Statistica Sinica **32** (2022), 2095-2117
doi:https://doi.org/10.5705/ss.202020.0427

PRIOR KNOWLEDGE GUIDED ULTRA-HIGH DIMENSIONAL VARIABLE SCREENING WITH APPLICATION TO NEUROIMAGING DATA

Jie He and Jian Kang

University of Michigan

Abstract: Variable screening is a powerful and efficient tool for dimension reduction under ultrahigh-dimensional settings. However, most existing methods overlook useful prior knowledge in specific applications. In this work, from a Bayesian modeling perspective, we develop a unified variable screening procedure for linear regression models. We discuss different constructions of posterior mean screening (PMS) statistics to incorporate different types of prior knowledge according to specific applications. With non-informative prior specifications, PMS is equivalent to the high-dimensional ordinary least-square projection (HOLP). We establish the screening consistency property for PMS with different types of prior knowledge. We show that PMS is robust to prior misspecifications. Furthermore, when the prior knowledge provides correct information on the true parameter settings, PMS can substantially improve the selection accuracy over that of the HOLP and other existing methods. We illustrate our method using extensive simulation studies and an analysis of neuroimaging data.

Key words and phrases: Linear regression, posterior mean screening, prior knowledge, screening consistency.

1. Introduction

Modern technologies have produced a vast amount of high-throughput data, in which the number of variables far outweighs the sample size. This has motivated the development of feature learning and screening methods: a powerful and efficient tool for dimension reduction (Fan and Fan (2008); Fan and Song (2010); Bühlmann and van de Geer (2011); Zhao and Li (2012)) in regression.

The pioneering work on variable screening was that on sure independence screening (SIS) (Fan and Lv (2008)), which has been extended to generalized linear models (Fan and Fan (2008); Fan, Samworth and Wu (2009); Fan and Song (2010)), generalized additive models (Fan, Feng and Song (2011)), quantile regression (He, Wang and Hong (2013); Ma, Li and Tsai (2017)), and the propor-

Corresponding author: Jian Kang, Department of Biostatistics, University of Michigan, Ann Arbor, MI 48109-2029, USA. E-mail: jiankang@umich.edu.

tional hazards model (Zhao and Li (2012); Gorst-Rasmussen and Scheike (2013)). By extending screening criteria based solely on the marginal correlations between the outcome and the predictors, a variety of statistics have been proposed that account for the dependence between the predictors, thus improving the screening accuracy and robustness (Li et al. (2012); Zhu et al. (2011); Cho and Fryzlewicz (2012); Hall and Miller (2009)). In particular, the high-dimensional ordinary least squares projection (HOLP) (Wang and Leng (2016)), which uses the generalized inverse of the design matrix in lieu of marginal correlations, exhibits good theoretical properties and high computational efficiency. In addition to the above methods based on model assumptions, variable screening has been generalized to model-free cases. As a result, the corresponding screening statistics are applicable to a general model class without a specific expression (Zhu et al. (2011); Li, Zhong and Zhu (2012); Cui, Li and Zhong (2015); Zhu et al. (2017); Pan et al. (2020)).

Variable screening has a wide range of applications in biomedical sciences, such as brain imaging and genetics. For instance, functional magnetic resonance imaging (fMRI) has been broadly employed to measure neural activities related to brain functions (Huettel, Song and McCarthy (2004); Smith and Fahrmeir (2007)). There is a growing interest in selecting important voxels with strong fMRI signals in order to identify brain regions that are highly associated with certain brain function behaviors or psychiatric disorders. The standard brain template for fMRI images contains 200,000 spatially contiguous voxels, which can be partitioned into groups based on the brain anatomy. In addition, existing studies may have identified locations or voxels where the brain activity is strongly associated with a response variable of interest, such as cognitive behaviors or disease status. This poses an interesting question on how to incorporate such prior information, including prior important knowledge and the prior spatial structure, into variable screening methods. Several methods have been developed to address this questions. For example, conditional sure independence screening (CSIS) (Barut, Fan and Verhasselt (2016); Hong, Kang and Li (2018)) directly includes predetermined important features into the model when screening variables. Similarly, the partition-based screening (PartS) method (Kang, Hong and Li (2017)) incorporates a spatial-guided partition structure into generalized linear models, and performs variable screening by dividing all covariates into groups, whereas the covariance-insured screening (CIS) method (He et al. (2019)) applies prior information through inter-feature dependence. However, these methods are all developed from a frequentist perspective, and only focus on incorporating one type of prior information. As straightforward approaches to integrating prior knowledge, Bayesian variable selection methods have been developed for neuroimaging applications (Smith and Fahrmeir (2007); Goldsmith, Huang and Crainiceanu (2014); Li et al. (2015); Kang, Reich and Staicu (2018)). These methods often aim to make a fully Bayesian inference on model selection, that is, selecting important predictor variables into the model, and then estimate the posterior distribution of the parameters, along with the posterior inclusion probability of each predictor variable. Thus, they incur large computational costs for high-dimensional problems, and many existing methods are not feasible for ultrahigh–dimensional problems. In contrast, the variable screening methods screen out predictor variables that are not strongly associated with the response variable, thus achieving efficient dimension reduction with less of a computational burden. There is a need to develop Bayesian variable screening methods to systematically incorporate prior knowledge and structural information in science.

In this work, from a Bayesian modeling perspective, we propose a unified feature screening procedure for the linear regression model. We construct the screening statistics using the posterior mean of the coefficients, which can incorporate prior information according to specific applications. Many prior models are available for linear regression, for example, the spike-slab priors (Ishwaran and Rao (2005)), non-local priors (Rossell and Telesca (2017)) and global-local shrinkage priors (Bhattacharya, Chakraborty and Mallick (2016)), for which efficient posterior computation methods are available for fully Bayesian inferences in the high-dimensional case. However, our focus is on variable screening. Thus we choose the normal prior for simplicity and good interpretations. The normal distribution is a conjugate prior, the closed form of the posterior mean is available, and the computation only involves non-iterative matrix operations. In addition, it is more straightforward to incorporate useful knowledge into the screening procedure using normal priors. To illustrate our idea, we consider a simple example. We simulate data from a linear regression model with 100 samples and generate 10,000 predictors from a multivariate normal distribution with mean zero and variance one. The correlation between any two predictors is set to 0.5. The true values of the regression coefficients are specified as $\beta_1 = \beta_2 = 3$, $\beta_3 = -7.5$, and $\beta_j = 0$, for $j = 4, \ldots, 10000$. Suppose we have prior knowledge that predictors 1 and 3 are more likely to be selected. We assign the normal priors to the regression coefficients and incorporate the prior selection information into the mean parameters using the empirical Bayes method. In this case, we obtain the closed form of the posterior mean of the regression coefficients; refer to Section 3.1 for more details. This is equal to our proposed posterior mean screening (PMS) statistics. We repeat the experiments 500 times and obtain the distributions of the screening

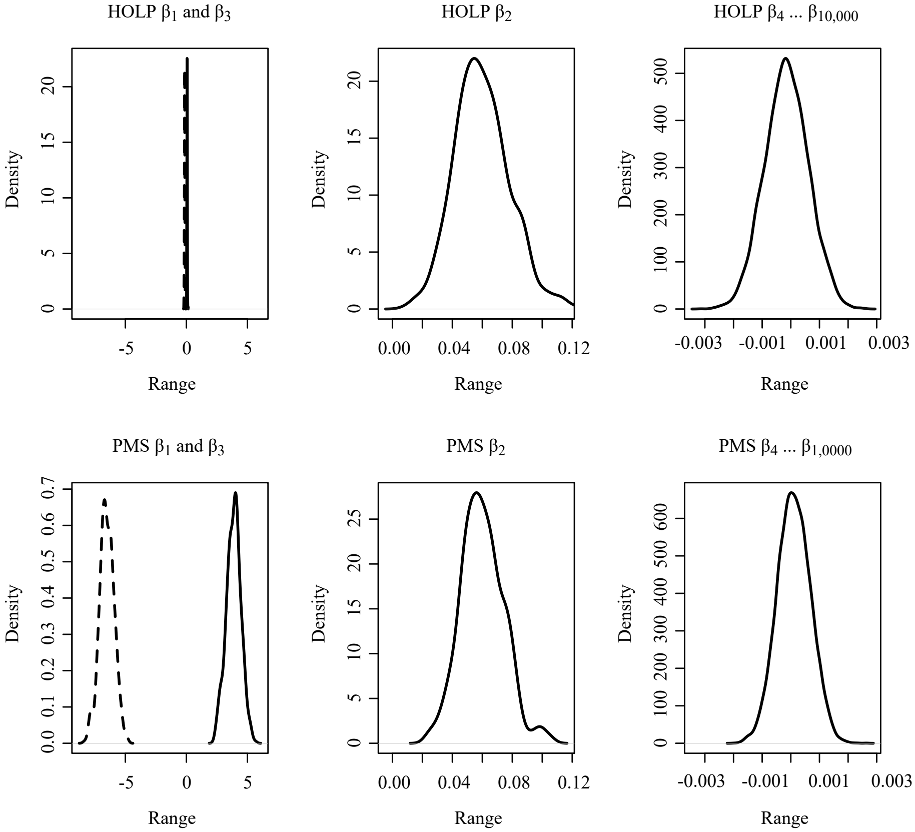

Figure 1. A simulated example for the distributions of the screening statistics by HOLP and posterior mean sceerning (PMS) based on 500 replicates. The true values are specified as follows: $\beta_1 = \beta_2 = 3, \beta_3 = -7.5$, and $\beta_j = 0$, for $j = 4, \ldots, 10000$. PMS incorporates the prior selection knowledge for β_1 and β_3 and thus obtains better screening statistics. PMS also improves the screening statistics for β_2 compared to HOLP.

