol{k}}^\top \boldsymbol{U}_{\boldsymbol{k}} \boldsymbol{\beta}_{\boldsymbol{k}}}{2\tau}\right) \prod_{i=1}^{|\boldsymbol{k}|} \beta_{k_i}^{2r}, \tag{2}$$

$$\pi(\boldsymbol{k}) \propto I(|\boldsymbol{k}| \leq m_n). \tag{3}$$

Here $\boldsymbol{U}$ is a $p \times p$ nonsingular matrix, r is a positive integer referred to as the order of the density and $d_{\boldsymbol{k}}$ is the normalizing constant independent of the positive constant τ. Please note that prior (2) is obtained as the product of the density of multivariate normal distribution and even powers of parameters, $\prod_{i=1}^{|\boldsymbol{k}|} \beta_{k_i}^{2r}$. This results in $\pi(\boldsymbol{\beta}_{\boldsymbol{k}} \mid \tau, \boldsymbol{k}) = 0$ at $\boldsymbol{\beta}_{\boldsymbol{k}} = 0$, which is desirable because (2) is a prior for the nonzero elements of $\boldsymbol{\beta}$. Some standard regularity assumptions on the hyperparameters will be provided later in Section 3. In (3), $m_n \in [p]$ is a positive integer restricting the size of the largest model, and a uniform prior is placed on the model space restricting our analysis to models having size less than or equal to m_n. Similar structure has also been considered in [5,9,14]. An alternative is to use a complexity prior [15] that takes the form of

$$\pi(\boldsymbol{k}) \propto c_1^{-|\boldsymbol{k}|} p^{-c_2|\boldsymbol{k}|},$$

for some positive constants c_1, c_2. The essence is to force the estimated model to be sparse by penalizing dense models. As noted in [9], the model selection consistency result based on the nonlocal priors derives strength directly from the marginal likelihood and does not require strong penalty over model size. This is indeed reflected in the simulation studies in [14], where the authors compare the model selection performance under uniform prior and complexity prior. The result under uniform prior is much better than that under complexity prior, as the complexity prior always tends to prefer the sparse models.

By the hierarchical model (1) to (3) and Bayes' rule, the resulting posterior probability for model $\boldsymbol{k}$ is denoted by,

$$\pi(\boldsymbol{k}|\boldsymbol{y}) = \frac{\pi(\boldsymbol{k})}{\pi(\boldsymbol{y})} m_{\boldsymbol{k}}(\boldsymbol{y}), \tag{4}$$

where $\pi(\boldsymbol{y})$ is the marginal density of $\boldsymbol{y}$, and $m_{\boldsymbol{k}}(\boldsymbol{y})$ is the marginal density of $\boldsymbol{y}$ under model $\boldsymbol{k}$ given by

$$\begin{aligned} m_{\boldsymbol{k}}(\boldsymbol{y}) &= \int \exp\left\{L_n(\boldsymbol{\beta}_{\boldsymbol{k}})\right\} \pi\left(\boldsymbol{\beta}_{\boldsymbol{k}} \mid \boldsymbol{k}\right) d\boldsymbol{\beta}_{\boldsymbol{k}} \\ &= \int \exp\left\{L_n(\boldsymbol{\beta}_{\boldsymbol{k}})\right\} d_{\boldsymbol{k}} (2\pi)^{-|\boldsymbol{k}|/2} (\tau)^{-r|\boldsymbol{k}| - |\boldsymbol{k}|/2} |\boldsymbol{U}_{\boldsymbol{k}}|^{\frac{1}{2}} \exp\left(-\frac{\boldsymbol{\beta}_{\boldsymbol{k}}^\top \boldsymbol{U}_{\boldsymbol{k}} \boldsymbol{\beta}_{\boldsymbol{k}}}{2\tau}\right) \prod_{i=1}^{|\boldsymbol{k}|} \beta_{k_i}^{2r} d\boldsymbol{\beta}_{\boldsymbol{k}}, \end{aligned} \tag{5}$$

where

$$L_n(\boldsymbol{\beta}_{\boldsymbol{k}}) = \log\left(\prod_{i=1}^{n}\left\{\frac{\exp\left(\boldsymbol{x}_{i\boldsymbol{k}}^{\top}\boldsymbol{\beta}_{\boldsymbol{k}}\right)}{1+\exp\left(\boldsymbol{x}_{i\boldsymbol{k}}^{\top}\boldsymbol{\beta}_{\boldsymbol{k}}\right)}\right\}^{y_i}\left\{\frac{1}{1+\exp\left(\boldsymbol{x}_{i\boldsymbol{k}}^{\top}\boldsymbol{\beta}_{\boldsymbol{k}}\right)}\right\}^{1-y_i}\right) \tag{6}$$

is the log likelihood function. In particular, these posterior probabilities can be used to select a model by computing the posterior mode defined by

$$\hat{\boldsymbol{k}} = arg\, max_{\boldsymbol{k}}\pi(\boldsymbol{k}|\boldsymbol{y}). \tag{7}$$

Of course, the closed form of these posterior probabilities cannot be obtained due to not only the nature of logistic regression but also the structure of nonlocal prior. Therefore, special efforts need to be devoted to both consistency analysis and computational strategy as we shall see in the following sections.

2.2. Extension to Generalized Linear Model

We can easily extend our previous discussion on logistic regression to a generalized linear model (GLM) [16]. Given predictors $\boldsymbol{x}_i$ and an outcome y_i for $1 \leq i \leq n$, a probability density function (or probability mass function) of a generalized linear model has the following form of the exponential family

$$p(y_i|\theta) = \exp\left\{a(\theta)y_i + b(\theta) + c(y_i)\right\},$$

in which $a(\cdot)$ is a continuously differentiable function with respect to θ with nonzero derivative, $b(\cdot)$ is also a continuously differentiable function of θ, $c(\cdot)$ is some constant function of y, and θ is also known as the natural parameter that relates the response to the predictors through the linear function $\theta_i = \theta_i(\boldsymbol{\beta}) = \boldsymbol{x}_i^{\top}\boldsymbol{\beta}$. The mean function is $\mu = E(y_i|\boldsymbol{x}_i) = -b'(\theta_i)/a'(\theta_i) \triangleq \phi(\theta_i)$, where $\phi(\cdot)$ is the inverse of some chosen link function.

The class of pMOM densities specified in (2) can still be used for model selection in this generalized setting by noting that the log likelihood function in (5) and (6) now takes the general form of

$$L_n(\boldsymbol{\beta}_{\boldsymbol{k}}) = \sum_{i=1}^{n}\left\{a(\theta_i(\boldsymbol{\beta}_{\boldsymbol{k}}))y_i + b(\theta_i(\boldsymbol{\beta}_{\boldsymbol{k}})) + c(y_i)\right\}. \tag{8}$$

After obtaining the posterior probabilities in (4) with the log likelihood substituted as (8), we can select a model by computing the posterior mode. In Section 4, we will adopt some search algorithm that use these posterior probabilities to target the mode in a more efficient way.

3. Main Results

In this section, we show that the proposed Bayesian model enjoys desirable theoretical properties. Let $\boldsymbol{t} \subseteq [p]$ be the true model, which means that the nonzero locations of the true coefficient vector are $\boldsymbol{t} = (j, j \in \boldsymbol{t})$. We consider $\boldsymbol{t}$ to be a fixed vector. Let $\boldsymbol{\beta}_0 \in \mathbb{R}^p$ be the true coefficient vector and $\boldsymbol{\beta}_{0,\boldsymbol{t}} \in \mathbb{R}^{|\boldsymbol{t}|}$ be the vector of the true nonzero coefficients. In the following analysis, we will focus on logistic regression, but our argument can be easily extended to any other GLM as well. In particular,

$$\boldsymbol{H}_n(\boldsymbol{\beta}_{\boldsymbol{k}}) \;=\; -\frac{\partial^2 L_n(\boldsymbol{\beta}_{\boldsymbol{k}})}{\partial\boldsymbol{\beta}_{\boldsymbol{k}}\partial\boldsymbol{\beta}_{\boldsymbol{k}}^{\top}} \;=\; \sum_{i=1}^{n}\sigma_i^2(\boldsymbol{\beta}_{\boldsymbol{k}})\boldsymbol{x}_i\boldsymbol{x}_i^{\top} \;=\; \boldsymbol{X}_{\boldsymbol{k}}^{\top}\boldsymbol{\Sigma}(\boldsymbol{\beta}_{\boldsymbol{k}})\boldsymbol{X}_{\boldsymbol{k}}$$

as the negative Hessian of $L_n(\boldsymbol{\beta}_{\boldsymbol{k}})$, where $\boldsymbol{\Sigma}(\boldsymbol{\beta}_{\boldsymbol{k}}) \equiv \boldsymbol{\Sigma}_{\boldsymbol{k}} = diag(\sigma_1^2(\boldsymbol{\beta}_{\boldsymbol{k}}), \ldots, \sigma_n^2(\boldsymbol{\beta}_{\boldsymbol{k}}))$, $\sigma_i^2(\boldsymbol{\beta}_{\boldsymbol{k}}) = \mu_i(\boldsymbol{\beta}_{\boldsymbol{k}})(1-\mu_i(\boldsymbol{\beta}_{\boldsymbol{k}}))$ and

$$\mu_i(\boldsymbol{\beta}_{\boldsymbol{k}}) \;=\; \frac{\exp\left(\boldsymbol{x}_{i\boldsymbol{k}}^{\top}\boldsymbol{\beta}_{\boldsymbol{k}}\right)}{1+\exp\left(\boldsymbol{x}_{i\boldsymbol{k}}^{\top}\boldsymbol{\beta}_{\boldsymbol{k}}\right)}.$$

In the rest of the paper, we denote $\boldsymbol{\Sigma} = \boldsymbol{\Sigma}(\boldsymbol{\beta}_{\boldsymbol{t}})$ and $\sigma_i^2 = \sigma_i^2(\boldsymbol{\beta}_{\boldsymbol{t}})$ for simplicity.

Before we establish our main results, the following notations are needed for stating our assumptions. For any $a, b \in \mathbb{R}$, $a \vee b$ and $a \wedge b$ mean the maximum and minimum of a and b, respectively. For any sequences a_n and b_n, we denote $a_n \lesssim b_n$, or equivalently $a_n = O(b_n)$, if there exists a constant $C > 0$ such that $|a_n| \leq C|b_n|$ for all large n. We denote $a_n \ll b_n$, or equivalently $a_n = o(b_n)$, if $a_n/b_n \longrightarrow 0$ as $n \to \infty$. Without loss of generality, if $a_n \geq b_n > 0$ and there exist constants $C_1 > C_2 > 0$ such that $C_2 < b_n/a_n \leq a_n/b_n < C_1$, we denote $a_n \sim b_n$. For a given vector $\boldsymbol{v} = (v_1, \ldots, v_p)^\top \in \mathbb{R}^p$, the vector ℓ_2-norm is denoted as $\|\boldsymbol{v}\|_2 = (\sum_{j=1}^p v_j^2)^{1/2}$. For any real symmetric matrix $\boldsymbol{A}$, $\lambda_{\max}(\boldsymbol{A})$ and $\lambda_{\min}(\boldsymbol{A})$ are the maximum and minimum eigenvalue of $\boldsymbol{A}$, respectively. To attain desirable asymptotic properties of our posterior, we assume the following conditions:

Condition (A1) $\log n \lesssim \log p = o(n^{1/2})$ and $m_n = O\Big((n/\log p)^{\frac{1-d'}{2}} \wedge \log p\Big)$ for some $0 \leq d < (1+d)/2 \leq d' \leq 1$.

Condition (A1) ensures our proposed method can accommodate high dimensions where the number of predictors grows at a sub-exponential rate of n. Condition (A1) also specifies the parameter m_n in the uniform prior (3) that restricts our analysis on a set of *reasonably large* models. Similar assumptions restricting the model size have been commonly assumed in the sparse estimation literature [4,5,9,17].

Condition (A2) For some constant $C \in (0, \infty)$ and $0 \leq d < (1+d)/2 \leq d' \leq 1$,

$$\begin{aligned} \max_{i,j} |x_{ij}| &\leq C, \\ 0 < \lambda \leq \min_{\boldsymbol{k}:|\boldsymbol{k}| \leq m_n + |\boldsymbol{t}|} \lambda_{\min}\Big(n^{-1} H_n(\boldsymbol{\beta}_{0,\boldsymbol{k}})\Big) &\leq \Lambda_{m_n+|\boldsymbol{t}|} \leq C^2 (\log p)^d, \end{aligned}$$

and $\Lambda_\zeta = \max_{\boldsymbol{k}:|\boldsymbol{k}| \leq \zeta} \lambda_{\max}(n^{-1} \boldsymbol{X}_{\boldsymbol{k}}^\top \boldsymbol{X}_{\boldsymbol{k}})$ for any integer $\zeta > 0$. Furthermore, $\|\boldsymbol{\beta}_{0,\boldsymbol{t}}\|_2^2 = O\big((\log p)^d\big)$.