statistics using PMS and HOLP; see Figure 1. For the active predictors 1 and 3, for which the prior selection information is incorporated, the screening statistics by PMS are clearly away from zero, whereas those by HOLP are concentrated at zero. For the active predictor 2, for which no prior selection information is incorporated, the PMS and the HOLP screening statistics are both much smaller than those for predictors 1 and 3. However, the PMS screening statistic is still clearly larger than zero, with probability one, while the lower tail of the HOLP screening statistic touches zero. This result clearly indicates that PMS is slightly better. For all other predictors ($j = 4, \ldots, 10000$), the distributions of both screening statistics are concentrated around zero (See the third column of Figure 1 for the mixtures of all the distributions), but the PMS screening statistic has a smaller variance. In summary, the normal prior can effectively incorporate prior selection knowledge and improve the screening accuracy. In this paper, we primarily discuss how to incorporate two types of prior knowledge: the prior selection, and the prior group structure. With a non-informative prior specification, PMS is equivalent to HOLP (Wang and Leng (2016)). We study the theoretical foundations of the proposed method. We discuss the technical conditions of the formulations of the prior knowledge to establish the screening consistency property. We show that our proposed feature screening method is very robust to prior misspecification. When the prior knowledge is consistent with the true parameter setting, the proposed method outperforms HOLP and other existing methods.

The remainder of this paper is organized as follows. In Section 2, we propose the unified framework of Bayesian feature screening for a linear regression model and establish the theoretical properties. In Section 3, we develop constructions of the prior mean and covariance information under specific cases, with their theoretical properties. In Section 4, we propose PMS variable screening-based ensemble learning to combine different types of prior knowledge. In Section 5, we evaluate the performance of the proposed method using a series of simulation studies. We apply the PMS method by analyzing neuroimaging data in Section 6. Section 7 concludes the paper. The Supplementary Material contains all technical proofs.

2. Posterior Mean Variable Screening

2.1. Notation and model specification

Let $\mathbb{R}^d$ be a d-dimensional vector space of real numbers, where $\mathbf{1}_d$ and $\mathbf{0}_d$ are d-dimensional vectors of all ones and all zeros, respectively. Denote by Sym_+^d the space of $d \times d$ symmetric positive-definite matrices with identity matrix $\mathbf{I}_d$. $\mathrm{N}_d(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ represents a d-dimensional normal distribution with mean $\boldsymbol{\mu} \in \mathbb{R}^d$ and covariance matrix $\boldsymbol{\Sigma} \in \mathrm{Sym}_+^d$. In addition, $I(\mathcal{A}) : \mathcal{F} \to \{0, 1\}$ refers to an event indicator, where $I(\mathcal{A}) = 1$ if event $\mathcal{A}$ occurs, and $I(\mathcal{A}) = 0$ otherwise. Notation $\|\cdot\|$ denotes the Euclidean norm. To any set $\mathcal{A}$, $|\mathcal{A}|$ represents the cardinal number of $\mathcal{A}$. For sequences of numbers x_n and y_n, $x_n = o(y_n)$ implies that $\lim_{n\to\infty} x_n/y_n = 0$, and $x_n = O(y_n)$ denotes that x_n/y_n is bounded.

Suppose the data set includes n observations of an outcome, along with p predictors. We consider the linear regression model

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}, \tag{2.1}$$

where $\mathbf{Y} = (y_1, \ldots, y_n)^{\mathrm{T}} \in \mathbb{R}^n$ denotes the outcome variable, $\mathbf{X} = (\mathbf{x}_1, \ldots, \mathbf{x}_p) \in \mathbb{R}^{n\times p}$ is a design matrix with rank $\min\{n, p\}$, $\boldsymbol{\epsilon} \in \mathbb{R}^n$ denotes independent and identically distributed (i.i.d.) errors with marginal distribution $\mathrm{N}(0, \sigma^2)$, and σ^2 is an unknown nuisance parameter. In addition, $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_p)^{\mathrm{T}} \in \mathbb{R}^p$ is a vector of coefficients for the predictors $\{x_j\}_{j=1}^p$.

2.2. PMS statistics

We assign a multivariate normal prior to $\boldsymbol{\beta}$:

$$\boldsymbol{\beta} \sim \mathrm{N}_p(\boldsymbol{\mu}, \tau^2\boldsymbol{\Lambda}), \tag{2.2}$$

where $\boldsymbol{\mu} = (\mu_1, \ldots, \mu_p)^{\mathrm{T}} \in \mathbb{R}^p$, $\boldsymbol{\Lambda} = (\lambda_{j,k}) \in \mathrm{Sym}_+^p$ with $\mathrm{Var}(\beta_j) = \tau^2\lambda_{j,j}$, and $\mathrm{Cov}(\beta_j, \beta_k) = \tau^2\lambda_{j,k}$. It is obvious that the parameter $\tau^2 > 0$ controls the overall prior variability of $\boldsymbol{\beta}$. Because the prior (2.2) is a conjugate prior, the posterior distribution of $\boldsymbol{\beta}$ given the other parameters is also a multivariate normal distribution: $(\boldsymbol{\beta} \mid \boldsymbol{\mu}, \boldsymbol{\Lambda}, \bullet) \sim \mathrm{N}_p(\boldsymbol{\nu}, \mathbf{K})$, where $\bullet = \{\sigma^2, \tau^2, \mathbf{X}, \mathbf{y}\}$, $\boldsymbol{\nu} = (\theta\boldsymbol{\Lambda}^{-1} + \mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1}(\theta\boldsymbol{\Lambda}^{-1}\boldsymbol{\mu} + \mathbf{X}^{\mathrm{T}}\mathbf{y})$ and $\mathbf{K} = (\theta\boldsymbol{\Lambda}^{-1} + \mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1}\sigma^2$. Here, $\theta = \sigma^2/\tau^2 > 0$ reflects the precision of the prior knowledge on the structure of the predictors.

We propose using the posterior mean $\boldsymbol{\nu}$ as the variable screening statistics, which incorporates prior knowledge $\boldsymbol{\mu}$ and $\boldsymbol{\Lambda}$. This method is referred to as posterior mean screening (PMS) in the rest of this paper, and is represented as

$$\widehat{\boldsymbol{\beta}}^{\mathrm{PMS}} = (\theta\boldsymbol{\Lambda}^{-1} + \mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1}(\theta\boldsymbol{\Lambda}^{-1}\boldsymbol{\mu} + \mathbf{X}^{\mathrm{T}}\mathbf{y}). \tag{2.3}$$

However, computing the screening statistics from (2.3) is not possible when p is on the scale of millions. From Proposition 1 in the Supplementary Material, we know that the PMS statistic is equivalent to:

$$\widehat{\boldsymbol{\beta}}^{\mathrm{PMS}} = \boldsymbol{\mu} + \boldsymbol{\Lambda}\mathbf{X}^{\mathrm{T}}\boldsymbol{\Omega}(\mathbf{Y} - \mathbf{X}\boldsymbol{\mu}), \tag{2.4}$$

where $\boldsymbol{\Omega} = (\mathbf{X}\boldsymbol{\Lambda}\mathbf{X}^{\mathrm{T}} + \theta\mathbf{I}_n)^{-1}$. Because the inverse of $\boldsymbol{\Omega}$ is an $n \times n$ matrix, the computation can be simplified to a large degree, even though $p \gg n$. From equation (2.4), we observe that PMS reduces to the HOLP when $\boldsymbol{\mu} = \mathbf{0}_p$, $\boldsymbol{\Lambda} = \mathbf{I}_p$, and $\theta = 0$. Thus, the HOLP is a special case of PMS with an uninformative prior.

In practise, we compute the PMS statistic using a fast algorithm, which includes the following three steps: (1) compute $\widetilde{\mathbf{X}} = \mathbf{X}\boldsymbol{\Lambda}$; (2) compute $\mathbf{b} = \boldsymbol{\Omega}(\mathbf{y} - \mathbf{X}\boldsymbol{\mu})$ by solving $(\widetilde{\mathbf{X}}\mathbf{X}^{\mathrm{T}} + \theta\mathbf{I}_n)\mathbf{b} = (\mathbf{y} - \mathbf{X}\boldsymbol{\mu})$; and (3) compute $\widehat{\boldsymbol{\beta}}^{\mathrm{PMS}} = \boldsymbol{\mu} + \widetilde{\mathbf{X}}^{\mathrm{T}}\mathbf{b}$.

Then, we obtain output $\widehat{\boldsymbol{\beta}}^{\mathrm{PMS}}$. Given a thresholding parameter α, the selected index set can be expressed as $\widehat{\mathcal{M}}_\alpha = \{j : |\hat{\beta}_j^{\mathrm{PMS}}| > \alpha\}$.

For a general case, when $p \gg n$, the time complexity to obtain $\widehat{\mathcal{M}}_\alpha$ is $O(np^2)$. The main computing bottleneck is in the matrix multiplication between $\boldsymbol{\Lambda}$ and $\mathbf{X}$. Existing matrix parallel computing methods can be applied directly to this step, thus reducing the computational time. When $\boldsymbol{\Lambda}$ is specified as sparse, the computational cost can be further reduced. For example, if the number of nonzero elements in $\boldsymbol{\Lambda}$ is $O(p)$, the computational complexity can be reduced to $O(n^2p)$, which is the same as that of the HOLP (Wang and Leng (2016)). In many applications, a sparse specification on $\boldsymbol{\Lambda}$ is common and sensible for two reasons. First, in many high-dimensional problems, it is common to assume that the number of true features is much smaller than the number of candidate features. This assumption implies that the true values of $\boldsymbol{\beta}$ are sparse. Thus, one may specify the nonzero prior covariance on a small set of predictors that are considered to be the most likely possible candidates on the true features. This specification leads to sparse $\boldsymbol{\Lambda}$. Second, the sufficient prior knowledge on the dependence between all pairs of β_j is usually limited. It is common to specify only a few connected region pairs. Note that $\boldsymbol{\Lambda}$ is the prior covariance on $\boldsymbol{\beta}$ and does not have a true value. Thus, simple and sparse specifications are preferred.

2.3. Screening consistency

In this section, we establish the theoretical properties of the proposed PMS variable screening method. Denote $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_p)^{\mathrm{T}}$ as the true parameter vector of interest, and $\mathcal{M}_0 = \{j : \beta_j \neq 0\}$ as the index set composed of nonzero coefficients. Throughout this paper, $x_n = o_p(y_n)$ indicates that x_n/y_n converges to zero in probability as $n \to \infty$, and $x_n = O_p(y_n)$ means that x_n/y_n is bounded in probability.

We need some regularity conditions to establish the theoretical property, where Conditions A1–A3 are similar to those introduced by the HOLP (Wang and Leng (2016)) and are listed in the Supplementary Material. We list a few other conditions that are related to prior specifications.

*A*4. For some $c_7 > 0$ and $\gamma \geq 0$, $\max_{j\in\{1,\ldots,p\}} \left|\mathbf{e}_j^{\mathrm{T}} \boldsymbol{\Lambda}^{-1} (\boldsymbol{\mu} - \boldsymbol{\beta})\right| \leq c_7 n^{\gamma}/p$.

*A*5. For the matrix $\boldsymbol{\Lambda}$, let λ_{ij} be the element in the ith row and the jth column. Then, there exists $\nu > 0$, such that $\max_{i\in\{1,\ldots,p\}} \lambda_{ii} = O(n^{\nu})$, and $\sum_{j=1}^{p} |\lambda_{ij}| = O\left(n^{\nu-2\tau_3-2\tau_4-2\gamma}/\sqrt{\log n}\right)$, for $j \neq i; i, j = 1, \ldots, p$. For some $c_8 > 1$, $\min_{j\in\mathcal{M}_0} |\beta_j| \geq c_8 n^{1-(\tau_3+\tau_4)-\gamma+\nu}/p$. In addition, denote the ith row of matrix $\boldsymbol{\Lambda}^{1/2}$ as $\bar{\boldsymbol{\lambda}}_i = \left(\bar{\lambda}_{i1}, \ldots, \bar{\lambda}_{ip}\right)^{\mathrm{T}}$, for $i = 1, \ldots, p$. Then,

$\sum_{u=1}^{p}\sum_{v\neq u}\left|\bar{\lambda}_{iu}\bar{\lambda}_{jv}\right| = O(n^{\nu})$, and $\sum_{u=1}^{p}\left|\bar{\lambda}_{iu}\bar{\lambda}_{ju}\right| = O(n^{1-(\tau_3+\tau_4)-2\gamma+\nu}/\sqrt{\log n})$, for $j\neq i; i,j=1,\ldots,p$. Here, $p > cn^{1+\nu}$, for some $c > 1$.

*A*6. Assume $\log p = o[\min\{n^{1-\xi_1}/\log n, q(\tilde{C}n^{1/2-\xi_2}/\sqrt{\log n})\}]$, for some $0<\xi_1<1$, $0<\xi_2<1/2$, and $\tilde{C}>0$, and there exists α_n, such that $\sqrt{p}\alpha_n/n^{1-(\tau_3+\tau_4)-\gamma+\nu} \to 0$ and $\alpha_n\sqrt{p\log n}/n^{1-(\tau_3+\tau_4)-\gamma+\nu}\to\infty$.

Condition A4 provides an upper bound on the difference between the prior mean parameter and the true values of the regression coefficients. This condition implies that PMS may still enjoy screening consistency, even when the prior mean deviates slightly from the true settings. Condition A5 imposes some upper bound constraints on the diagonal elements and off-diagonal elements of $\mathbf{\Lambda}$, which are related to the prior variance and prior correlations of the predictor effects $\boldsymbol{\beta}$, respectively. This condition implies that, to ensure the screening consistency, neither the prior variance nor the correlation can increase too fast as the sample size n increases. This condition can be straightforwardly verified when $\mathbf{\Lambda}$ is an identity matrix, where the predictor effects are assumed to be prior independent. When $\mathbf{\Lambda}$ is a sparse matrix, such as a band matrix or a block diagonal matrix, condition A5 can be simplified and converted to conditions related to the upper bounds of the bandwidth or the block size, providing insights on prior specifications in practice. For example, in the scalar-on-image regression for the neuroimaging application, sparse block diagonal covariance structures can be adopted based on the brain function and anatomical region partitions. Condition A5 provides insights on upper bounds of the number of brain regions, number of voxels within regions, and maximum correlation within each region in the order of the sample size.

Theorem 1 (Screening Consistency)**.** *Under Conditions A*1*–A*6*, we have*

$$P\left(\min_{j\in\mathcal{M}_0}\left|\hat{\beta}_j^{\mathrm{PMS}}\right| > \alpha_n > \max_{j\notin\mathcal{M}_0}\left|\hat{\beta}_j^{\mathrm{PMS}}\right|\right)$$
$$= 1 - O\left[\exp\left(\frac{-Cn^{1-\xi_1}}{2\log n}\right) + \exp\left\{1-\frac{1}{2}q\left(\frac{\tilde{C}n^{1/2-\xi_2}}{\sqrt{\log n}}\right)\right\}\right],$$

for some constants $C, \tilde{C} > 0$.

Theorem 1 indicates that, under some mild regularity conditions, PMS can perfectly separate the important and unimportant variables with probability tending to one. The HOLP has a similar result, but the order of the convergence rate is different to that of PMS.

2.4. Choice of thresholding parameters

The thresholding parameter α is critical to the performance of the variable screening procedure. Overestimating α inflates the false positive rate, while underestimating α hinders sure screening. We adopt the random decoupling method (Barut, Fan and Verhasselt (2016)) to select the thresholding parameter α.

To ensure the stability of α, we replicate the decoupling procedure K times. Given the data $\{(\mathbf{x}_i, Y_i)\}_{i=1}^n$, we allow a $(1-\tau_r)100\%$ proportion ($\tau_r \in [0,1]$) of inactive variables to be included in the model when $\mathbf{X}$ and $\mathbf{Y}$ are not related, which corresponds to the null model. We randomly permute the rows of the design matrix $\mathbf{X}$ and obtain the pseudo data $\{(\widetilde{\mathbf{x}}_i^{(k)}, Y_i)\}_{i=1}^n$, with $\widetilde{\mathbf{x}}_i^{(k)} = \mathbf{x}_{\pi_k(i)}$, where $\pi_{k(i)}$ is a permutation of index i, for $i = 1, \ldots, n$, and $k = 1, \ldots, K$. Based on the above expression, we obtain the values of the PMS statistics

$$\{|\hat{\beta}_j^{\mathrm{PMS}(k)}|, j = 1, \ldots, p; k = 1, \ldots, K\},$$

where α_k^* is the τ_r-quantile of $\{|\hat{\beta}_j^{\mathrm{PMS}(k)}|, j = 1, \ldots, p\}$. Finally, we choose $\alpha^* = \max_{1 \le k \le K} \alpha_k^*$ as the thresholding parameter.

3. Incorporating Prior Knowledge

When specific information on $\boldsymbol{\mu}$ and $\boldsymbol{\Lambda}$ is available, we can straightforwardly carry out PMS by computing $\widehat{\boldsymbol{\beta}}^{\mathrm{PMS}}$ based on (2.4). However, in many cases, the prior knowledge is relatively vague, which means it is not trivial to specify the values of $\boldsymbol{\mu}$ and $\boldsymbol{\Lambda}$. In this section, we discuss how to construct different types of prior knowledge systematically based on the PMS framework.

3.1. Priors on selection

Suppose we are interested in incorporating prior knowledge on which features should be selected into the PMS framework. For instance, in a functional neuroimaging study, we are interested in selecting important brain locations that are highly associated with the intelligence quotient (IQ). According to some studies, we may know some brain regions are more likely to be selected than the other regions. This type of information can be used to classify all features into two groups: prior-selected and prior-unselected.