Condition (A2) gives lower and upper bounds of $n^{-1}\boldsymbol{H}_n(\boldsymbol{\beta}_{0,\boldsymbol{k}})$ and $n^{-1}\boldsymbol{X}_{\boldsymbol{k}}^\top \boldsymbol{X}_{\boldsymbol{k}}$, respectively, where $\boldsymbol{k}$ is a large model satisfying $|\boldsymbol{k}| \leq m_n + |\boldsymbol{t}|$. The lower bound condition can be regarded as a restricted eigenvalue condition for ℓ_0-sparse vectors. Restricted eigenvalue conditions are routinely assumed in high-dimensional theory to guarantee some level of curvature of the objective function and are satisfied with high probability for sub-Gaussian design matrices [5]. Similar conditions have also been used in the linear regression literature [18–20]. The last assumption in Condition (A2) says that the magnitude of true signals is bounded above $(\log p)^d$ up to some constant, which allows the magnitude of signals to increase to infinity.

Condition (A3) For some constant $c_0 > 0$,

$$\min_{j \in \boldsymbol{t}} \beta_{0,j}^2 \geq c_0 \Big(\frac{|\boldsymbol{t}| \Lambda_{|\boldsymbol{t}|} \log p}{n} \vee \frac{1}{\log p} \Big). \tag{9}$$

Condition (A3) gives a lower bound for nonzero signals, which is called the *beta-min* condition. In general, this type of condition is necessary for catching every nonzero signal. Please note that due to Conditions (A1) and (A2), the right-hand side of (9) decreases to zero as $n \to \infty$. Thus, it allows the smallest nonzero coefficients to tend to zero as we observe more data.

Condition (A4) For some small constant $\delta > 0$, the hyperparameters τ and r satisfy

$$\tau^{r+1/2} \sim n^{-1/2} p^{2+\delta}.$$

Condition (A4) suggests appropriate conditions for the hyperparameter τ in (2). A similar assumption has also been considered in [9]. The scale parameter τ in the nonlocal prior density reflects the dispersion of the nonlocal prior density around zero, and implicitly determines the size of the regression coefficients that will be shrunk to zero [8,9]. For the below theoretical results, we assume

that $\boldsymbol{U} = \boldsymbol{I}$ for simplicity, but our results are still valid for other choices of $\boldsymbol{U}$ as long as $\lambda_{\max}(\boldsymbol{U}) = O(1)$ and $\lambda_{\min}(\boldsymbol{U}) = O(1)$.

Theorem 1 (No super set). *Under conditions (A1), (A2) and (A4),*

$$\pi(\boldsymbol{k} \supsetneq \boldsymbol{t} \mid \boldsymbol{y}) \xrightarrow{P} 0, \text{ as } n \to \infty.$$

Theorem 1 says that, asymptotically, our posterior will not overfit the model, i.e., not include unnecessarily many variables. Of course, it does not guarantee that the posterior will concentrate on the true model. To capture every significant variable, we require the magnitudes of nonzero entries in $\boldsymbol{\beta}_{0,t}$ not to be too small. Theorem 2 shows that with an appropriate lower bound specified in Condition (A3), the true model $\boldsymbol{t}$ will be the mode of the posterior.

Theorem 2 (Posterior ratio consistency). *Under conditions (A1)–(A4) with* $c_0 = \{(1-\epsilon_0)\lambda\}^{-1}\{2(3+\delta) + 5\{(1-\epsilon_0)\lambda\}^{-1}\}$ *for some small constant* $\epsilon_0 > 0$,

$$\max_{\boldsymbol{k} \neq \boldsymbol{t}} \frac{\pi(\boldsymbol{k} \mid \boldsymbol{y})}{\pi(\boldsymbol{t} \mid \boldsymbol{y})} \xrightarrow{P} 0, \text{ as } n \to \infty.$$

Posterior ratio consistency is a useful property especially when we are interested in the point estimation with the posterior mode, but does not provide how large probability the posterior puts on the true model. In the following theorem, we state that our posterior achieves *strong selection consistency*. By strong selection consistency, we mean that the posterior probability assigned to the true model t converges to 1. Since strong selection consistency implies posterior ratio consistency, it requires a slightly stronger condition on the lower bound for the magnitudes of nonzero entries in $\boldsymbol{\beta}_{0,t}$, i.e., a larger value of c_0, compared to that in Theorem 2.

Theorem 3 (Strong selection consistency). *Under conditions (A1)–(A4) with* $c_0 = \{(1-\epsilon_0)\lambda\}^{-1}\{2(9+2\delta) + 5\{(1-\epsilon_0)\lambda\}^{-1}\}$ *for some small constant* $\epsilon_0 > 0$, *the following holds:*

$$\pi(\boldsymbol{t} \mid \boldsymbol{y}) \xrightarrow{P} 1, \text{ as } n \to \infty.$$

4. Computational Strategy

In this section, we describe how to approximate the marginal density of the data and to conduct the model selection procedure. The integral formulation in (4) leads to the posterior probabilities not available in closed form. Hence, we use Laplace approximation to compute $m_{\boldsymbol{k}}(\boldsymbol{y})$ and $\pi(\boldsymbol{k}|\boldsymbol{y})$. A similar approach to compute posterior probabilities has been used in [8–10].

Please note that for any model $\boldsymbol{k}$, when $\boldsymbol{U}_{\boldsymbol{k}} = \boldsymbol{I}_{\boldsymbol{k}}$, the normalization constant $d_{\boldsymbol{k}}$ in (2) is given by $d_{\boldsymbol{k}} = \{(2r-1)!!\}^{-|\boldsymbol{k}|}$. Let

$$\begin{aligned} f(\boldsymbol{\beta}_{\boldsymbol{k}}) &= \log\left(\exp\{L_n(\boldsymbol{\beta}_{\boldsymbol{k}})\}\pi(\boldsymbol{\beta}_{\boldsymbol{k}} \mid \boldsymbol{k})\right) \\ &= \textstyle\sum_{i=1}^n \left\{y_i \boldsymbol{x}_{i\boldsymbol{k}}^\top \boldsymbol{\beta}_{\boldsymbol{k}} - \log\left(1 + \exp(\boldsymbol{x}_{i\boldsymbol{k}}^\top \boldsymbol{\beta}_{\boldsymbol{k}})\right)\right\} - |\boldsymbol{k}| \log\left((2r-1)!!\right) - \frac{|\boldsymbol{k}|}{2}\log(2\pi) - \left(r|\boldsymbol{k}| + \frac{|\boldsymbol{k}|}{2}\right)\log\tau \\ &\quad - \frac{\boldsymbol{\beta}_{\boldsymbol{k}}^\top \boldsymbol{\beta}_{\boldsymbol{k}}}{2\tau} + \textstyle\sum_{i=1}^{|\boldsymbol{k}|} 2r \log\left(|\beta_{k_i}|\right). \end{aligned}$$

For any model $\boldsymbol{k}$, the Laplace approximation of $m_{\boldsymbol{k}}(\boldsymbol{y})$ is given by

$$(2\pi)^{\frac{|\boldsymbol{k}|}{2}} \exp\{f(\hat{\boldsymbol{\beta}}_{\boldsymbol{k}})\} |\boldsymbol{V}(\hat{\boldsymbol{\beta}}_{\boldsymbol{k}})|^{-\frac{1}{2}}, \tag{10}$$

where $\hat{\boldsymbol{\beta}}_{\boldsymbol{k}} = arg\,max_{\boldsymbol{\beta}_{\boldsymbol{k}}} f(\boldsymbol{\beta}_{\boldsymbol{k}})$ obtained via the optimization function optim in R using a quasi-Newton method and $\boldsymbol{V}(\hat{\boldsymbol{\beta}}_{\boldsymbol{k}})$ is a $|\boldsymbol{k}| \times |\boldsymbol{k}|$ symmetric matrix which can be calculated as:

$$-\sum_{i=1}^{n} \frac{\boldsymbol{x}_{i\boldsymbol{k}} \boldsymbol{x}_{i\boldsymbol{k}}^{\top} \exp(\boldsymbol{x}_{i\boldsymbol{k}}^{\top} \boldsymbol{\beta}_{\boldsymbol{k}})}{\{1 + \exp(\boldsymbol{x}_{i\boldsymbol{k}}^{\top} \boldsymbol{\beta}_{\boldsymbol{k}})\}^2} - \frac{1}{\tau} \boldsymbol{I}_{\boldsymbol{k}} - diag\left(\frac{2r}{\beta_{k_1}^2}, \ldots, \frac{2r}{\beta_{k_{|\boldsymbol{k}|}}^2}\right).$$

The above Laplace approximation can be used to compute the log of the posterior probability ratio between any given model $\boldsymbol{k}$ and true model $\boldsymbol{t}$, and select a model $\boldsymbol{k}$ with the highest probability.

We then adopt the shotgun stochastic search (SSS) algorithm [9,12] to efficiently navigate through the massive model space and identify the global maxima. Using the Laplace approximations of the marginal probabilities in (10), the SSS method aims at exploring "interesting" regions of the resulting high-dimensional model spaces and quickly identifies regions of high posterior probability over models. Let $\text{nbd}(\boldsymbol{k}) = \{\Gamma^{+}, \Gamma^{-}, \Gamma^{0}\}$ containing all the neighbors of model $\boldsymbol{k}$, in which $\Gamma^{+} = \{\boldsymbol{k} \cup \{j\} : j \notin \boldsymbol{k}\}$, $\Gamma^{-} = \{\boldsymbol{k} \setminus \{j\} : j \in \boldsymbol{k}\}$ and $\Gamma^{0} = \{\boldsymbol{k} \setminus \{j\} \cup \{l\} : j \in \boldsymbol{k}, l \notin \boldsymbol{k}\}$. The SSS procedure is described in Algorithm 1.

Algorithm 1 Shotgun Stochastic Search (SSS)

Choose an initial model $\boldsymbol{k}^{(1)}$
for $i = 1$ to $i = N - 1$ **do**
  Compute $\pi(\boldsymbol{k}|\boldsymbol{y})$ for all $\boldsymbol{k} \in \text{nbd}(\boldsymbol{k}^{(i)})$
  Sample $\boldsymbol{k}^{+}$, $\boldsymbol{k}^{-}$ and $\boldsymbol{k}^{0}$, from Γ^{+}, Γ^{-} and Γ^{0} with probabilities proportional to $\pi(\boldsymbol{k}|\boldsymbol{y})$
  Sample the next model $\boldsymbol{k}^{(i+1)}$ from $\{\boldsymbol{k}^{+}, \boldsymbol{k}^{-}, \boldsymbol{k}^{0}\}$ with probability proportional to $\{\pi(\boldsymbol{k}^{+}|\boldsymbol{y}), \pi(\boldsymbol{k}^{-}|\boldsymbol{y}), \pi(\boldsymbol{k}^{0}|\boldsymbol{y})\}$
end for

5. Simulation Studies

In this section, we demonstrate the performance of the proposed method in various settings. Let $\boldsymbol{X}$ be the design matrix whose first $|\boldsymbol{t}|$ columns correspond to the active covariates for which we have nonzero coefficients, while the rest correspond to the inactive ones with zero coefficients. In all the simulation settings, we generate $\boldsymbol{x}_i \overset{i.i.d.}{\sim} N_p(0, \boldsymbol{\Sigma})$ for $i = 1, \ldots, n$ under the following two different cases of $\boldsymbol{\Sigma}$:

- Case 1: Isotropic design, where $\boldsymbol{\Sigma} = \boldsymbol{I}_p$, i.e., no correlation imposed between different covariates.
- Case 2: Autoregressive correlated design, where $\Sigma_{ij} = 0.3^{|i-j|}$, for all $1 \leq i \leq j \leq p$. The correlations among different covariates are set to different values.

Following the simulation settings in [9,10], we consider the following two designs, each with the same sample size $n = 100$ and number of predictors being either $p = 100$ or 150:

- Design 1 (Dense model): The number of predictors $p = 100$ and $|\boldsymbol{t}| = 8$.
- Design 2 (High-dimensional): The number of predictors $p = 150$ and $|\boldsymbol{t}| = 4$.

We investigate the following two settings for the true coefficient vector $\boldsymbol{\beta}_{0,\boldsymbol{t}}$ to include different combinations of small and large signals.

- Setting 1: All the entries of $\boldsymbol{\beta}_{0,\boldsymbol{t}}$ are set to 3.
- Setting 2: All the entries of $\boldsymbol{\beta}_{0,\boldsymbol{t}}$ are generated from Unif$(1.5, 3)$.

Finally, for given $\boldsymbol{X}$ and $1 \leq i \leq n$, we sample y_i from the following logistic model as in (1)

$$\mathbb{P}(y_i = 1 \mid \boldsymbol{x}_i, \boldsymbol{\beta}_0) = \frac{\exp(\boldsymbol{x}_i^{\top} \boldsymbol{\beta}_0)}{1 + \exp(\boldsymbol{x}_i^{\top} \boldsymbol{\beta}_0)}.$$

We will refer to our proposed method as "nonlocal" and its performance will then be compared with other existing methods including Spike and Slab prior-based model selection [21], empirical Bayesian LASSO (EBLasso) [22], Lasso [23] and SCAD [24]. The tuning parameters in the regularization approaches are chosen by 5-fold cross-validation. Spike and slab prior method is implemented via the BoomSpikeSlab package in R. For the nonlocal prior, the hyperparameters are set at $\boldsymbol{U} = \boldsymbol{I}$, $r = 1$ and we tune $\tau = 10^{-i} n^{-1/2} p^{2+0.05}$ for four different values of $i = 0, 1, 2, 3$. We choose the optimal τ by the mean squared prediction error through 5-fold cross-validation. Please note that this implies that τ is data-dependent and the resulting procedure is similar to an empirical-Bayesian approach in the high-dimensional Bayesian literature given the prior knowledge about the sparse true model [13]. For the SSS procedure, the initial model was set by randomly taking three coefficients to be active and the remaining to be inactive. The detailed steps for our method are coded in R and publicly available at https://github.com/xuan-cao/Nonlocal-Logistic-Selection. In particular, the stochastic search is implemented via the SSS function in the R package BayesS5.

To evaluate the performance of variable selection, the precision, sensitivity, specificity, Matthews correlation coefficient (MCC) [25] and mean-squared prediction error (MSPE) are reported at Tables 1–4, where each simulation setting is repeated for 20 times. The criteria are defined as

$$\text{Precision} = \frac{TP}{TP+FP}, \quad \text{Sensitivitiy} = \frac{TP}{TP+FN}, \quad \text{Specificity} = \frac{TN}{TN+FP},$$
$$\text{MCC} = \frac{TP \times TN - FP \times FN}{\sqrt{(TP+FP)(TP+FN)(TN+FP)(TN+FN)}}, \text{MSPE} = \frac{1}{n_{\text{test}}} \sum_{i=1}^{n_{\text{test}}} \left(\hat{y}_i - y_{\text{test},i}\right)^2,$$

where TP, TN, FP and FN are true positive, true negative, false positive and false negative, respectively. Here we denote $\hat{y}_i = \boldsymbol{x}_i^\top \hat{\boldsymbol{\beta}}$, where $\hat{\boldsymbol{\beta}}$ is the estimated coefficient based on each method. For Bayesian methods, the usual GLM estimates based on the selected support are used as $\hat{\boldsymbol{\beta}}$. We generated test samples $y_{\text{test},1}, \ldots, y_{\text{test},n_{\text{test}}}$ with $n_{\text{test}} = 50$ to calculate the MSPE.

Table 1. The summary statistics for Design 1 (Dense model design) are represented for each setting of the true regression coefficients under the first isotropic covariance case. Different setting means different choice of the true coefficient β_0.

	Setting 1				
	Precision	**Sensitivity**	**Specificity**	**MCC**	**MSPE**
Nonlocal	1	1	1	1	0.02
Spike and Slab	1	0.38	1	0.60	0.21
Lasso	0.67	1	0.96	0.80	0.17
EBLasso	1	0.38	1	0.60	0.22
SCAD	0.57	1	0.93	0.73	0.14
	Setting 2				
	Precision	**Sensitivity**	**Specificity**	**MCC**	**MSPE**
Nonlocal	0.73	1	0.97	0.84	0.18
Spike and Slab	1	0.13	1	0.34	0.23
Lasso	0.54	0.88	0.93	0.65	0.15
EBLasso	1	0.63	1	0.78	0.22
SCAD	0.47	0.88	0.91	0.60	0.13

Table 2. The summary statistics for Design 1 (Dense model design) are represented for each setting of the true regression coefficients under the second autoregressive covariance case. Different setting means different choice of the true coefficient β_0.

	Setting 1				
	Precision	**Sensitivity**	**Specificity**	**MCC**	**MSPE**
Nonlocal	0.89	1	0.99	0.94	0.13
Spike and Slab	0.71	0.63	0.98	0.64	0.20
Lasso	0.70	0.88	0.98	0.76	0.16
EBLasso	1	0.50	1	0.69	0.23
SCAD	0.67	0.75	0.97	0.68	0.17
	Setting 2				
	Precision	**Sensitivity**	**Specificity**	**MCC**	**MSPE**
Nonlocal	0.88	0.88	0.99	0.86	0.14
Spike and Slab	0.83	0.63	0.99	0.70	0.13
Lasso	0.63	0.88	0.96	0.72	0.14
EBLasso	1	0.38	1	0.60	0.22
SCAD	0.47	0.88	0.91	0.60	0.13

Table 3. The summary statistics for Design 2 (High-dimensional design) are represented for each setting of the true regression coefficients under the first isotropic covariance case. Different setting means different choice of the true coefficient β_0.

	Setting 1				
	Precision	**Sensitivity**	**Specificity**	**MCC**	**MSPE**
Nonlocal	1	1	1	1	0.08
Spike and Slab	0.75	0.75	0.99	0.74	0.09
Lasso	0.80	1	0.99	0.89	0.14
EBLasso	1	0.75	1	0.86	0.21
SCAD	0.67	1	0.99	0.81	0.12
	Setting 2				
	Precision	**Sensitivity**	**Specificity**	**MCC**	**MSPE**
Nonlocal	1	1	1	1	0.10
Spike and Slab	0.75	0.75	0.99	0.74	0.11
Lasso	0.67	1	0.99	0.81	0.14
EBLasso	1	0.75	1	0.86	0.23
SCAD	0.44	1	0.97	0.66	0.12

Table 4. The summary statistics for Design 2 (High-dimensional design) are represented for each setting of the true regression coefficients under the second autoregressive covariance case. Different setting means different choice of the true coefficient β_0.

	Setting 1				
	Precision	**Sensitivity**	**Specificity**	**MCC**	**MSPE**
Nonlocal	1	0.75	1	0.86	0.11
Spike and Slab	1	0.50	1	0.71	0.10
Lasso	0.57	1	0.98	0.75	0.10
EBLasso	1	0.50	1	0.70	0.18
SCAD	0.44	1	0.97	0.66	0.12
	Setting 2				
	Precision	**Sensitivity**	**Specificity**	**MCC**	**MSPE**
Nonlocal	1	0.75	1	0.86	0.15
Spike and Slab	0.50	0.50	0.99	0.49	0.14
Lasso	0.44	1	0.97	0.66	0.13
EBLasso	1	0.50	1	0.70	0.21
SCAD	0.40	1	0.96	0.62	0.14

Based on the above simulation results, we notice that under the first isotropic covariance case, the nonlocal-based approach overall works better than other methods especially in the strong signal setting (i.e., Setting 1), where our method is able to consistently achieve perfect estimation accuracy. This is because as signal strength gets stronger, the consistency conditions of our method are easier to satisfy which leads to better performance. When the covariance is autoregressive, our method suffers from lower sensitivity compared with the frequentist approaches in high-dimensional design (Table 4), but still has higher precision, specificity and MCC. The poor precision of the regularization methods has also been discussed in previous literature in the sense that selection of the regularization parameter using cross-validation is optimal with respect to prediction but tends to include too many noise predictors [26]. Again we observe under the autoregressive design, the performance of our method improves as the true signals strengthen. To sum up, the above simulation studies indicate that the proposed method can perform well under a variety of configurations with different data generation mechanisms.

6. Application to fMRI Data Analysis

In this section, we apply the proposed model selection method to an fMRI data set for identifying aberrant functional brain activities to aid the diagnosis of Parkinson's Disease (PD) [27]. Data consists of 70 PD patients and 50 healthy controls (HC). All the demographic characteristics and clinical symptom ratings have been collected before MRI scanning. In particular, we adopt the mini-mental state examination (MMSE) for cognitive evaluation and the Hamilton Depression Scale (HAMD) for measuring the severity of depression.

6.1. Image Feature Extraction

Functional imaging data for all subjects are collected and retrieved from the archive by neuroradiologists. Image preprocessing procedure is carried out via Statistical Parametric Mapping (SPM12) operated on the Matlab platform. For each subject, we first discard the first 5 time points for signal equilibrium and the remaining 135 images underwent slice-timing and head motion corrections. Four subjects with more than 2.5 mm maximum displacement in any of the three dimensions or 2.5° of any angular motion are removed. The functional images are spatially normalized to the Montreal Neurological Institute space with $3 \times 3 \times 3$ mm^3 cubic voxels and smoothed with a 4 mm full width at half maximum (FWHM) Gaussian kernel. We further regress out nuisance covariates and applied temporal filter ($0.01 \text{ Hz} < f < 0.08 \text{ Hz}$) to diminish high-frequency noise.

Zang et al. [28] proposed the method of Regional Homogeneity (ReHo) to analyze characteristics of regional brain activity and to reflect the temporal homogeneity of neural activity. Since some preprocessing methods especially spatial smoothing fMRI time series may significantly change the ReHo magnitudes [29], preprocessed fMRI data without the spatial smoothing step are used for calculating ReHo. In particular, we focus on the mReHo maps obtained by dividing the mean ReHo of the whole brain within each voxel in the ReHo map. We further segment the mReHo maps and extract all the 112 ROI signals based on the Harvard-Oxford atlas (HOA) using the Resting-State fMRI Data Analysis Toolkit.

Slow fluctuations in activity are fundamental features of the resting brain for determining correlated activity between brain regions and resting state networks. The relative magnitude of these fluctuations can discriminate between brain regions and subjects. Amplitude of Low Frequency Fluctuations (ALFF) [30] are related measures that quantify the amplitude of these low frequency oscillations. Leveraging the preprocessed data, we retain the standardized mALFF maps after dividing the ALFF of each voxel by the global mean ALFF. Using the HOA, we again obtain 112 mALFF values via extracting the ROI signals based on the mALFF maps. Voxel-Mirrored Homotopic Connectivity (VMHC) quantifies functional homotopy by providing a voxel-wise measure of connectivity between hemispheres [31]. By segmenting the VMHC maps according to HOA, we also extract 112 VHMC values.

6.2. Results

Our candidate features consist of 336 radiomic variables along with all the clinical characteristics. We now consider a standard logistic regression model with the binary disease indicator as the outcome and all the radiomic variables together with five clinical factors as predictors. Various models including the proposed and other competing methods will then be implemented for classifying subjects based on these extracted features. The dataset is randomly divided into a training set (80%) and a testing set (20%) while maintaining the PD:HC ratio in both sets. For Bayesian methods, we first obtain the identified variables, and then evaluate the testing set performance using standard GLM estimates based on the selected features. The penalty parameters in all frequentist methods are tuned via 5-fold cross validation in the training set. The hyperparameters for the proposed method are set as in simulation studies.

The HAMD score and nine radiomic features including five mALFFs, two ReHos, two VHMCs are selected by the SSS procedure under pMOM prior. In Figure 1, we plot the histograms of selected radiomic features with different colors representing different groups. The predictive performance of various methods in the test set is summarized in Table 5. We can tell from Table 5 that the nonlocal prior-based approach has overall better prediction performance compared with other methods. Our nonlocal approach has higher precision and specificity compared with all the other methods, but yields a lower sensitivity than the frequentist approaches. Based on the most comprehensive measure MCC, our method outperforms all the other methods.

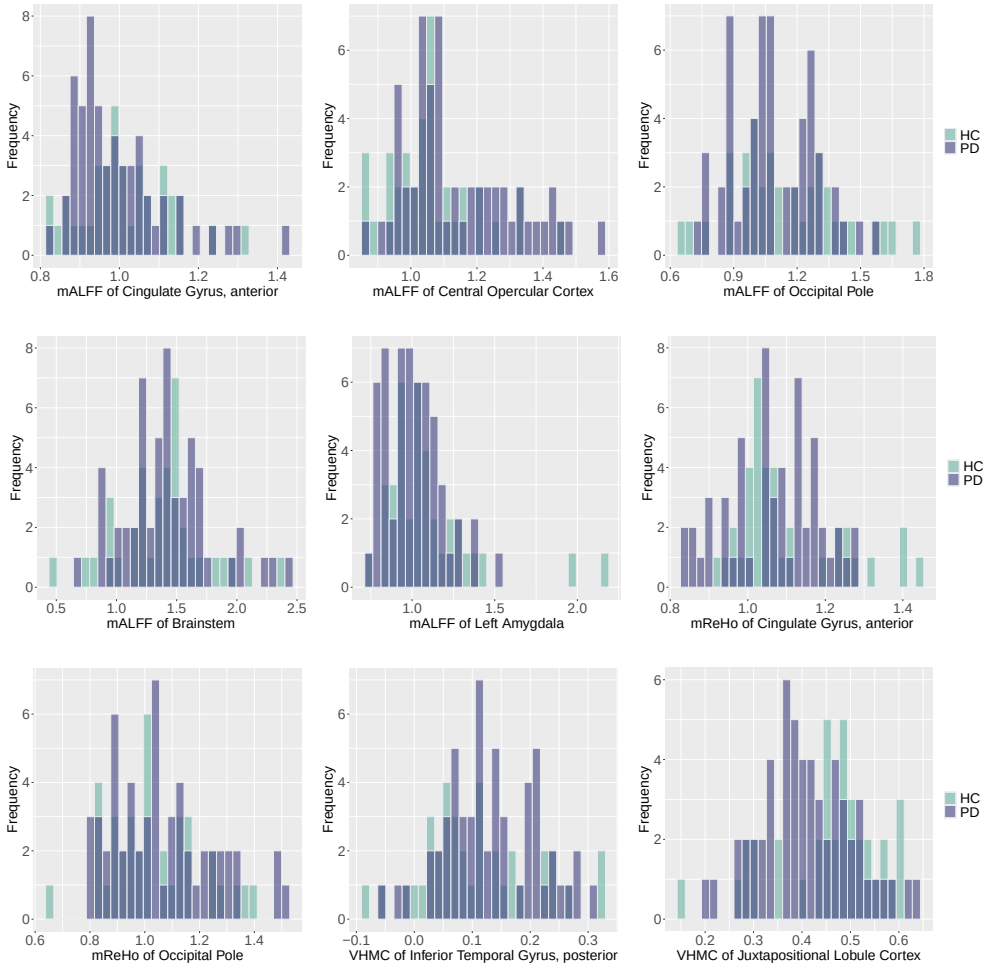

Figure 1. Histograms of selected radiomic features for PD and HC subjects with darker color representing overlapping values. Purple: PD group; Green: HC group.