In general, denote by $\mathcal{S}$ the indices of the prior-selected features, where $q = |\mathcal{S}|$ is the number of elements in $\mathcal{S}$. Assume that $\mu_j = 0$ for feature $j \notin \mathcal{S}$ and $\boldsymbol{\mu}_{\mathcal{S}} = (\mu_j : j \in \mathcal{S})$. Introducing a hyper-prior normal distribution to $\boldsymbol{\mu}_{\mathcal{S}}$, the sampling distribution of $\mathbf{Y}$ given the prior-selected features is $\mathrm{N}_n(\mathbf{X}_{\mathcal{S}}\boldsymbol{\mu}_{\mathcal{S}}, \boldsymbol{\Omega}_{\mathcal{S}}^{-1})$, where $\mathbf{X}_{\mathcal{S}} = (\mathbf{x}_j, j \in \mathcal{S})$, $\boldsymbol{\Omega}_{\mathcal{S}} = (\mathbf{X}_{\mathcal{S}}\boldsymbol{\Lambda}_{\mathcal{S}}\mathbf{X}_{\mathcal{S}}^{\mathrm{T}} + \theta\mathbf{I}_n)^{-1}$, and $\boldsymbol{\Lambda}_{\mathcal{S}}$ is the prior correlation matrix of coefficients for the features in $\mathcal{S}$. When $q \leq n$, we can assign $\boldsymbol{\mu}_{\mathcal{S}}$ the uninformative hyper-prior, that is, $\pi(\boldsymbol{\mu}_{\mathcal{S}}) \propto 1$. When $q > n$, we assign $\boldsymbol{\mu}_{\mathcal{S}}$ a normal prior, that is, $\boldsymbol{\mu}_{\mathcal{S}} \sim \mathrm{N}(\mathbf{0}_q, \tilde{\tau}^2\mathbf{I}_q)$ for $\tilde{\tau}^2 > 0$. We construct the PMS selection statistics, $\widehat{\boldsymbol{\beta}}_{\mathcal{S}}^{\mathrm{PMS}}$, in Proposition 1.

Proposition 1. *The PMS selection statistic is*

$$\widehat{\boldsymbol{\beta}}_{\mathcal{S}}^{\mathrm{PMS}} = \tilde{\boldsymbol{\mu}} + \boldsymbol{\Lambda}\mathbf{X}^{\mathrm{T}}\boldsymbol{\Omega}(\mathbf{Y} - \mathbf{X}\tilde{\boldsymbol{\mu}}), \tag{3.1}$$

where $\tilde{\boldsymbol{\mu}} = (\tilde{\mu}_1, \ldots, \tilde{\mu}_p)^{\mathrm{T}}$ *with* $\tilde{\mu}_j = 0$, *for* $j \notin \mathcal{S}$, *and* $(\tilde{\mu}_j : j \in \mathcal{S}) = \tilde{\boldsymbol{\mu}}_{\mathcal{S}}$ *with*

$$\tilde{\boldsymbol{\mu}}_{\mathcal{S}} = \mathrm{E}(\boldsymbol{\mu}_{\mathcal{S}} \mid \boldsymbol{\Lambda}, \bullet) = \begin{cases} (\mathbf{X}_{\mathcal{S}}^{\mathrm{T}}\boldsymbol{\Omega}_{\mathcal{S}}\mathbf{X}_{\mathcal{S}})^{-1}\mathbf{X}_{\mathcal{S}}^{\mathrm{T}}\boldsymbol{\Omega}_{\mathcal{S}}\mathbf{Y} & q \leq n \\ (\mathbf{X}_{\mathcal{S}}^{\mathrm{T}}\boldsymbol{\Omega}_{\mathcal{S}}\mathbf{X}_{\mathcal{S}} + \tilde{\tau}^{-2}\mathbf{I}_q)^{-1}\mathbf{X}_{\mathcal{S}}^{\mathrm{T}}\boldsymbol{\Omega}_{\mathcal{S}}\mathbf{Y} & q > n. \end{cases}$$

Given $\boldsymbol{\Omega}$, the complexity of computing $\widehat{\boldsymbol{\beta}}_{\mathcal{S}}^{\mathrm{PMS}}$ is no more than $O(nq^2)$. This additional computational cost is moderate. When $q \leq n$, we can establish the screening consistency property for $\widehat{\boldsymbol{\beta}}_{\mathcal{S}}^{\mathrm{PMS}}$. Let $\mathcal{S}_3 = \mathcal{S}^c \cap \mathcal{M}_0$ be a set composed of all nonselected features. The following is an essential regularization condition for the screening consistency of $\widehat{\boldsymbol{\beta}}_{\mathcal{S}}^{\mathrm{PMS}}$.

B1. There exists a constant $\bar{\gamma}_1 > 0$, such that $\|\mathbf{X}_{\mathcal{S}_3}\boldsymbol{\beta}_{\mathcal{S}_3}\| = O\left(n^{\bar{\gamma}_1}\right)$, where $\mathbf{X}_{\mathcal{S}_3} = (X_j, j \in \mathcal{S}_3)$ and $\boldsymbol{\beta}_{\mathcal{S}_3} = (\beta_j, j \in \mathcal{S}_3)$.

Condition B1 indicates that the number of prior-not-selected features should not be too large. At the same time, the signal strength of those features cannot be too strong and the upper bound is on the polynomial order of n. Other regularity conditions, including B2 and B3, are listed in the Supplementary Material. B2 imposes restrictions on the eigenvalues of the matrix $\boldsymbol{\Lambda}_{\mathcal{S}}$, and B3 makes assumptions on the structure of a sub-matrix of $\boldsymbol{\Lambda}^{-1}$, a matrix composed of elements in $\boldsymbol{\Lambda}^{-1}$, the rows and columns of which are in $\mathcal{S}$.

Theorem 2 (Screening Consistency for $\widehat{\boldsymbol{\beta}}_{\mathcal{S}}^{\mathrm{PMS}}$)**.** *Under Conditions* B1–B3*, when* $q \leq n$*, the PMS statistic* (3.1) *enjoys the screening consistency property.*

When $q > n$, additional assumptions on $\tilde{\tau}^2$ are needed to ensure screening consistency. For example, if we assume $\tilde{\tau}^2 \to 0$ as $n \to \infty$, then $\tilde{\boldsymbol{\mu}} \to \mathbf{0}_p$ with probability one. In this case, the prior selection information vanishes when the sample size increases. However, we can verify Condition A4 by making other mild assumptions, and use Theorem1 to establish the screening consistency. For more general and weaker assumptions, a rigorous proof is nontrivial. Thus we leave this to future research. In practice, with a choice of $\tilde{\tau}^{-2} = 10^{-3}$, in which case the prior selection information plays an important role, we show that PMS performs very well in simulations and data applications. See Sections 5.2 and 6.

3.2. Priors on group-level importance

For some applications, prior knowledge can be straightforwardly used to determine the groups of features, but within each group, the importance of features may be difficult to distinguish. For example, in the analysis of brain imaging data, the whole brain region can be partitioned into a set of exclusive regions according to brain anatomical structures. The commonly used brain atlas includes the automated anatomical labeling (Tzourio-Mazoyer et al. (2002, AAL)) system, consisting of 90 regions. A standard brain template with a single voxel size of 2 mm^3 contains more than 180,000 voxels in AAL 116 regions. The goal is to select voxel-level features by incorporating region-level information. It is reasonable to assume that within the same region, voxels have a prior with the same level of importance.

In general, suppose the prior knowledge can partition all the features into m groups. For many applications, we can assume $m < n$. Let $\mathbf{B} = (b_{g,j})$ be an $m \times p$ group indicator matrix with $b_{g,j} \in \{0, 1\}$, where $b_{g,j} = 1$ indicates that feature j belongs to group g, and $b_{g,j} = 0$ otherwise. Note that the column sum of $\mathbf{B}$ is equal to one, that is, $\mathbf{1}_m^{\mathrm{T}}\mathbf{B} = \mathbf{1}_p$, which indicates that feature j has to be uniquely assigned to one group. Let $\bar{\mu}_g$ represent the level of importance for group g, and write $\bar{\boldsymbol{\mu}} = (\bar{\mu}_1, \ldots, \bar{\mu}_m)^{\mathrm{T}}$. We assume that within-group features have the same levels of importance, that is, $\boldsymbol{\mu} = \mathbf{B}^{\mathrm{T}}\bar{\boldsymbol{\mu}}$. Then, the sampling distribution of $\mathbf{Y}$ given $\mathbf{X}$ and the prior selected group structure is $\mathrm{N}_n(\mathbf{X}\mathbf{B}^{\mathrm{T}}\bar{\boldsymbol{\mu}}, \boldsymbol{\Omega}_{\mathbf{B}}^{-1})$, where $\boldsymbol{\Omega}_B = (\mathbf{X}\mathbf{B}^{\mathrm{T}}\boldsymbol{\Lambda}_{\mathbf{B}}\mathbf{B}\mathbf{X}^{\mathrm{T}} + \theta\mathbf{I}_n)^{-1}$ and $\boldsymbol{\Lambda}_{\mathbf{B}}$ is the prior correlation matrix for $\mathbf{B}\boldsymbol{\beta}$. Under this setting, we obtain the PMS group statistics ($\widehat{\boldsymbol{\beta}}_{\mathbf{B}}^{\mathrm{PMS}}$) from the following Proposition 2.