Table 5. The summary statistics for prediction performance on the testing set for all methods.

	Precision	Sensitivity	Specificity	MCC	MSPE
Nonlocal	0.77	0.83	0.73	0.56	0.21
Spike and Slab	0.53	0.75	0.27	0.40	0.29
Lasso	0.67	1	0.45	0.55	0.18
EBLasso	0.57	1	0.18	0.32	0.28
SCAD	0.58	1	0.37	0.41	0.19

7. Conclusions

In this paper, we propose a Bayesian hierarchical model with a pMOM prior specification over regression coefficients to perform variable selection in high-dimensional generalized linear models. The model selection consistency of our method is established under mild conditions and the shotgun stochastic search algorithm can be used for the implementation of our proposed approach. Our simulation and real data studies indicate that the proposed method has better performance for variable selection compared to a variety of state-of-the-art competing methods. In the fMRI data analysis, our method is able to identify abnormal functional brain activities for PD that occur in the regions of interest including cingulate gyrus, central opercular cortex, occipital pole, brainstem, left amygdala, occipital pole, inferior temporal gyrus, and juxtapositional lobule cortex. These findings suggest disease-related alterations of functional activities that provide physicians sufficient information to get involved with early diagnosis and treatment. Our findings are also coherent with the alternative functional features in cortical regions, brainstem, and limbic regions discovered in previous studies [32–35].

Our fMRI study certainly has limitations. First, we would like to note that fMRI data are typically treated as spatio-temporal objects and a generalized linear model with spatially varying coefficients can be implemented for brain decoding [36]. However, in our application, for each subject, a total of 135 fMRI scans were obtained, each with the dimension of $64 \times 64 \times 31$. If we take each voxel as a covariate to perform the whole-brain functional analysis, it would be computationally challenging and impractical given the extremely high dimension. Hence, we adopt the radiomics approach to extract three different types of features that can summarize the functional activity of the brain, and take these radiomic features as covariates in our generalized linear model. For future studies, we will focus on several regions of interest rather than the entire brain and take the spatio-temporal dependency among voxels into consideration.

Second, although ReHo, ALFF, and VHMC are different types of radiomic features that quantify the functional activity of the brain, it is definitely possible that in some regions, three measures are highly correlated with each other. Our current theoretical and computational strategy can accommodate a reasonable amount of correlations among covariates, but might not work in the presence of high correlation structure. For future studies, we will first carefully examine the potential correlations among features and might only retain one feature for each region if significant correlations are detected.

One possible extension of our methodology is to address the potential misspecification of the hyperparameter τ. The scale parameter τ is of particular importance in the sense that it can reflect the dispersion of the nonlocal density around zero, and implicitly determine the size of the regression coefficients that will be shrunk to zero [8]. Cao et al. [14] investigated the model selection consistency for the hyper-pMOM priors in linear regression setting, where an additional inverse-gamma prior is placed over τ. Wu et al. [11] proved the model selection consistency using hyper-pMOM prior in generalized linear models, but assumed a fixed dimension p. For future study, we will consider this fully Bayesian approach to carefully examine the theoretical and empirical properties for such hyper-pMOM prior in the context of high-dimensional generalized linear regression. We can also extend our method to develop a Bayesian approach for growth models in the context of non-linear regression [37], where the log-transformation is typically used to recover the additive structure.

However, then the model does not fall into the category of GLMs, which is beyond the current setting in this paper. Therefore, we leave it as a future work.

Author Contributions: Conceptualization, X.C.; Methodology, X.C. and K.L.; Software, X.C.; Supervision, K.L.; Validation, K.L.; Writing—original draft, X.C.; Writing—review & editing, K.L. Both authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by Simons Foundation: No.635213; University of Cincinnati: Taft Summer Research Fellowship; National Research Foundation of Korea: No.2019R1F1A1059483; INHA UNIVERSITY Research Grant.

Acknowledgments: We would like to thank two referees for their valuable comments which have led to improvements of an earlier version of the paper. This research was supported by Simons Foundation's collaboration grant (No.635213), Taft Summer Research Fellowship at University of Cincinnati and the National Research Foundation of Korea (NRF) grant funded by the Korea government (MSIT) (No.2019R1F1A1059483).

Conflicts of Interest: The authors declare no conflict of interest.

Abbreviations

The following abbreviations are used in this manuscript:

Lasso	least absolute shrinkage and selection operator
SCAD	smoothly clipped absolute deviation
MCP	minimum concave penalty
pMOM	product moment
SSS	shotgun stochastic search
MCMC	Markov chain Monte Carlo
fMRI	functional Magnetic Resonance Imaging
GLM	generalized linear model
EBLasso	empirical Bayesian LASSO
MCC	Matthews correlation coefficient
MSPE	mean-squared prediction error
PD	Parkinson's Disease
HC	healthy controls
MMSE	mini-mental state examination
HAMD	Hamilton Depression Scale
SPM12	Statistical Parametric Mapping
FWHM	full width at half maximum
HOA	Harvard-Oxford atlas
ALFF	Amplitude of Low Frequency Fluctuations
VMHC	Voxel-Mirrored Homotopic Connectivity

Appendix A

Throughout the Supplementary Material, we assume that for any

$$\boldsymbol{u} \in \{\boldsymbol{u} \in \mathbb{R}^n : \boldsymbol{u} \text{ is in the space spanned by the columns of } \boldsymbol{\Sigma}^{1/2}\boldsymbol{X}_{\boldsymbol{k}}\}$$

and any model $\boldsymbol{k} \in \{\boldsymbol{k} \subseteq [r] : |\boldsymbol{k}| \leq m_n + |\boldsymbol{t}|\}$, there exists $\delta^* > 0$ such that

$$\mathbb{E}\Big[\exp\left\{\boldsymbol{u}^\top \boldsymbol{\Sigma}^{-1/2}(\boldsymbol{y}-\boldsymbol{\mu})\right\}\Big] \quad \leq \quad \exp\Big\{\frac{(1+\delta^*)\boldsymbol{u}^\top \boldsymbol{u}}{2}\Big\}, \tag{A1}$$

for any $n \geq N(\delta^*)$. However, as stated in [5], there always exists $\delta^* > 0$ satisfying inequality (A1), so it is not really a restriction. Since we will focus on sufficiently large n, δ^* can be considered an arbitrarily small constant, so we can always assume that $\delta > \delta^*$.

Proof of Theorem 1. Let $M_1 = \{\boldsymbol{k} : \boldsymbol{k} \supsetneq \boldsymbol{t}, |\boldsymbol{k}| \leq m_n\}$ and

$$PR(\boldsymbol{k}, \boldsymbol{t}) \quad = \quad \frac{\pi(\boldsymbol{k} \mid \boldsymbol{y})}{\pi(\boldsymbol{t} \mid \boldsymbol{y})},$$

where $\boldsymbol{t} \subseteq [r]$ is the true model. We will show that

$$\sum_{\boldsymbol{k}:\boldsymbol{k} \in M_1} PR(\boldsymbol{k}, \boldsymbol{t}) \quad \xrightarrow{P} \quad 0 \quad \text{as } n \to \infty. \tag{A2}$$

By Taylor's expansion of $L_n(\boldsymbol{\beta}_{\boldsymbol{k}})$ around $\widehat{\boldsymbol{\beta}}_{\boldsymbol{k}}$, which is the MLE of $\boldsymbol{\beta}_{\boldsymbol{k}}$ under the model $\boldsymbol{k}$, we have

$$L_n(\boldsymbol{\beta}_{\boldsymbol{k}}) \quad = \quad L_n(\widehat{\boldsymbol{\beta}}_{\boldsymbol{k}}) - \frac{1}{2}(\boldsymbol{\beta}_{\boldsymbol{k}} - \widehat{\boldsymbol{\beta}}_{\boldsymbol{k}})^\top \boldsymbol{H}_n(\tilde{\boldsymbol{\beta}}_{\boldsymbol{k}})(\boldsymbol{\beta}_{\boldsymbol{k}} - \widehat{\boldsymbol{\beta}}_{\boldsymbol{k}})$$

for some $\tilde{\boldsymbol{\beta}}_{\boldsymbol{k}}$ such that $\|\tilde{\boldsymbol{\beta}}_{\boldsymbol{k}} - \widehat{\boldsymbol{\beta}}_{\boldsymbol{k}}\|_2 \leq \|\boldsymbol{\beta}_{\boldsymbol{k}} - \widehat{\boldsymbol{\beta}}_{\boldsymbol{k}}\|_2$. Furthermore, by Lemmas A.1 and A.3 in [5] and Condition (A2), with probability tending to 1,

$$L_n(\boldsymbol{\beta}_{\boldsymbol{k}}) - L_n(\widehat{\boldsymbol{\beta}}_{\boldsymbol{k}}) \quad \leq \quad -\frac{1-\epsilon}{2}(\boldsymbol{\beta}_{\boldsymbol{k}} - \widehat{\boldsymbol{\beta}}_{\boldsymbol{k}})^\top \boldsymbol{H}_n(\boldsymbol{\beta}_{0,\boldsymbol{k}})(\boldsymbol{\beta}_{\boldsymbol{k}} - \widehat{\boldsymbol{\beta}}_{\boldsymbol{k}})$$

for any $\boldsymbol{k} \in M_1$ and $\boldsymbol{\beta}_{\boldsymbol{k}}$ such that $\|\boldsymbol{\beta}_{\boldsymbol{k}} - \boldsymbol{\beta}_{0,\boldsymbol{k}}\|_2 < c\sqrt{|\boldsymbol{k}|\Lambda_{|\boldsymbol{k}|}\log p/n} =: cw_n$, where $\epsilon = \epsilon_n := c'\sqrt{m_n^2 \Lambda_{m_n} \log p/n} = o(1)$, for some constants $c, c' > 0$. Please note that for $\boldsymbol{\beta}_{\boldsymbol{k}}$ such that $\|\boldsymbol{\beta}_{\boldsymbol{k}} - \widehat{\boldsymbol{\beta}}_{\boldsymbol{k}}\|_2 = cw_n/2$,

$$\begin{aligned} L_n(\boldsymbol{\beta}_{\boldsymbol{k}}) - L_n(\widehat{\boldsymbol{\beta}}_{\boldsymbol{k}}) \quad &\leq \quad -\frac{1-\epsilon}{2}\|\boldsymbol{\beta}_{\boldsymbol{k}} - \widehat{\boldsymbol{\beta}}_{\boldsymbol{k}}\|_2^2 \lambda_{\min}\{\boldsymbol{H}_n(\boldsymbol{\beta}_{0,\boldsymbol{k}})\} \\ &\leq \quad -\frac{1-\epsilon}{2}\frac{c^2 w_n^2}{4} n\lambda = -\frac{1-\epsilon}{8}c^2\lambda|\boldsymbol{k}|\Lambda_{|\boldsymbol{k}|}\log p \quad \longrightarrow \quad -\infty \quad \text{as } n \to \infty, \end{aligned}$$

where the second inequality holds due to Condition (A2). It also holds for any $\boldsymbol{\beta}_{\boldsymbol{k}}$ such that $\|\boldsymbol{\beta}_{\boldsymbol{k}} - \widehat{\boldsymbol{\beta}}_{\boldsymbol{k}}\|_2 > cw_n/2$ by concavity of $L_n(\cdot)$ and the fact that $\widehat{\boldsymbol{\beta}}_{\boldsymbol{k}}$ maximizes $L_n(\boldsymbol{\beta}_{\boldsymbol{k}})$.

Define the set $B := \{\boldsymbol{\beta}_{\boldsymbol{k}} : \|\boldsymbol{\beta}_{\boldsymbol{k}} - \widehat{\boldsymbol{\beta}}_{\boldsymbol{k}}\|_2 \leq cw_n/2\}$, then we have $B \subset \{\boldsymbol{\beta}_{\boldsymbol{k}} : \|\boldsymbol{\beta}_{\boldsymbol{k}} - \boldsymbol{\beta}_{0,\boldsymbol{k}}\|_2 \leq cw_n\}$ for some large $c > 0$ and any $\boldsymbol{k} \in M_1$, with probability tending to 1.

$$\begin{aligned} m_{\boldsymbol{k}}(\boldsymbol{y}) \quad &= \quad \int \exp\{L_n(\boldsymbol{\beta}_{\boldsymbol{k}})\}\pi(\boldsymbol{\beta}_{\boldsymbol{k}} \mid \boldsymbol{k})\, d\boldsymbol{\beta}_{\boldsymbol{k}} \\ &= \quad \int \exp\{L_n(\boldsymbol{\beta}_{\boldsymbol{k}})\} d_{\boldsymbol{k}}(2\pi)^{-|\boldsymbol{k}|/2}(\tau)^{-r|\boldsymbol{k}|-|\boldsymbol{k}|/2}|\boldsymbol{U}_{\boldsymbol{k}}|^{\frac{1}{2}} \exp\Big(-\frac{\boldsymbol{\beta}_{\boldsymbol{k}}^\top \boldsymbol{U}_{\boldsymbol{k}}\boldsymbol{\beta}_{\boldsymbol{k}}}{2\tau}\Big)\prod_{i=1}^{|\boldsymbol{k}|}\beta_{k_i}^{2r} d\boldsymbol{\beta}_{\boldsymbol{k}} \\ &\leq \quad d_{\boldsymbol{k}}(2\pi)^{-|\boldsymbol{k}|/2}(\tau)^{-r|\boldsymbol{k}|-|\boldsymbol{k}|/2}|\boldsymbol{U}_{\boldsymbol{k}}|^{\frac{1}{2}} \exp\{L_n(\widehat{\boldsymbol{\beta}}_{\boldsymbol{k}})\} \\ &\quad \times\Bigg[\int_B \exp\Big\{-\frac{1-\epsilon}{2}(\boldsymbol{\beta}_{\boldsymbol{k}} - \widehat{\boldsymbol{\beta}}_{\boldsymbol{k}})^\top \boldsymbol{H}_n(\boldsymbol{\beta}_{0,\boldsymbol{k}})(\boldsymbol{\beta}_{\boldsymbol{k}} - \widehat{\boldsymbol{\beta}}_{\boldsymbol{k}}) - \frac{\boldsymbol{\beta}_{\boldsymbol{k}}^\top \boldsymbol{U}_{\boldsymbol{k}}\boldsymbol{\beta}_{\boldsymbol{k}}}{2\tau}\Big\}\prod_{i=1}^{|\boldsymbol{k}|}\beta_{k_i}^{2r} d\boldsymbol{\beta}_{\boldsymbol{k}} \\ &\qquad + \exp\Big(-\frac{1-\epsilon}{8}c^2\lambda|\boldsymbol{k}|\Lambda_{|\boldsymbol{k}|}\log p\Big)\int_{B^c}\exp\Big(-\frac{\boldsymbol{\beta}_{\boldsymbol{k}}^\top \boldsymbol{U}_{\boldsymbol{k}}\boldsymbol{\beta}_{\boldsymbol{k}}}{2\tau}\Big)\prod_{i=1}^{|\boldsymbol{k}|}\beta_{k_i}^{2r} d\boldsymbol{\beta}_{\boldsymbol{k}}\Bigg] \end{aligned} \tag{A3}$$

Please note that for $\boldsymbol{A}_{\boldsymbol{k}} = (1-\epsilon)\boldsymbol{H}_n(\boldsymbol{\beta}_{0,\boldsymbol{k}})$ and $\boldsymbol{\beta}^*_{\boldsymbol{k}} = (\boldsymbol{A}_{\boldsymbol{k}} + \boldsymbol{U}_{\boldsymbol{k}}/\tau)^{-1}\boldsymbol{A}_{\boldsymbol{k}}\widehat{\boldsymbol{\beta}}_{\boldsymbol{k}}$, we have

$$\begin{aligned}
&\int_B \exp\Big\{-\frac{1-\epsilon}{2}(\boldsymbol{\beta}_{\boldsymbol{k}}-\widehat{\boldsymbol{\beta}}_{\boldsymbol{k}})^\top \boldsymbol{H}_n(\boldsymbol{\beta}_{0,\boldsymbol{k}})(\boldsymbol{\beta}_{\boldsymbol{k}}-\widehat{\boldsymbol{\beta}}_{\boldsymbol{k}}) - \frac{\boldsymbol{\beta}_{\boldsymbol{k}}^\top \boldsymbol{U}_{\boldsymbol{k}}\boldsymbol{\beta}_{\boldsymbol{k}}}{2\tau}\Big\}\prod_{i=1}^{|\boldsymbol{k}|}\beta_{k_i}^{2r}d\boldsymbol{\beta}_{\boldsymbol{k}} \\
\leq\ & \int \exp\Big\{-\frac{1}{2}(\boldsymbol{\beta}_{\boldsymbol{k}}-\boldsymbol{\beta}^*_{\boldsymbol{k}})^\top(\boldsymbol{A}_{\boldsymbol{k}}+\boldsymbol{U}_{\boldsymbol{k}}/\tau)(\boldsymbol{\beta}_{\boldsymbol{k}}-\boldsymbol{\beta}^*_{\boldsymbol{k}})\Big\}\prod_{i=1}^{|\boldsymbol{k}|}\beta_{k_i}^{2r}d\boldsymbol{\beta}_{\boldsymbol{k}} \\
&\times\ \exp\Big\{-\frac{1}{2}\widehat{\boldsymbol{\beta}}_{\boldsymbol{k}}^\top(\boldsymbol{A}_{\boldsymbol{k}}-\boldsymbol{A}_{\boldsymbol{k}}(\boldsymbol{A}_{\boldsymbol{k}}+\boldsymbol{U}_{\boldsymbol{k}}/\tau)^{-1}\boldsymbol{A}_{\boldsymbol{k}})\widehat{\boldsymbol{\beta}}_{\boldsymbol{k}}\Big\} \\
=\ & (2\pi)^{|\boldsymbol{k}|/2}\det(\boldsymbol{A}_{\boldsymbol{k}}+\boldsymbol{U}_{\boldsymbol{k}}/\tau)^{-1/2}\exp\Big\{-\frac{1}{2}\widehat{\boldsymbol{\beta}}_{\boldsymbol{k}}^\top(\boldsymbol{A}_{\boldsymbol{k}}-\boldsymbol{A}_{\boldsymbol{k}}(\boldsymbol{A}_{\boldsymbol{k}}+\boldsymbol{U}_{\boldsymbol{k}}/\tau)^{-1}\boldsymbol{A}_{\boldsymbol{k}})\widehat{\boldsymbol{\beta}}_{\boldsymbol{k}}\Big\}E_{\boldsymbol{k}}\Big(\prod_{i=1}^{|\boldsymbol{k}|}\beta_{k_i}^{2r}\Big),
\end{aligned}$$

where $E_{\boldsymbol{k}}(.)$ denotes the expectation with respect to a multivariate normal distribution with mean $\boldsymbol{\beta}^*_{\boldsymbol{k}}$ and covariance matrix $(\boldsymbol{A}_{\boldsymbol{k}}+\boldsymbol{U}_{\boldsymbol{k}}/\tau)^{-1}$. It follows from Lemma 6 in the supplementary material for [8] that

$$\begin{aligned}
E_{\boldsymbol{k}}\Big(\prod_{i=1}^{|\boldsymbol{k}|}\beta_{k_i}^{2r}\Big) &\leq \left(\frac{n\Lambda_{|\boldsymbol{k}|}+\tau^{-1}}{n\lambda+\tau^{-1}}\right)^{|\boldsymbol{k}|/2}\left\{\frac{4V}{|\boldsymbol{k}|}+\frac{4\left[(2r-1)!!\right]^{\frac{1}{r}}}{n(\lambda+\tau^{-2})}\right\}^{r|\boldsymbol{k}|} \\
&\leq \left(\frac{n\Lambda_{|\boldsymbol{k}|}+\tau^{-1}}{n\lambda+\tau^{-1}}\right)^{|\boldsymbol{k}|/2}2^{r|\boldsymbol{k}|-1}\left\{\left(\frac{4V}{|\boldsymbol{k}|}\right)^{r|\boldsymbol{k}|}+\left(\frac{4\left[(2r-1)!!\right]^{\frac{1}{r}}}{n(\lambda+\tau^{-1})}\right)^{r|\boldsymbol{k}|}\right\},
\end{aligned}$$

where $V = \|\boldsymbol{\beta}^*_{\boldsymbol{k}}\|_2^2$, and

$$\int_{B^c}\exp\Big\{-\frac{\boldsymbol{\beta}_{\boldsymbol{k}}^\top\boldsymbol{U}_{\boldsymbol{k}}\boldsymbol{\beta}_{\boldsymbol{k}}}{2\tau}\Big\}\prod_{i=1}^{|\boldsymbol{k}|}\beta_{k_i}^{2r}d\boldsymbol{\beta}_{\boldsymbol{k}} \leq (C\tau)^{|\boldsymbol{k}|/2},$$

for some constant $C > 0$. Further note that

$$\begin{aligned}
\det\Big\{\boldsymbol{H}_n(\boldsymbol{\beta}_{0,\boldsymbol{k}})(1-\epsilon)+\tau^{-1}I\Big\}^{1/2} &\leq \left(n\Lambda_{|\boldsymbol{k}|}+\tau^{-1}\right)^{|\boldsymbol{k}|/2} \\
&\leq \exp\left\{C|\boldsymbol{k}|\log n\right\} \\
&\ll \exp\Big\{\frac{1-\epsilon}{8}c^2\lambda|\boldsymbol{k}|\Lambda_{|\boldsymbol{k}|}\log p\Big\}
\end{aligned}$$

for some constant $C > 0$ and some large constant $c > 0$, by Conditions (A1), (A2) and (A4). Therefore, it follows from (A3) that

$$\begin{aligned}
m_{\boldsymbol{k}}(\boldsymbol{y}) &\leq C^{-|\boldsymbol{k}|/2}(\tau)^{-r|\boldsymbol{k}|-|\boldsymbol{k}|/2}\exp\left\{L_n(\widehat{\boldsymbol{\beta}}_{\boldsymbol{k}})\right\}\det\Big\{\boldsymbol{H}_n(\boldsymbol{\beta}_{0,\boldsymbol{k}})(1-\epsilon)+\tau^{-1}\boldsymbol{U}_{\boldsymbol{k}}\Big\}^{-1/2} \\
&\quad\times\left[\Lambda_{|\boldsymbol{k}|}{}^{|\boldsymbol{k}|/2}\left\{\left(\frac{V}{|\boldsymbol{k}|}\right)^{r|\boldsymbol{k}|}+n^{-r|\boldsymbol{k}|}\right\}+o(1)\right] \qquad \text{(A4)} \\
&\leq C^{-|\boldsymbol{k}|/2}(\tau)^{-r|\boldsymbol{k}|-|\boldsymbol{k}|/2}\exp\left\{L_n(\widehat{\boldsymbol{\beta}}_{\boldsymbol{k}})\right\}\det\Big\{\boldsymbol{H}_n(\boldsymbol{\beta}_{0,\boldsymbol{k}})(1-\epsilon)+\tau^{-1}\boldsymbol{U}_{\boldsymbol{k}}\Big\}^{-1/2} \\
&\quad\times\Lambda_{|\boldsymbol{k}|}{}^{|\boldsymbol{k}|/2}\left\{\left(\frac{V}{|\boldsymbol{k}|}\right)^{r|\boldsymbol{k}|}+n^{-r|\boldsymbol{k}|}\right\},
\end{aligned}$$

for some constant $C > 0$. Next, note that it follows from Lemma A.3 in the supplementary material for [5] that

$$\begin{aligned} V = \|\boldsymbol{\beta}_{\boldsymbol{k}}^*\|_2^2 \le \|\widehat{\boldsymbol{\beta}}_{\boldsymbol{k}}\|_2^2 &\le \left(\|\widehat{\boldsymbol{\beta}}_{\boldsymbol{k}} - \boldsymbol{\beta}_{0,\boldsymbol{k}}\|_2 + \|\boldsymbol{\beta}_{0,\boldsymbol{k}}\|_2\right)^2 \\ &\le \left(\sqrt{\frac{|\boldsymbol{k}|\Lambda_{|\boldsymbol{k}|}\log p}{n}} + \sqrt{\log p}\right)^2 \\ &\le 2\left(\frac{|\boldsymbol{k}|\Lambda_{|\boldsymbol{k}|}\log p}{n} + \log p\right). \end{aligned}$$