Proposition 2. *Under the non-informative prior assumption* $\pi(\bar{\boldsymbol{\mu}}) \propto 1$, *the posterior mean of* $\bar{\boldsymbol{\mu}}$ *is* $\widetilde{\boldsymbol{\mu}} = \mathrm{E}(\bar{\boldsymbol{\mu}} \mid \boldsymbol{\Lambda}, \bullet) = (\mathbf{B}\mathbf{X}^{\mathrm{T}}\boldsymbol{\Omega}_{\mathbf{B}}\mathbf{X}\mathbf{B}^{\mathrm{T}})^{-1}\mathbf{B}\mathbf{X}^{\mathrm{T}}\boldsymbol{\Omega}_{\mathbf{B}}\mathbf{Y}$. *Then the PMS group statistics can be expressed as*

$$\widehat{\boldsymbol{\beta}}_{\mathbf{B}}^{\mathrm{PMS}} = \mathbf{B}^{\mathrm{T}}\widetilde{\boldsymbol{\mu}} + \boldsymbol{\Lambda}\mathbf{X}^{\mathrm{T}}\boldsymbol{\Omega}(\mathbf{Y} - \mathbf{X}\mathbf{B}^{\mathrm{T}}\widetilde{\boldsymbol{\mu}}). \tag{3.2}$$

To establish the screening consistency for $\widehat{\boldsymbol{\beta}}_{\mathbf{B}}^{\mathrm{PMS}}$, we impose several regularity conditions. In particular, we need the following condition.

C2. Let $\mathbf{B}^*$ be the true group indicator matrix. Then, there exists some constant $\bar{\gamma}_2 > 0$, such that $\|(\mathbf{B} - \mathbf{B}^*)\bar{\boldsymbol{\beta}}\| = O(n^{\bar{\gamma}_2}/\sqrt{p})$.

Condition C2 indicates that the group structure $\mathbf{B}$ for active features should not be too far away from the truth. Furthermore, the PMS group statistics tolerate a certain extent of false negative rate. Other conditions, including C1, C3 and C4, are listed in the Supplementary Material. C1 and C3 impose assumptions on the eigenvalues of the matrix $\mathbf{B}\mathbf{B}^{\mathrm{T}}$ and $\mathbf{\Lambda}_{\mathbf{B}}$; C4 adds constraints on the structure of the matrix $\mathbf{B}\mathbf{\Lambda}^{-1}\mathbf{B}^{\mathrm{T}}$.

Theorem 3 (Screening Consistency for $\widehat{\boldsymbol{\beta}}_{\mathbf{B}}^{\mathrm{PMS}}$)**.** *Under Conditions C1–C4, the PMS group statistics* (3.2) *enjoy the screening consistency property.*

Theorems 1, 2, and 3 indicate that our proposed PMS method is robust to prior knowledge. Even though the prior knowledge is not exactly correct, as long as it is not too far away from the truth, the screening results are still consistent. We also demonstrate this property in our simulations in Section 5.

4. PMS Screening-Based Ensemble Learning

In many applications, multiple sources of prior knowledge may be available, but none is clearly better than the others. We propose tackling this issue by combining PMS screening (CPMS), the screening statistics that integrate prior knowledge from different sources simultaneously. Assume that we have $K (K < \infty)$ types of prior knowledge, and denote by $\widehat{\boldsymbol{\beta}}^{(1)}, \ldots, \widehat{\boldsymbol{\beta}}^{(K)}$ the corresponding PMS statistics. We introduce each entry of the CPMS statistics ($\widehat{\boldsymbol{\beta}}^{\mathrm{CPMS}}$) as $\hat{\beta}_j^{\mathrm{CPMS}} = \max\{|\hat{\beta}_j^{(1)}|, \ldots, |\hat{\beta}_j^{(K)}|\}$, for $j = 1, \ldots, p$. Given a thresholding parameter α, the selected feature set can be expressed as $\widetilde{\mathcal{M}}_\alpha = \{j : \hat{\beta}_j^{\mathrm{CPMS}} > \alpha\}$. For CPMS, the thresholding parameter α can also be selected using the random decoupling method. Similarly to PMS, we demonstrate the theoretical properties of CPMS in the following Theorem 4.

Theorem 4 (Screening Consistency for CPMS)**.** *If Conditions A1–A5 hold for all $k, k = 1, \ldots, K$, there exist α_n^K, C' and $\tilde{C}' > 0$, such that*

$$P\left(\min_{j\in\mathcal{M}_0}\left|\hat{\beta}_j^{\mathrm{CPMS}}\right| > \alpha_n^K > \max_{j\notin\mathcal{M}_0}\left|\hat{\beta}_j^{\mathrm{CPMS}}\right|\right)$$
$$= 1 - O\left[\exp\left(\frac{-C' n^{1-\tilde{\xi}_1}}{2\log n}\right) + \exp\left\{1 - \frac{1}{2}q\left(\frac{\tilde{C}' n^{1/2-\tilde{\xi}_2}}{\sqrt{\log n}}\right)\right\}\right],$$

for some $0 < \tilde{\xi}_1 < 1$ and $0 < \tilde{\xi}_2 < 1/2$.

5. Simulation Studies

We compared the performance of the proposed PMS method with that of existing variable screening methods, such as the SIS, HOLP, CIS, and PartS methods, using a series of simulation studies. We assigned two settings to a p-dimensional linear regression model (2.1): a group structure, and an image regression case. All simulation results are based on 200 replicates and a signal-to-noise ratio $R^2 = 0.5$ and 0.9. We evaluate the screening accuracy using three criteria: the false positive rate (FPR) when the power of detecting the true signals is 80%, the false negative rate (FNR) when the FPR is controlled at 10%, and the median number of variables needed to include all true signals (Model Size). In all the tables, we report the FPR and the FNR multiplied by 1,000, with the standard deviations in the parentheses, and the model size, with the corresponding 25% and 75% quantile intervals in parentheses. The R package `PMS` (`https://github.com/kangjian2016/PMS.git`) is available.

5.1. Group structure case

Similarly to a scenario in Zou and Hastie (2005), we introduced a group structure to all covariates in this setting. The specific distribution information is summarized as $x_{j+3m} = z_j + \mathrm{N}(0, \delta^2)$, where $z_j \sim \mathrm{N}(0,1)$, for $j = 1, 2, 3$ are independent, $\delta^2 = 0.01$, and $m = 0, \ldots, 4$. In addition, $x_j \sim \mathrm{N}(0, \delta^2)$, for $j = 16, \ldots, p$, are independent. Thus, the first 15 features are divided into three groups, denoted as $\mathcal{G}_k^0 = \{x_{k+3m}, m = 0, \ldots, 4\}$, for $k = 1, 2, 3$, and the regression coefficients are set as $\beta_{\mathcal{G}_1^0} = 0.5$, $\beta_{\mathcal{G}_2^0} = 3$, $\beta_{\mathcal{G}_3^0} = 5$, and $\beta_j = 0$, for $j = 16, \ldots, p$.

Assume we have another group series $\mathcal{G}_k = \{x_{k+3m} : m = 0, \ldots, 9\}$, for $k = 1, 2, 3$. Obviously, half of the elements in $\mathcal{G}_k$ are active features, whereas the others are inactive. For each $\mathcal{G}_k$, we exchanged two active features with two active ones from the next group, and obtained new group series $\{\tilde{\mathcal{G}}_k\}_{k=1}^3$. We constructed the prior covariance information of PMS based on $\{\tilde{\mathcal{G}}_k\}_{k=1}^3$. Letting $\tilde{\mathcal{G}}_4 = \{x_j, j = 31, \ldots, p\}$ and $\mathbf{L}$ be the normalized graph Laplacian matrix for $\{\tilde{\mathcal{G}}_k\}_{k=1}^4$, we introduced a network structure to the prior covariance matrix, of the form $\mathbf{\Lambda} = (\mathbf{L} + \varepsilon \mathbf{I}_p)^{-1}$, with $\varepsilon = 10^{-3}$. Because $\{\tilde{\mathcal{G}}_k\}_{k=1}^3$ are inconsistent with $\{\mathcal{G}_k\}_{k=1}^3$, the matrix $\mathbf{\Lambda}$ is misspecified. For PMS with prior selection ("PMS-selection"), the pre-selected set is $\mathcal{S} = \bigcup_{k=1}^3 \tilde{\mathcal{G}}_k$. For PMS with prior group information ("PMS-group"), $\{\tilde{\mathcal{G}}_k\}_{k=1}^4$ is chosen as the group partition. We considered two PartS methods with different partition structures. One is based on the true group structure $\{\mathcal{G}_k^0\}_{k=1}^4$, and referred to as "PartS-I." The other structure, called "PartS-II," is obtained by randomly partition-

Table 1. Screening accuracy for predictors, with group structure.