Therefore,

$$\begin{aligned} \left(\frac{V}{|\boldsymbol{k}|}\right)^{r|\boldsymbol{k}|} &\le \left(\frac{2\log p}{|\boldsymbol{k}|}\right)^{r|\boldsymbol{k}|} \exp\left(r|\boldsymbol{k}|^2\frac{\Lambda_{|\boldsymbol{k}|}}{n}\right) \\ &\lesssim \left(\frac{2\log p}{|\boldsymbol{k}|}\right)^{r|\boldsymbol{k}|} \end{aligned}$$

by Conditions (A1) and (A2). Combining with (A4), we obtain the following upper bound for $m_{\boldsymbol{k}}(\boldsymbol{y})$,

$$\begin{aligned} m_{\boldsymbol{k}}(\boldsymbol{y}) \le\ & (C_1\tau)^{-r|\boldsymbol{k}|-|\boldsymbol{k}|/2}\exp\left\{L_n(\widehat{\boldsymbol{\beta}}_{\boldsymbol{k}})\right\}\det\left\{\boldsymbol{H}_n(\boldsymbol{\beta}_{0,\boldsymbol{k}})(1-\epsilon) + \tau^{-1}\boldsymbol{U}_{\boldsymbol{k}}\right\}^{-1/2} \\ & \times\Lambda_{|\boldsymbol{k}|}^{|\boldsymbol{k}|/2}\left(\frac{\log p}{|\boldsymbol{k}|}\right)^{r|\boldsymbol{k}|}, \end{aligned} \tag{A5}$$

for any $\boldsymbol{k} \in M_1$ and some constant $C_1 > 0$. Similarly, by Lemma 4 in the supplementary material for [8] and the similar arguments leading up to (A5), with probability tending to 1, we have

$$\begin{aligned} m_{\boldsymbol{t}}(\boldsymbol{y}) &\gtrsim (C_1\tau)^{-r|\boldsymbol{t}|-|\boldsymbol{t}|/2}\exp\left\{L_n(\widehat{\boldsymbol{\beta}}_{\boldsymbol{t}})\right\}\det\left\{\boldsymbol{H}_n(\boldsymbol{\beta}_{0,\boldsymbol{t}})(1+\epsilon) + \tau^{-1}\boldsymbol{U}_{\boldsymbol{t}}\right\}^{-1/2} \\ &\quad\times\ \exp\left\{-\frac{1}{2}\widehat{\boldsymbol{\beta}}_{\boldsymbol{t}}^\top\left(\boldsymbol{A}_{\boldsymbol{t}} - \boldsymbol{A}_{\boldsymbol{t}}(\boldsymbol{A}_{\boldsymbol{t}} + \tau^{-2}\boldsymbol{I}_{\boldsymbol{t}})^{-1}\boldsymbol{A}_{\boldsymbol{t}}\right)\widehat{\boldsymbol{\beta}}_{\boldsymbol{t}}\right\}(\log p)^{-r|\boldsymbol{t}|} \\ &\gtrsim (C_1\tau)^{-r|\boldsymbol{t}|-|\boldsymbol{t}|/2}\exp\left\{L_n(\widehat{\boldsymbol{\beta}}_{\boldsymbol{t}})\right\}\det\left\{\boldsymbol{H}_n(\boldsymbol{\beta}_{0,\boldsymbol{t}})(1+\epsilon) + \tau^{-1}\boldsymbol{U}_{\boldsymbol{t}}\right\}^{-1/2}(\log p)^{-r|\boldsymbol{t}|} \end{aligned}$$

by Lemma A1, where $\boldsymbol{A}_{\boldsymbol{t}} = (1+\epsilon)\boldsymbol{H}_n(\boldsymbol{\beta}_{0,\boldsymbol{t}})$. Therefore, with probability tending to 1,

$$\begin{aligned} \frac{m_{\boldsymbol{k}}(\boldsymbol{y})}{m_{\boldsymbol{t}}(\boldsymbol{y})} &\lesssim \left\{C_1 n^{1/2}\tau^{r+1/2}\right\}^{-(|\boldsymbol{k}|-|\boldsymbol{t}|)}\frac{\det\left\{n^{-1}\boldsymbol{H}_n(\boldsymbol{\beta}_{0,\boldsymbol{t}})(1+\epsilon) + (n\tau)^{-1}\boldsymbol{U}_{\boldsymbol{t}}\right\}^{1/2}}{\det\left\{n^{-1}\boldsymbol{H}_n(\boldsymbol{\beta}_{0,\boldsymbol{k}})(1-\epsilon) + (n\tau)^{-1}\boldsymbol{U}_{\boldsymbol{k}}\right\}^{1/2}} \\ &\quad\times\ \exp\left\{L_n(\widehat{\boldsymbol{\beta}}_{\boldsymbol{k}}) - L_n(\widehat{\boldsymbol{\beta}}_{\boldsymbol{t}})\right\}\Lambda_{|\boldsymbol{k}|}^{\frac{|\boldsymbol{k}|}{2}}\left(\frac{\log p}{|\boldsymbol{k}|}\right)^{r|\boldsymbol{k}|}(\log p)^{r|\boldsymbol{t}|} \\ &\lesssim \left\{C_1 p^{2+\delta}\right\}^{-(|\boldsymbol{k}|-|\boldsymbol{t}|)}\left(\frac{2}{\lambda}\right)^{|\boldsymbol{k}|-|\boldsymbol{t}|}\exp\left\{L_n(\widehat{\boldsymbol{\beta}}_{\boldsymbol{k}}) - L_n(\widehat{\boldsymbol{\beta}}_{\boldsymbol{t}})\right\}(\log p)^{r(2|\boldsymbol{k}|+|\boldsymbol{t}|)} \\ &\lesssim (C_1 p)^{-(2+\delta)(|\boldsymbol{k}|-|\boldsymbol{t}|)}\left(\frac{2}{\lambda}\right)^{|\boldsymbol{k}|-|\boldsymbol{t}|}p^{(1+\delta^*)(1+2w)(|\boldsymbol{k}|-|\boldsymbol{t}|)}(\log p)^{r(2|\boldsymbol{k}|+|\boldsymbol{t}|)} \end{aligned} \tag{A6}$$

for any $k \in M_1$, where the second inequality holds by Lemma 2 in [38], Conditions (A2) and (A4), and the third inequality follows from Lemma 3 in [38], which implies

$$L_n(\widehat{\boldsymbol{\beta}}_{\boldsymbol{k}}) - L_n(\widehat{\boldsymbol{\beta}}_{\boldsymbol{t}}) \le b_n(|\boldsymbol{k}| - |\boldsymbol{t}|) \tag{A7}$$

for any $\boldsymbol{k} \in M_1$ with probability tending to 1, where $b_n = (1+\delta^*)(1+2w)\log p$ with some small constant $w > 0$ satisfying $1 + \delta > (1+\delta^*)(1+2w)$.

Hence, with probability tending to 1, it follows from (A6) that

$$\sum_{\boldsymbol{k}\in M_1} PR(\boldsymbol{k},\boldsymbol{t}) = \sum_{\boldsymbol{k}\supsetneq\boldsymbol{t}} \frac{\pi(\boldsymbol{k})m_{\boldsymbol{k}}(\boldsymbol{y})}{\pi(\boldsymbol{t})m_{\boldsymbol{t}}(\boldsymbol{y})} \le \sum_{\boldsymbol{k}\supsetneq\boldsymbol{t}} \frac{m_{\boldsymbol{k}}(\boldsymbol{y})}{m_{\boldsymbol{t}}(\boldsymbol{y})} \le \sum_{|\boldsymbol{k}|-|\boldsymbol{t}|=1}^{m_n-|\boldsymbol{t}|} \binom{p-|\boldsymbol{t}|}{|\boldsymbol{k}|-|\boldsymbol{t}|} p^{-(1+c)(|\boldsymbol{k}|-|\boldsymbol{t}|)}. \tag{A8}$$

for some constant $c > 0$. Using $\binom{p-|\boldsymbol{t}|}{|\boldsymbol{k}|-|\boldsymbol{t}|} \le p^{|\boldsymbol{k}|-|\boldsymbol{t}|}$ and (A8), we get

$$\sum_{\boldsymbol{k}\in M_1} PR(\boldsymbol{k},\boldsymbol{t}) \;=\; o(1).$$

Thus, we have proved the desired result (A2). □

Proof of Theorem 2. Let $M_2 = \{\boldsymbol{k} : \boldsymbol{k} \not\supseteq \boldsymbol{t}, |\boldsymbol{k}| \le m_n\}$. For any $\boldsymbol{k} \in M_2$, let $\boldsymbol{k}^* = \boldsymbol{k} \cup \boldsymbol{t}$, so that $\boldsymbol{k}^* \in M_1$. Let $\boldsymbol{\beta}_{\boldsymbol{k}^*}$ be the $|\boldsymbol{k}^*|$-dimensional vector including $\boldsymbol{\beta}_{\boldsymbol{k}}$ for $\boldsymbol{k}$ and zeros for $\boldsymbol{t} \setminus \boldsymbol{k}$. Then by Taylor's expansion and Lemmas A.1 and A.3 in [5], with probability tending to 1,

$$\begin{aligned} L_n(\boldsymbol{\beta}_{\boldsymbol{k}^*}) &= L_n(\widehat{\boldsymbol{\beta}}_{\boldsymbol{k}^*}) - \frac{1}{2}(\boldsymbol{\beta}_{\boldsymbol{k}^*} - \widehat{\boldsymbol{\beta}}_{\boldsymbol{k}^*})^\top \boldsymbol{H}_n(\tilde{\boldsymbol{\beta}}_{\boldsymbol{k}^*})(\boldsymbol{\beta}_{\boldsymbol{k}^*} - \widehat{\boldsymbol{\beta}}_{\boldsymbol{k}^*}) \\ &\le L_n(\widehat{\boldsymbol{\beta}}_{\boldsymbol{k}^*}) - \frac{1-\epsilon}{2}(\boldsymbol{\beta}_{\boldsymbol{k}^*} - \widehat{\boldsymbol{\beta}}_{\boldsymbol{k}^*})^\top \boldsymbol{H}_n(\boldsymbol{\beta}_{0,\boldsymbol{k}^*})(\boldsymbol{\beta}_{\boldsymbol{k}^*} - \widehat{\boldsymbol{\beta}}_{\boldsymbol{k}^*}) \\ &\le L_n(\widehat{\boldsymbol{\beta}}_{\boldsymbol{k}^*}) - \frac{n(1-\epsilon)\lambda}{2}\|\boldsymbol{\beta}_{\boldsymbol{k}^*} - \widehat{\boldsymbol{\beta}}_{\boldsymbol{k}^*}\|_2^2 \end{aligned}$$

for any $\boldsymbol{\beta}_{\boldsymbol{k}^*}$ such that $\|\boldsymbol{\beta}_{\boldsymbol{k}^*} - \boldsymbol{\beta}_{0,\boldsymbol{k}^*}\|_2 \le c\sqrt{|\boldsymbol{k}^*|\Lambda_{|\boldsymbol{k}^*|}\log p/n} = cw_n$ for some large constant $c > 0$. Please note that

Let $\boldsymbol{B}_{\boldsymbol{k}} = n(1-\epsilon)\lambda \boldsymbol{I}_{\boldsymbol{k}}$ and $\boldsymbol{\beta}_{\boldsymbol{k}}^* = (\boldsymbol{B}_{\boldsymbol{k}} + \boldsymbol{U}_{\boldsymbol{k}}/\tau)^{-1}\boldsymbol{B}_{\boldsymbol{k}}\widehat{\boldsymbol{\beta}}_{\boldsymbol{k}}$,

$$\begin{aligned} &\int \exp\Big\{-\frac{n(1-\epsilon)\lambda}{2}\|\boldsymbol{\beta}_{\boldsymbol{k}^*} - \widehat{\boldsymbol{\beta}}_{\boldsymbol{k}^*}\|_2^2\Big\} \exp\Big(-\frac{\boldsymbol{\beta}_{\boldsymbol{k}}^\top \boldsymbol{U}_{\boldsymbol{k}}\boldsymbol{\beta}_{\boldsymbol{k}}}{2\tau}\Big) \prod_{i=1}^{|\boldsymbol{k}|}\beta_{k_i}^{2r} d\boldsymbol{\beta}_{\boldsymbol{k}} \\ = &\int \exp\Big\{-\frac{n(1-\epsilon)\lambda}{2}\|\boldsymbol{\beta}_{\boldsymbol{k}} - \widehat{\boldsymbol{\beta}}_{\boldsymbol{k}}\|_2^2 - \frac{1}{2\tau}\boldsymbol{\beta}_{\boldsymbol{k}}^\top \boldsymbol{U}_{\boldsymbol{k}}\boldsymbol{\beta}_{\boldsymbol{k}}\Big\} \prod_{i=1}^{|\boldsymbol{k}|}\beta_{k_i}^{2r} d\boldsymbol{\beta}_{\boldsymbol{k}} \exp\Big\{-\frac{n(1-\epsilon)\lambda}{2}\|\widehat{\boldsymbol{\beta}}_{\boldsymbol{t}\setminus\boldsymbol{k}}\|_2^2\Big\} \\ = &(2\pi)^{\frac{|\boldsymbol{k}|}{2}}|\boldsymbol{B}_{\boldsymbol{k}} + \boldsymbol{U}_{\boldsymbol{k}}/\tau|^{-1/2} \exp\Big\{-\frac{1}{2}\widehat{\boldsymbol{\beta}}_{\boldsymbol{k}}^\top(\boldsymbol{B}_{\boldsymbol{k}} - \boldsymbol{B}_{\boldsymbol{k}}(\boldsymbol{B}_{\boldsymbol{k}} + \boldsymbol{U}_{\boldsymbol{k}}/\tau)^{-1}\boldsymbol{B}_{\boldsymbol{k}})\widehat{\boldsymbol{\beta}}_{\boldsymbol{k}}\Big\} E_{\boldsymbol{k}}\Big(\prod_{i=1}^{|\boldsymbol{k}|}\beta_{k_i}^{2r}\Big) \\ &\times \exp\Big\{-\frac{n(1-\epsilon)\lambda}{2}\|\widehat{\boldsymbol{\beta}}_{\boldsymbol{t}\setminus\boldsymbol{k}}\|_2^2\Big\}. \end{aligned}$$

where $E_{\boldsymbol{k}}(.)$ denotes the expectation with respect to a multivariate normal distribution with mean $\boldsymbol{\beta}_{\boldsymbol{k}}^*$ and covariance matrix $(\boldsymbol{B}_{\boldsymbol{k}} + \boldsymbol{U}_{\boldsymbol{k}}/\tau)^{-1}$. It follows from Lemma 6 in the supplementary material for [8] that