		$R^2 = 0.5$			$R^2 = 0.9$		
(n, p)	Method	FPR	FNR	Model Size	FPR	FNR	Model Size
$(100, 10000)$	SIS	430(286)	294(128)	5108(2270,7931)	429(300)	268(126)	5271(1543,8190)
	HOLP	434(283)	298(120)	5043(2550,8093)	430(298)	272(122)	5361(1599,8265)
	CIS	38(6)	0(0)	412(385,429)	24(7)	0(0)	287(257,309)
	PartS-I	296(236)	273(110)	8209(5809,9421)	242(206)	244(119)	7654(5525,8875)
	PartS-II	345(208)	314(106)	8212(5917,9422)	330(0.216)	285(0.082)	7675(5538,8878)
	PMS-selection	0(0)	0(5)	24(20,28)	0(0)	1(7)	24(20,27)
	PMS-group	11(58)	22(63)	30(29,30)	2(13)	16(46)	30(30,30)
$(100, 20000)$	SIS	434(278)	287(120)	10813(5088,16144)	421(291)	277(121)	9516(4181,16428)
	HOLP	442(280)	190(116)	10877(5119,16330)	424(288)	277(121)	9707(4113,16189)
	CIS	22(3)	0(0)	458(434,478)	15(3)	0(0)	332(299,351)
	PartS-I	290(186)	328(97)	17266(14546,18544)	249(183)	270(87)	17229(14198,18725)
	PartS-II	347(161)	368(96)	17395(14569,18546)	310(177)	300(67)	17293(14270,18729)
	PMS-selection	0(0)	0(5)	24(19,28)	0(0)	1(8)	24(20,28)
	PMS-group	6(38)	10(45)	30(29,30)	2(20)	10(37)	30(30,30)
$(200, 20000)$	SIS	363(292)	248(140)	7912(2675,14709)	305(284)	225(151)	5948(1668,11813)
	HOLP	370(285)	256(133)	8074(3082,14809)	301(0.274)	227(151)	5574(1803,12178)
	CIS	29(5)	0(0)	636(615,654)	13(7)	0(0)	404(290,433)
	PartS-I	219(213)	230(119)	15588(8152,18236)	183(201)	201(125)	13689(6148,17489)
	PartS-II	305(224)	269(97)	15611(8159,18239)	249(215)	247(112)	13745(6161,17497)
	PMS-selection	0(0)	2(10)	25(19,28)	0(0)	1(8)	23(18,27)
	PMS-group	5(34)	22(58)	30(30,30)	1(0)	10(30)	30(30,31)

ing all features into 357 groups and assigning features in $\{\tilde{\mathcal{G}}_k\}_{k=1}^3$ into the first three groups. We considered three combinations of sample sizes and dimensions $(n, p) = (100, 10000), (100, 20000), (200, 20000)$. The screening accuracy is summarized in Table 1, indicating that PMS-selection and PMS-group outperformed all other methods in terms of all criteria. We report the computation times of the different methods with varied sample sizes and dimensions in Figure 2, where both PMS methods are comparable to SIS and HOLP, but much faster than PartS and CIS.

5.2. Scalar-on-image regression

Here, we apply PMS to a scalar-on-image regression model. Two different scenarios were considered. The first scenario is a real data-based simulation study. We applied the local functional connectivity density (LFCD) in the ABIDE data set as covariates. The sample size is $n = 441$, and the dimension of all features is 38,547, which can be partitioned into 90 different regions based on the AAL partition criterion. We assumed that regions "Postcentral_L," "Precuneus_R," and "Cuneus_L" were selected regions, with number of active features 99, 107, and 90, respectively. We generated the effects of region "Postcentral_L" from

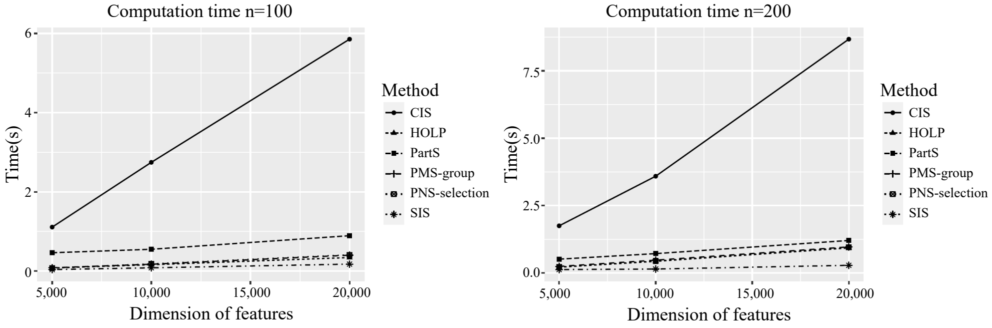

Figure 2. Comparisons of computational times with varied sample sizes and dimensions. The computations were performed on a Macbook Pro with a 2.3 GHz Quad-Core CPU and 16 GB memory.

Table 2. Screening accuracy based on the simulated data from the ABIDE study.

	$R^2 = 0.5$			$R^2 = 0.9$		
Method	FPR	FNR	Model Size	FPR	FNR	Model Size
SIS	798(43)	735(15)	37860(37307,38396)	814(16)	727(5)	37729(37535,37974)
HOLP	809(26)	944(17)	38475(38391,38515)	800(26)	933(19)	38459(38378,38510)
CIS	836(3)	958(16)	38471(38470,38473)	816(13)	893(19)	38473(38471,38475)
PartS	800(26)	930(18)	38364(38277,38423)	826(22)	971(9)	38398(38315,38427)
PMS	7(0)	2(2)	1331(620,9112)	7(0)	2(2)	1370(619,7120)

uniform distribution U$[4,5]$, those of region "Precuneus_R" from U$[-1,-0.5]$, and those of region "Cuneus_L" from U$[0,0.5]$. We chose the union set of a small neighborhood of each feature as a preselected feature set, and finished the variable screening procedure using the proposed PMS method. The screening accuracy results for the various methods are summarized in Table 2.

To further understand the performance of the different methods, we conducted another simulation study for the scalar-on-image regression. We generated p-dimensional features from Gaussian processes on equally spaced grids $\{\mathbf{s}_j\}_{j=1}^{p}$ in $[-1,1]^2$. The covariance function of $\mathbf{x}$ takes the form $\mathrm{Cov}(x_i, x_j) = \exp(-0.5\|\mathbf{s}_i - \mathbf{s}_j\|^2)$, for $i, j \in \{1,\ldots,p\}, i \neq j$. We assumed that the true signal is concentrated at one circle and one equal-sized triangle, as shown in Case II in Figure 4. The true signal from the circle region is generated from U$[-0.5,0]$, and the true signal in the triangle region followes U$[0,0.5]$. The true size is 217. To apply the proposed PMS method, we introduced the sparsity spatial structure $\mathbf{\Lambda}_1(i,j) = \exp(-\rho\|\mathbf{s}_i - \mathbf{s}_j\|_2) I\{\|\mathbf{s}_i - \mathbf{s}_j\|_2 \leq 0.4\}$ to the prior covariance matrix, with $\rho = 0.3$.

Table 3. Screening accuracy for a scalar-on-image regression with simulated data from different activation shapes.

	$R^2 = 0.5$			$R^2 = 0.9$		
Method	FPR	FNR	Model Size	FPR	FNR	Model Size
SIS	345(53)	951(76)	4529(4020,5104)	313(23)	985(38)	4394(4153,4651)
HOLP	312(67)	951(82)	4091(3718,4817)	269(37)	919(45)	3785(3548,4075)
CIS	784(9)	994(28)	8313(8270,8377)	786(5)	1(0)	8343(8315,8372)
PartS	752(0)	976(11)	8029(8028,8029)	752(0)	976(11)	8029(8028,8029)
PMS-spatial	189(62)	189(134)	3080(2483,3766)	137(23)	135(97)	2356(2103,2652)
PMS-group	24(114)	16(87)	268(263,270)	10(57)	7(59)	269(266,271)

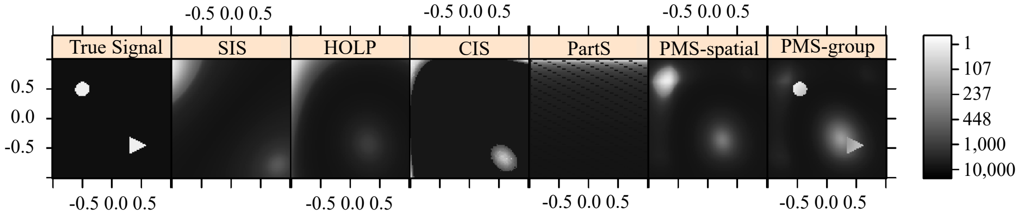

Figure 3. Rankings of screening statistics obtained from different methods for a scalar-on-image regression with signal-to-noise ratio $R^2 = 0.1$.

To implement PMS with prior group information, that is, PMS-group, we designed a circle region and a triangle region, each of which covered 125% of the corresponding true region. See Case III in Figure 4. We considered all features in the union of the two regions as one group, and the others as another group. We also considered PMS incorporating only a prior covariance matrix without selection and group information, as "PMS-spatial." All results with $(n, p) = (200, 10000)$ are summarized in Table 3. To provide more intuitive comparisons of the methods, in Figure 3, we also visualize the rankings of all imaging predictors based on the screening statistics.