$$\begin{aligned} E_{\boldsymbol{k}}\Big(\prod_{i=1}^{|\boldsymbol{k}|}\beta_{k_i}^{2r}\Big) &\le \Big(\frac{n\lambda + \tau^{-1}}{n\lambda + \tau^{-1}}\Big)^{|\boldsymbol{k}|/2}\left\{\frac{4V}{|\boldsymbol{k}|} + \frac{4\left[(2r-1)!!\right]^{\frac{1}{r}}}{n(\lambda + \tau^{-1})}\right\}^{r|\boldsymbol{k}|} \\ &\le \Big(\frac{n\lambda + \tau^{-1}}{n\lambda + \tau^{-1}}\Big)^{|\boldsymbol{k}|/2} 2^{r|\boldsymbol{k}|-1}\left\{\Big(\frac{4V}{|\boldsymbol{k}|}\Big)^{r|\boldsymbol{k}|} + \left(\frac{4\left[(2r-1)!!\right]^{\frac{1}{r}}}{n(\lambda + \tau^{-1})}\right)^{r|\boldsymbol{k}|}\right\}, \end{aligned}$$

where $V = \|\boldsymbol{\beta}_{\boldsymbol{k}}^*\|_2^2$. Define the set $B_* := \{\boldsymbol{\beta}_{\boldsymbol{k}} : \|\boldsymbol{\beta}_{\boldsymbol{k}^*} - \widehat{\boldsymbol{\beta}}_{\boldsymbol{k}^*}\|_2 \le cw_n/2\}$, for some large constant $c > 0$, then by similar arguments used for super sets, with probability tending to 1,

$$\begin{aligned}
\pi(\boldsymbol{k} \mid \boldsymbol{y}) &= d_{\boldsymbol{k}}(2\pi)^{-|\boldsymbol{k}|/2}(\tau)^{-r|\boldsymbol{k}|-|\boldsymbol{k}|/2}|\boldsymbol{U}_{\boldsymbol{k}}|^{\frac{1}{2}} \int_{B_* \cup B_*^c} \exp\{L_n(\beta_{G_{k^*}})\} \exp\Big(-\frac{\boldsymbol{\beta}_{\boldsymbol{k}}^\top \boldsymbol{U}_{\boldsymbol{k}} \boldsymbol{\beta}_{\boldsymbol{k}}}{2\tau}\Big) \prod_{i=1}^{|\boldsymbol{k}|} \beta_{k_i}^{2r} d\boldsymbol{\beta}_{\boldsymbol{k}} \\
&\lesssim (C_1\tau)^{-r|\boldsymbol{k}|-|\boldsymbol{k}|/2} \exp\{L_n(\widehat{\boldsymbol{\beta}}_{\boldsymbol{k}^*})\} \det\left(\boldsymbol{B}_{\boldsymbol{k}} + \boldsymbol{U}_{\boldsymbol{k}}/\tau\right)^{-1/2} \\
&\quad \times \Big[\exp\Big\{-\frac{n(1-\epsilon)\lambda}{2}\|\widehat{\boldsymbol{\beta}}_{\boldsymbol{t}\setminus\boldsymbol{k}}\|_2^2\Big\}\Big(\frac{\log p}{|\boldsymbol{k}|}\Big)^{r|\boldsymbol{k}|} + \exp\{-cC|\boldsymbol{k}^*|\Lambda_{|\boldsymbol{k}^*|}\log p\}\Big]
\end{aligned}$$

for any $\boldsymbol{k} \in M_2$ and for some constant $C > 0$.

Since the lower bound for $\pi(\boldsymbol{t} \mid \boldsymbol{y})$ can be derived as before, it leads to

$$\begin{aligned}
PR(\boldsymbol{k}, \boldsymbol{t}) &\lesssim \{C_1 n^{1/2}\tau^{r+1/2}\}^{-(|\boldsymbol{k}|-|\boldsymbol{t}|)} \frac{\det\{(1+\epsilon)n^{-1}\boldsymbol{H}_n(\boldsymbol{\beta}_{0,\boldsymbol{t}}) + (n\tau)^{-1}\boldsymbol{U}_{\boldsymbol{t}}\}^{1/2}}{\det\{(1-\epsilon)\lambda\boldsymbol{I}_{\boldsymbol{k}} + (n\tau)^{-1}\boldsymbol{U}_{\boldsymbol{k}}\}^{1/2}} \\
&\quad \times \exp\{L_n(\widehat{\boldsymbol{\beta}}_{\boldsymbol{k}^*}) - L_n(\widehat{\boldsymbol{\beta}}_{\boldsymbol{t}})\} \exp\Big\{-\frac{n(1-\epsilon)\lambda}{2}\|\widehat{\boldsymbol{\beta}}_{\boldsymbol{t}\setminus\boldsymbol{k}}\|_2^2\Big\}\Big(\frac{\log p}{|\boldsymbol{k}|}\Big)^{r|\boldsymbol{k}|}(\log p)^{r|\boldsymbol{t}|} \quad \text{(A9)} \\
&+ \{C_1 n^{1/2}\tau^{r+1/2}\}^{-(|\boldsymbol{k}|-|\boldsymbol{t}|)} \det\{(1+\epsilon)n^{-1}\boldsymbol{H}_n(\boldsymbol{\beta}_{0,\boldsymbol{t}}) + (n\tau)^{-1}\boldsymbol{U}_{\boldsymbol{t}}\}^{1/2} \\
&\quad \times \exp\{L_n(\widehat{\boldsymbol{\beta}}_{\boldsymbol{k}^*}) - L_n(\widehat{\boldsymbol{\beta}}_{\boldsymbol{t}})\} \exp\{-c\,C|\boldsymbol{k}^*|\Lambda_{|\boldsymbol{k}^*|}\log p\}(\log p)^{r|\boldsymbol{t}|} \quad \text{(A10)}
\end{aligned}$$

for any $\boldsymbol{k} \in M_2$ with probability tending to 1.

We first focus on (A9). Please note that

$$\begin{aligned}
&\frac{\det\{(1+\epsilon)n^{-1}\boldsymbol{H}_n(\boldsymbol{\beta}_{0,\boldsymbol{t}}) + (n\tau)^{-1}\boldsymbol{U}_{\boldsymbol{t}}\}^{1/2}}{\det\{(1-\epsilon)\lambda\boldsymbol{I}_{\boldsymbol{k}} + (n\tau)^{-1}\boldsymbol{U}_{\boldsymbol{k}}\}^{1/2}} \\
\le\ & \frac{\{(1+\epsilon)\Lambda_{|\boldsymbol{t}|} + (n\tau)^{-1}\}^{|\boldsymbol{t}|/2}}{\{(1-\epsilon)\lambda + (n\tau)^{-1}\}^{|\boldsymbol{k}|/2}} \\
=\ & \Big\{\frac{(1+\epsilon)\Lambda_{|\boldsymbol{t}|} + (n\tau)^{-1}}{(1-\epsilon)\lambda + (n\tau)^{-1}}\Big\}^{|\boldsymbol{t}|/2} \Big\{\frac{1}{(1-\epsilon)\lambda + (n\tau)^{-1}}\Big\}^{(|\boldsymbol{k}|-|\boldsymbol{t}|)/2} \\
\lesssim\ & \exp\{C|\boldsymbol{t}|\log\Lambda_{|\boldsymbol{t}|}\}\Big\{\frac{1}{(1-\epsilon)\lambda + (n\tau)^{-1}}\Big\}^{(|\boldsymbol{k}|-|\boldsymbol{t}|)/2}
\end{aligned}$$

for some constant $C > 0$. Furthermore, by the same arguments used in (A7), we have

$$\begin{aligned}
L_n(\widehat{\boldsymbol{\beta}}_{\boldsymbol{k}^*}) - L_n(\widehat{\boldsymbol{\beta}}_{\boldsymbol{t}}) &\lesssim C_*(|\boldsymbol{k}^*| - |\boldsymbol{t}|)\log p \\
&= C_*|\boldsymbol{t}\setminus\boldsymbol{k}|\log p + C_*(|\boldsymbol{k}| - |\boldsymbol{t}|)\log p
\end{aligned}$$

for some constant $C_* > 0$ and for any $\boldsymbol{k} \in M_2$ with probability tending to 1. Here we choose $C_* = (1+\delta^*)(1+2w)$ if $|\boldsymbol{k}| > |\boldsymbol{t}|$ or $C_* = 3+\delta$ if $|\boldsymbol{k}| \le |\boldsymbol{t}|$ so that

$$\begin{aligned}
&\{C_1 n^{1/2}\tau^{r+1/2}\}^{-(|\boldsymbol{k}|-|\boldsymbol{t}|)}\Big\{\frac{1}{(1-\epsilon)\lambda + (n\tau)^{-1}}\Big\}^{(|\boldsymbol{k}|-|\boldsymbol{t}|)/2} p^{C_*(|\boldsymbol{k}|-|\boldsymbol{t}|)} \\
&\times \exp\Big\{r|\boldsymbol{k}|\log\Big(\frac{\log p}{|\boldsymbol{k}|}\Big) + r|\boldsymbol{t}|\log(\log p)\Big\} \\
\lesssim\ & p^{(C_*-2-\delta)(|\boldsymbol{k}|-|\boldsymbol{t}|)} = o(1),
\end{aligned}$$

where the inequality holds by Condition (A4). To be more specific, we divide M_2 into two disjoint sets $M_2' = \{\boldsymbol{k} : \boldsymbol{k} \in M_2, |\boldsymbol{t}| < |\boldsymbol{k}| \le m_n\}$ and $M_2^* = \{\boldsymbol{k} : \boldsymbol{k} \in M_2, |\boldsymbol{k}| \le |\boldsymbol{t}|\}$, and will show that

$\sum_{k\in M_2'} PR(k,t) + \sum_{k\in M_2^*} PR(k,t) \longrightarrow 0$ as $n \to \infty$ with probability tending to 1. Thus, we can choose different C_* for M_2' and M_2^* as long as $C_* \geq (1+\delta^*)(1+2w)$. On the other hand, with probability tending to 1, by Condition (A3),

$$\begin{aligned}
\exp\Big\{-\frac{n(1-\epsilon)\lambda}{2}\|\widehat{\beta}_{t\setminus k}\|_2^2\Big\} &\leq \exp\Big[-\frac{n(1-\epsilon)\lambda}{2}\Big\{\|\beta_{0,t\setminus k}\|_2^2 - \|\widehat{\beta}_{t\setminus k} - \beta_{0,t\setminus k}\|_2^2\Big\}\Big] \\
&\leq \exp\Big[-\frac{n(1-\epsilon)\lambda}{2}\Big\{|t\setminus k|^2 \min_{j\in t}\beta_{0,j}^2 - c' w_n'^2\Big\}\Big] \\
&\leq \exp\Big\{-\frac{(1-\epsilon)\lambda}{2}(c_0|t\setminus k|^2|t| - c'|t\setminus k|)\Lambda_{|t|}\log p\Big\} \\
&\leq \exp\Big\{-\frac{(1-\epsilon)\lambda}{2}(c_0-c')|t\setminus k|^2|t|\Lambda_{|t|}\log p\Big\}
\end{aligned}$$

for any $k \in M_2$ and some large constants $c_0 > c' > 0$, where $w_n'^2 = |t\setminus k|\Lambda_{|t\setminus k|}\log p/n$. Here, $c' = 5\lambda^{-2}(1-\epsilon)^{-2}$ by the proof of Lemma A.3 in [5].

Hence, (A9) for any $k \in M_2$ is bounded above by

$$\begin{aligned}
&\exp\Big\{|t|\log\Lambda_{|t|} + C_*|t\setminus k|\log p - \frac{(1-\epsilon)\lambda}{2}(c_0-c')|t\setminus k|^2|t|\Lambda_{|t|}\log p\Big\} \\
\lesssim\ &\exp\Big\{-\Big(\frac{(1-\epsilon)\lambda}{2}(c_0-c') - C_* - o(1)\Big)|t\setminus k|^2|t|\Lambda_{|t|}\log p\Big\} \\
\leq\ &\exp\Big\{-\Big(\frac{(1-\epsilon)\lambda}{2}(c_0-c') - C_* - o(1)\Big)|t\setminus k|^2|t|\Lambda_{|t|}\log p\Big\} \\
\leq\ &\exp\Big\{-\Big(\frac{(1-\epsilon)\lambda}{2}(c_0-c') - C_* - o(1)\Big)|t|\Lambda_{|t|}\log p\Big\}
\end{aligned}$$

with probability tending to 1, where the last term is of order $o(1)$ because we assume $c_0 = \frac{1}{(1-\epsilon_0)\lambda}\{2(3+\delta) + \frac{5}{(1-\epsilon_0)\lambda}\} > \frac{2}{(1-\epsilon)\lambda}(C_* + o(1)) + c'$ for some small $\epsilon_0 > 0$.