5.3. Sensitivity analysis

We performed a sensitivity analysis for PMS. The true signal is the same as that in Section 5.2. We varied the prior specifications and evaluated the screening accuracy. For the prior covariance matrix $\mathbf{\Lambda}$, we attempted the norms $\mathbf{\Lambda}_1$ and $\mathbf{\Lambda}_2(i,j) = \exp(-\rho\|\mathbf{s}_i - \mathbf{s}_j\|_2^2)I\{\|\mathbf{s}_i - \mathbf{s}_j\|_2 \leq 0.4\}$. We also considered $\mathbf{\Lambda}$ with different values of ρ. Because the results were quite similar with each other, only results for $\rho = 0.3$ are reported. For prior selection $\mathcal{S}$, we designed three cases, summarized in Figure 4. According to the prior selection set $\mathcal{S}$ and true active feature set $\mathcal{M}_0$, all features were partitioned into four categories:

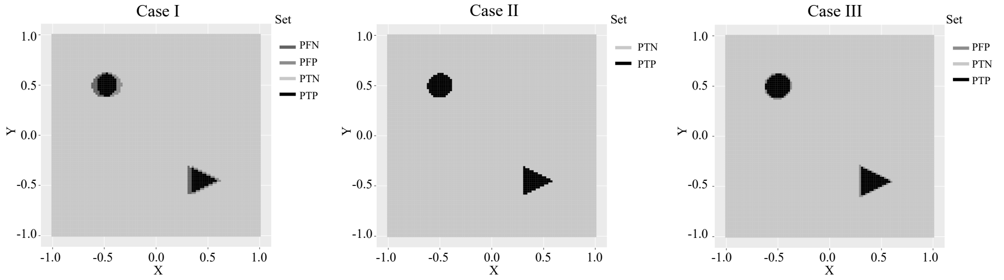

Figure 4. Prior selection regions and the true signal regions for the three cases in the sensitivity analysis.

Table 4. Sensitivity analysis on prior specifications for PMS.

			$R^2 = 0.5$			$R^2 = 0.9$		
$\mathbf{\Lambda}$	Method	$\mathcal{S}$	FPR	FNR	Model Size	FPR	FNR	Model Size
$\mathbf{\Lambda}_1$	PMS-spatial	N/A	189(62)	189(134)	3080(2483,3766)	137(23)	135(97)	2356(2103,2652)
	PMS-selection	I	20(4)	122(8)	3018(2608,3625)	19(1)	125(2)	3090(2902,3349)
		II	0(0)	1(2)	217(217,217)	0(0)	1(2)	217(217,270)
		III	5(0)	1(2)	272(272,272)	5(0)	2(4)	272(272,1180)
	PMS-group	I	76(166)	155(85)	5533(4021,6970)	29(64)	131(51)	4933(3788,6602)
		II	0(0)	0(0)	217(217,217)	0(0)	0(0)	217(217,217)
		III	24(114)	16(87)	268(263,270)	10(57)	7(59)	269(266,271)
$\mathbf{\Lambda}_2$	PMS-spatial	N/A	188(63)	188(135)	3068(2479,3762)	134(23)	134(99)	2351(2097,2642)
	PMS-selection	I	20(6)	122(10)	3061(2773,3378)	19(1)	125(3)	3049(2911,3216)
		II	0(0)	1(2)	217(217,217)	0(0)	1(2)	217(217,217)
		III	5(0)	1(3)	272(272,272)	5(1)	2(4)	272(272,1286)
	PMS-group	I	76(166)	155(84)	5515(4006,6943)	29(65)	130(51)	4916(3771,6581)
		II	0(0)	0(0)	217(217,217)	0(0)	0(0)	217(217,217)
		III	24(114)	16(87)	270(265,272)	10(57)	7(59)	269(266,271)

prior true positive (PTP), prior true negative (PTN), prior false positive (PFP), and prior false negative (PFN). In Case I, the prior selection set has both false positives and false negatives, where the PTP rate is $|\mathcal{M}_0 \cap \mathcal{S}|/|\mathcal{M}_0| \approx 75\%$, and the PFP rate is $|\mathcal{M}_0^c \cap \mathcal{S}|/|\mathcal{M}_0| \approx 25\%$. Case II is the ideal case in which the prior information is consistent with the true signal, that is, $\mathcal{M}_0 = \mathcal{S}$. In Case III, $\mathcal{M}_0 \subset \mathcal{S}$, with $|\mathcal{M}_0^c \cap \mathcal{S}|/|\mathcal{M}_0| \approx 25\%$. With the prior selected set, we used the method in Section 5.2 to specify prior group information for the PMS-group. The sensitivity analysis results are summarized in Table 4. “PMS-spatial” refers to PMS incorporating prior spatial covariance, without using prior selection and prior group information.

The results shown in the Tables 1- 4 indicate that the PMS method performs well under all settings; the advantages become more obvious when the signal is weak, which is a relatively difficult scenario. Under the case that the prior selected set contains all active features, PMS can select all important variables with very small model size, as well as nearly zero false positive and false negative rates. Furthermore, PMS is robust to prior mis-specification, as mentioned in Section 2. Even though the prior selection information is inconsistent with the true signals, PMS still performs very well. Moreover, even when the block structure for the covariance is misspecified, PMS still outperforms the other methods. In contrast, the performance of PartS highly depends on the accuracy of the prior partition structure, as shown in Figure 3 and the additional results in the Supplementary Material. From the sensitivity analysis, the PMS results are robust to a mild change of the prior knowledge. These results indicate that incorporating appropriate prior knowledge can substantially improve screening accuracy.

Some additional simulation studies are reported in the Supplementary Material, including a linear regression with a compound symmetry covariance matrix, applications of random decoupling in thresholding, and the CPMS results.

6. Data Application

We applied the proposed PMS method to resting-state fMRI (R-fMRI) data from the Autism Brain Imaging Data Exchange (ABIDE) Study (Di Martino et al. (2014)). The fMRI measures the blood oxygen level signal linked to neural activities, whereas the R-fMRI measures brain activity in a resting state. The ABIDE study comprises 20 resting-state functional magnetic resonance data sets from 17 experiment sites. The human brain is registered into the 3 mm standard Montreal Neurological Institute space composed of 38,547 voxels, which can be partitioned into 90 regions according to the AAL brain atlas. Removing all individuals with missing values, there are 441 healthy subjects. For each subject, the R-fMRI signal is recorded for each voxel over some time points. The intelligence quotient (IQ) and other demographic information, such as age and gender, are also collected.

Our main question of interest was to identify brain regions that are highly associated with IQ for healthy individuals, adjusted for age and gender. To select active imaging biomarkers for IQ prediction, we compared two types of imaging measures derived from the R-fMRI data: the fractional amplitude of low-frequency fluctuations (fALFF) and the local functional connectivity density (LFCD). In particular, the fALFF measures the spontaneous fluctuations in the fMRI signal intensity and reflects local brain activity. The LFCD mapping finds

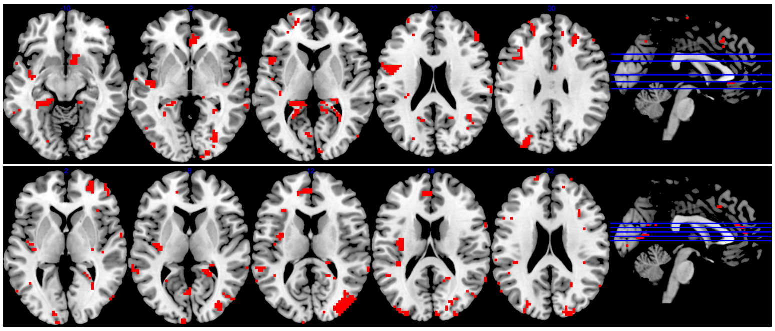

Figure 5. Selected 1,000 features that are shown on five axial slices.

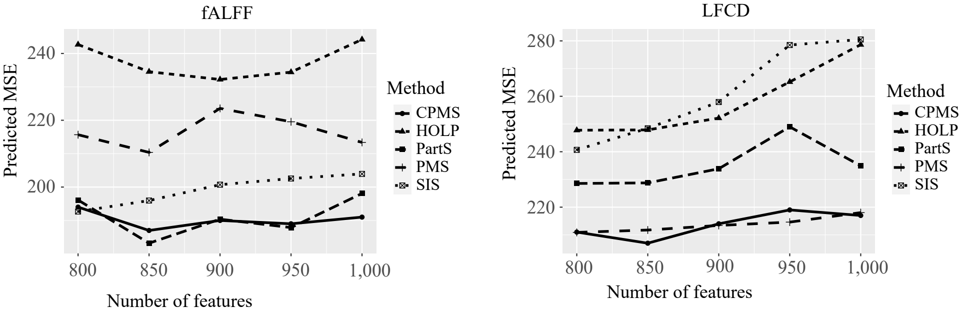

Figure 6. Predicted MSE for different methods in the fALFF and the LFCD data set.

the given neighbors and neighbors' neighbors until the edges become weaker than a given threshold value.