It is easy to see that the maximum (A10) over $k \in M_2$ is also of order $o(1)$ with probability tending to 1 by the similar arguments. Since we have (A2) in the proof of Theorem 1, it completes the proof. □

Proof of Theorem 3. Let $M_2 = \{k : k \not\supseteq t, |k| \leq m_n\}$. Since we have Theorem 1, it suffices to show that

$$\sum_{k:k\in M_2} PR(k,t) \quad \xrightarrow{P} \quad 0 \quad \text{as } n\to\infty. \tag{A11}$$

By the proof of Theorem 2, the summation of (A9) over $k \in M_2$ is bounded above by

$$\begin{aligned}
&\textstyle\sum_{k\in M_2} p^{(C_*-2-\delta)(|k|-|t|)}\exp\Big\{-\Big(\frac{(1-\epsilon)\lambda}{2}(c_0-c') - C_* - o(1)\Big)|t\setminus k|^2|t|\Lambda_{|t|}\log p\Big\} \\
\leq\ &\textstyle\sum_{|k|=0}^{r}\sum_{v=0}^{(|t|-1)\wedge|k|}\binom{|t|}{v}\binom{r-|t|}{|k|-v}p^{-(|k|-|t|)}\exp\Big\{-\Big(\frac{(1-\epsilon)\lambda}{2}(c_0-c') - C_* - o(1)\Big)(|t|-v)^2|t|\Lambda_{|t|}\log p\Big\} \\
\leq\ &\textstyle\sum_{|k|=0}^{r}\sum_{v=0}^{(|t|-1)\wedge|k|}(|t|r)^{|t|-v}\exp\Big\{-\Big(\frac{(1-\epsilon)\lambda}{2}(c_0-c') - C_* - o(1)\Big)(|t|-v)^2|t|\Lambda_{|t|}\log p\Big\} \\
\leq\ &\exp\Big\{-\Big(\frac{(1-\epsilon)\lambda}{2}(c_0-c') - C_* - o(1)\Big)|t|\Lambda_{|t|}\log p + 2(|t|+2)\log p\Big\} \\
\leq\ &\exp\Big\{-\Big(\frac{(1-\epsilon)\lambda}{2}(c_0-c') - C_* - 6 - o(1)\Big)|t|\Lambda_{|t|}\log p\Big\}
\end{aligned}$$

with probability tending to 1, where $C_* \leq 3+\delta$ is defined in the proof of Theorem 2. Please note that the last term is of order $o(1)$ because we assume $c_0 = \frac{1}{(1-\epsilon_0)\lambda}\{2(9+2\delta) + \frac{5}{(1-\epsilon_0)\lambda}\} > \frac{2}{(1-\epsilon)\lambda}(C_* + 6 + o(1)) + c'$ for some small $\epsilon_0 > 0$. It is easy to see that the summation of (A10) over $k \in M_2$ is also of order $o(1)$ with probability tending to 1 by the similar arguments. □

Lemma A1. *Under Condition (A2), we have*

$$\exp\{\frac{1}{2}\widehat{\boldsymbol{\beta}}_k^\top(\boldsymbol{A}_k - \boldsymbol{A}_k(\boldsymbol{A}_k + \tau^{-1}\boldsymbol{U}_k)^{-1}\boldsymbol{A}_k)\widehat{\boldsymbol{\beta}}_k\} \lesssim 1$$

for any $\boldsymbol{k} \in M_1$ *with probability tending to 1.*

Proof. Please note that by Condition (A2),

$$(\boldsymbol{A}_k + \tau^{-1}\boldsymbol{U}_k)^{-1} \geq (\boldsymbol{A}_t + (n\tau\lambda)^{-1}\boldsymbol{A}_k)^{-1} = \frac{n\tau\lambda}{n\tau\lambda+1}\boldsymbol{A}_k^{-1},$$

which implies that

$$\frac{1}{2}\widehat{\boldsymbol{\beta}}_k^\top(\boldsymbol{A}_k - \boldsymbol{A}_k(\boldsymbol{A}_k + \tau^{-1}\boldsymbol{U}_k)^{-1}\boldsymbol{A}_k)\widehat{\boldsymbol{\beta}}_k \leq \frac{1}{2(n\tau\lambda+1)}\widehat{\boldsymbol{\beta}}_k^\top\boldsymbol{A}_k\widehat{\boldsymbol{\beta}}_k.$$

Thus, we complete the proof if we show that

$$\frac{1}{n\tau\lambda}\widehat{\boldsymbol{\beta}}_k^\top\boldsymbol{H}_n(\boldsymbol{\beta}_{0,k})\widehat{\boldsymbol{\beta}}_k \leq C$$

for some constant $C > 0$ and any $k \in M_1$ with probability tending to 1. By Lemma A.3 in [5] and Condition (A2),

$$\begin{aligned}\frac{1}{n\tau}\widehat{\boldsymbol{\beta}}_k^\top\boldsymbol{H}_n(\boldsymbol{\beta}_{0,k})\widehat{\boldsymbol{\beta}}_k &\leq \frac{1}{\tau}\lambda_{\max}\{n^{-1}\boldsymbol{H}_n(\boldsymbol{\beta}_{0,k})\}\|\widehat{\boldsymbol{\beta}}_k\|_2^2 \\ &\leq \frac{1}{\tau}(\log p)^d(\|\boldsymbol{\beta}_{0,k}\|_2^2 + o(1)) = O(1)\end{aligned}$$

for any $k \in M_1$ with probability tending to 1. □

References

1. Tibshirani, R. Regression shrinkage and selection via the lasso. *J. R. Stat. Soc. Ser. (Methodol.)* **1996**, *58*, 267–288. [CrossRef]
2. Fan, J.; Li, R. Variable selection via nonconcave penalized likelihood and its oracle properties. *J. Am. Stat. Assoc.* **2001**, *96*, 1348–1360. [CrossRef]
3. Zhang, C. Nearly unbiased variable selection under minimax concave penalty. *Ann. Stat.* **2010**, *38*, 894–942. [CrossRef]
4. Liang, F.; Song, Q.; Yu, K. Bayesian subset modeling for high-dimensional generalized linear models. *J. Am. Stat. Assoc.* **2013**, *108*, 589–606. [CrossRef]
5. Narisetty, N.; Shen, J.; He, X. Skinny gibbs: A consistent and scalable gibbs sampler for model selection. *J. Am. Stat. Assoc.* **2018**, 1–13. [CrossRef]
6. Ročková, V.; Georg, E. The spike-and-slab lasso. *J. Am. Stat. Assoc.* **2018**, *113*, 431–444. [CrossRef]
7. Johnson, V.; Rossell, D. On the use of non-local prior densities in bayesian hypothesis tests hypothesis. *J. R. Statist. Soc. B* **2010**, *72*, 143–170. [CrossRef]
8. Johnson, V.; Rossell, D. Bayesian model selection in high-dimensional settings. *J. Am. Stat. Assoc.* **2012**, *107*, 649–660. [CrossRef]
9. Shin, M.; Bhattacharya, A.; Johnson, V. Scalable bayesian variable selection using nonlocal prior densities in ultrahigh-dimensional settings. *Stat. Sin.* **2018**, *28*, 1053–1078.
10. Shi, G.; Lim, C.; Maiti, T. Bayesian model selection for generalized linear models using non-local priors. *Comput. Stat. Data Anal.* **2019**, *133*, 285–296. [CrossRef]
11. Wu, Ho.; Ferreira, M.R.; Elkhouly, M.; Ji, T. Hyper nonlocal priors for variable selection in generalized linear models. *Sankhya A* **2020**, *82*, 147–185. [CrossRef]
12. Hans, C.; Dobra, A.; West, M. Shotgun stochastic search for "large p" regression. *J. Am. Stat. Assoc.* **2007**, *102*, 507–516. [CrossRef]

13. Yang, X.; Narisetty, N. Consistent group selection with bayesian high dimensional modeling. *Bayesian Anal.* **2018**. [CrossRef]
14. Cao, X.; Khare, K.; Ghosh, M. High-dimensional posterior consistency for hierarchical non-local priors in regression. *Bayesian Anal.* **2020**, *15*, 241—262. [CrossRef]
15. Castillo, I.; Schmidt-Hieber, J.; Van der Vaart, A. Bayesian linear regression with sparse priors. *Ann. Stat.* **2015**, *43*, 1986–2018. [CrossRef]
16. McCullagh, P.; Nelder, J.A. *Generalized Linear Models*, 2nd ed.; Chapman & Hall: London, UK, 1989.
17. Lee, K.; Lee, J.; Lin, L. Minimax posterior convergence rates and model selection consistency in high-dimensional dag models based on sparse cholesky factors. *Ann. Stat.* **2019**, *47*, 3413–3437. [CrossRef]
18. Ishwaran, H.; Rao, J. Spike and slab variable selection: Frequentist and bayesian strategies. *Ann. Stat.* **2005**, *33*, 730–773. [CrossRef]
19. Song, Q.; Liang, F. Nearly optimal bayesian shrinkage for high dimensional regression. *arXiv* **2017**, arXiv:1712.08964.
20. Yang, Y.; Wainwright, M.; Jordan, M. On the computational complexity of high-dimensional bayesian variable selection. *Ann. Stat.* **2016**, *44*, 2497–2532. [CrossRef]
21. Tüchler, R. Bayesian variable selection for logistic models using auxiliary mixture sampling. *J. Comput. Graph. Stat.* **2008**, *17*, 76–94. [CrossRef]
22. Cai, X.; Huang, A.; Xu, S. Fast empirical bayesian lasso for multiple quantitative trait locus mapping. *BMC Bioinform.* **2011**, *12*, 211. [CrossRef] [PubMed]
23. Friedman, J.; Hastie, T.; Tibshirani, R. Regularization paths for generalized linear models via coordinate descent. *J. Stat. Softw.* **2010**, *33*, 1–22. [CrossRef] [PubMed]
24. Breheny, P.; Huang, J. Coordinate descent algorithms for nonconvex penalized regression, with applications to biological feature selection. *Ann. Appl. Stat.* **2011**, *5*, 232–253. [CrossRef] [PubMed]
25. Matthews, B.W. Comparison of the predicted and observed secondary structure of T4 phage lysozyme. *Biochim. Biophys. Acta (BBA) Protein Struct.* **1975**, *405*, 442–451. [CrossRef]
26. Meinshausen, N.; Bühlmann, P. High-dimensional graphs and variable selection with the lasso. *Ann. Stat.* **2006**, *34*, 1436–1462. [CrossRef]
27. Wei, L.; Hu, X.; Zhu, Y.; Yuan, Y.; Liu, W.; Chen, H. Aberrant intra-and internetwork functional connectivity in depressed Parkinson's disease. *Sci. Rep.* **2017**, *7*, 1–12. [CrossRef]
28. Zang, Y.; Jiang, T.; Lu, Y.; He, Y.; Tian, L. Regional homogeneity approach to fmri data analysis. *NeuroImage* **2004**, *22*, 394–400. [CrossRef]
29. Zuo, Xi.; Xu, T.; Jiang, L.; Yang, Z.; Cao, X.; He, Y.; Zang, Y.; Castellanos, F.; Milham, M. Toward reliable characterization of functional homogeneity in the human brain: Preprocessing, scan duration, imaging resolution and computational space. *NeuroImage* **2013**, *65*, 374–386. [CrossRef]
30. Zang, Y.; He, Y.; Zhu, C.; Cao, Q.; Sui, M.; Liang, M., Tian, L.; Jiang, T.; Wang, Y. Altered baseline brain activity in children with adhd revealed by resting-state functional mri. *Brain Dev.* **2007**, *29*, 83–91.
31. Zuo, Xi.; Kelly, C.; Di Martino, A.; Mennes, M.; Margulies, D.; Bangaru, S.; Grzadzinski, R.; Evans, A.; Zang, Y.; Castellanos, F.; et al. Growing together and growing apart: Regional and sex differences in the lifespan developmental trajectories of functional homotopy. *J. Neurosci.* **2010**, *30*, 15034–15043. [CrossRef]
32. Liu, Y.; Li, M.; Chen, H.; Wei, X.; Hu, G.; Yu, S.; Ruan, X.; Zhou, J.; Pan, X.; Ze Li; et al. Alterations of regional homogeneity in parkinson's disease patients with freezing of gait: A resting-state fmri study. *Front. Aging Neurosci.* **2019**, *11*, 276. [CrossRef] [PubMed]
33. Mi, T.; Mei, S.; Liang, P.; Gao, L.; Li, K.; Wu, T.; Chan, P. Altered resting-state brain activity in parkinson's disease patients with freezing of gait. *Sci. Rep.* **2017**, *7*, 16711. [CrossRef] [PubMed]
34. Prell, T. Structural and functional brain patterns of non-motor syndromes in parkinson's disease. *Front. Neurol.* **2018**, *9*, 138. [CrossRef] [PubMed]
35. Wang, J.; Zhang, J.; Zang, Y.; Wu, T. Consistent decreased activity in the putamen in Parkinson's disease: A meta-analysis and an independent validation of resting-state fMRI. *GigaScience* **2018**, *7*, 6. [CrossRef]
36. Zhang, F.; Jiang, W.; Wong, P.; Wang, J. A Bayesian probit model with spatially varying coefficients for brain decoding using fMRI data. *Stat. Med.* **2016**, *35*, 4380–4397. [CrossRef]

37. Quintero, F.O.L.; Contreras-Reyes, J.E.; Wiff, R.; Arellano-Valle, R.B. Flexible Bayesian analysis of the von Bertalanffy growth function with the use of a log-skew-*t* distribution. *Fish. Bull.* **2017**, *115*, 13–26. [CrossRef]
38. Lee, K.; Cao, X. Bayesian group selection in logistic regression with application to mri data analysis. In *Biometrics, to Appear*; Wiley: Hoboken, NJ, USA, 2020. [CrossRef]