Using the AAL partition criterion, we constructed a block diagonal structure for the prior correlation matrix $\boldsymbol{\Lambda}$. Each block corresponds to one region with a sparsity spatial structure $\exp(-0.5\|\boldsymbol{s}_i - \boldsymbol{s}_j\|_2^2)$, where $\mathbf{s}_j$ represents the three-dimensional standardized coordinate of voxel j, for $j = 1, \ldots, 38547$. To add prior selected information, we chose the features in the brain regions in Table 7 of Li et al. (2009) as preselected features. For further comparison, we also applied CPMS, which was obtained by combining all prior selected information provided by SIS, HOLP, and PartS, and the preselected features. Because all parameter estimations by the CIS method are zero for these two data sets, we did not consider this method in this study. Table 5 summarizes the regions with more than 40 voxels when choosing the first 1,000 features by CPMS. The corresponding selected features are also shown on five axial slices in Figure 5.

We adopted 10-fold cross-validation to compare the performance of the different methods in terms of IQ prediction. We randomly split all data into 10

Table 5. Automated anatomical labeling of regions selected in the fALFF and the LFCD data set with more than 40 voxels selected by the CPMS selection method when choosing the first 1,000 features.

Dataset	Selected region	Voxel counts	Median rank	Selected region	Voxel counts	Median rank
fALFF	Frontal_Mid_R	72	454	Parietal_Inf_L	48	299
	Precentral_L	69	405	ParaHippocampal_L	45	360
	Frontal_Mid_L	63	551	Occipital_Mid_L	42	565
	Temporal_Sup_L	49	336	Fusiform_R	42	429
	Supp_Motor_Area_R	48	140			
LFCD	Frontal_Mid_R	88	526	Frontal_Inf_Tri_R	45	341
	Occipital_Mid_R	60	474	Temporal_Mid_R	41	601
	Insula_L	59	203	Occipital_Sup_R	40	803
	Precuneus_R	52	337			

subsets with approximately equal size. Each time, we chose one subset as the testing data, and the others as the training data. For the PartS method, we used the AAL partition criterion as the group partition. Predictions were performed using a ridge regression; the corresponding results are summarized in Figure 6. Figure 6 shows that the prior knowledge provided by Li et al. (2009) seems not to be consistent with the fALFF measure case; thus, the predicted MSE of PMS is relatively higher than those of PartS and SIS. However, we improved the prediction accuracy by using the CPMS method. Using all four sources of prior information simultaneously, the predicted MSE decreases significantly. In addition, under the LFCD measure, the predicted MSE of PMS is significantly smaller than those of other methods, indicating that PMS can select important features with higher accuracy if the prior knowledge is reasonable.

7. Discussion

We have proposed a prior knowledge-guided variable screening method for the linear regression model. We gave constructions of the proposed screening statistics under specific applications, and demonstrated the theoretical properties of PMS. We tested the performance of our method using a series of simulation studies, and applied it to the analysis of the ABIDE data. Being applicable to the linear regression model, PMS can be extended to the framework of generalized linear models. In recent years, variable screening methods based on model-free frameworks have been widely studied. Exploring an efficient way of incorporating prior knowledge into the variable screening procedure under the model-free setting is also an interesting topic, which we will explore in future work.

Supplementary Material

The online Supplementary Material includes Proposition 1 with its proof, the conditions and proofs of Theorems 1–4, and some additional simulation studies.

Acknowledgments

The authors thank the associate editor and two referees for their constructive and valuable comments. This research was partially supported by NIH grants R01DA048993, R01MH105561, R01GM124061, and R01CA249096.

References

Barut, E., Fan, J. and Verhasselt, A. (2016). Conditional sure independence screening. *Journal of the American Statistical Association* **111**, 1266–1277.

Bhattacharya, A., Chakraborty, A. and Mallick, B. K. (2016). Fast sampling with gaussian scale mixture priors in high-dimensional regression. *Biometrika* **103**, 985–991.

Bühlmann, P. and van de Geer, S. (2011). *Statistics for High-Dimensional Data: Methods, Theory and Applications*. Springer-Verlag, New York and Berlin.

Cho, H. and Fryzlewicz, P. (2012). High dimensional variable selection via tilting. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **74**, 593–622.

Cui, H., Li, R. and Zhong, W. (2015). Model-free feature screening for ultrahigh dimensional discriminant analysis. *Journal of the American Statistical Association* **110**, 630–41.

Di Martino, A., Yan, C.-G., Li, Q., Denio, E., Castellanos, F. X., Alaerts, K. et al. (2014). The autism brain imaging data exchange: towards a large-scale evaluation of the intrinsic brain architecture in autism. *Molecular Psychiatry* **19**, 659–667.

Fan, J. and Fan, Y. (2008). High dimensional classification using features annealed independence rules. *Annals of Statistics* **36**, 2605–2637.

Fan, J., Feng, Y. and Song, R. (2011). Nonparametric independence screening in sparse ultra-high dimensional additive models. *Journal of the American Statistical Association* **106**, 544–557.

Fan, J. and Lv, J. (2008). Sure independence screening for ultrahigh dimensional feature space. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **70**, 849–911.

Fan, J., Samworth, R. and Wu, Y. (2009). Ultrahigh dimensional feature selection: Beyond the linear model. *Journal of Machine Learning Research* **10**, 2013–2038.

Fan, J. and Song, R. (2010). Sure independence screening in generalized linear models with NP-dimensionality. *The Annals of Statistics* **38**, 3567–3604.

Goldsmith, J., Huang, L. and Crainiceanu, C. M. (2014). Smooth scalar-on-image regression via spatial bayesian variable selection. *Journal of Computational and Graphical Statistics* **23**, 46–64.

Gorst-Rasmussen, A. and Scheike, T. (2013). Independent screening for single-index hazard rate models with ultrahigh dimensional features. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **75**, 217–245.

Hall, P. and Miller, H. (2009). Using generalized correlation to effect variable selection in very high dimensional problems. *Journal of Computational and Graphical Statistics* **18**, 533–550.

He, K., Kang, J., Hong, H. G., Zhu, J., Li, Y., Lin, H. et al. (2019). Covariance-insured screening. *Computational Statistics & Data Analysis* **132**, 100–114.

He, X., Wang, L. and Hong, H. G. (2013). Quantile-adaptive model-free variable screening for high-dimensional heterogeneous data. *The Annals of Statistics* **41**, 342–369.

Hong, H. G., Kang, J. and Li, Y. (2018). Conditional screening for ultra-high dimensional covariates with survival outcomes. *Lifetime Data Analysis* **24**, 45–71.

Huettel, S. A., Song, A. W. and McCarthy, G. (2004). *Functional Magnetic Resonance Imaging.* Sinauer Associates, Sunderland.

Ishwaran, H. and Rao, J. S. (2005). Spike and slab variable selection: frequentist and Bayesian strategies. *The Annals of Statistics* **33**, 730–773.

Kang, J., Hong, H. G. and Li, Y. (2017). Partition-based ultrahigh-dimensional variable screening. *Biometrika* **104**, 785–800.

Kang, J., Reich, B. J. and Staicu, A.-M. (2018). Scalar-on-image regression via the soft-thresholded gaussian process. *Biometrika* **105**, 165–184.

Li, F., Zhang, T., Wang, Q., Gonzalez, M. Z., Maresh, E. L., Coan, J. A. et al. (2015). Spatial bayesian variable selection and grouping for high-dimensional scalar-on-image regression. *The Annals of Applied Statistics* **9**, 687–713.

Li, G., Peng, H., Zhang, J. and Zhu, L. (2012). Robust rank correlation based screening. *The Annals of Statistics* **40**, 1846–1877.

Li, R., Zhong, W. and Zhu, L. (2012). Feature screening via distance correlation learning. *Journal of the American Statistical Association* **107**, 1129–1139.

Li, Y., Liu, Y., Li, J., Qin, W., Li, K., Yu, C. et al.(2009). Brain anatomical network and intelligence. *PLoS Computational Biology* **5**, e1000395.

Ma, S., Li, R. and Tsai, C.-L. (2017). Variable screening via quantile partial correlation. *Journal of the American Statistical Association* **112**, 650–663.

Pan, W., Wang, X., Zhang, H., Zhu, H. and Zhu, J. (2020). Ball covariance: A generic measure of dependence in banach space. *Journal of the American Statistical Association.*

Rossell, D. and Telesca, D. (2017). Nonlocal priors for high-dimensional estimation. *Journal of the American Statistical Association* **112**, 254–265.

Smith, M. and Fahrmeir, L. (2007). Spatial Bayesian variable selection with application to functional magnetic resonance imaging. *Journal of the American Statistical Association* **102**, 417–431.

Tzourio-Mazoyer, N., Landeau, B., Papathanassiou, D., Crivello, F., Etard, O., Delcroix, N. et al. (2002). Automated anatomical labeling of activations in spm using a macroscopic anatomical parcellation of the mni mri single-subject brain. *Neuroimage* **15**, 273–289.

Wang, X. and Leng, C. (2016). High dimensional ordinary least squares projection for screening variables. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **78**, 589–611.

Zhao, S. D. and Li, Y. (2012). Principled sure independence screening for Cox models with ultra-high-dimensional covariates. *Journal of Multivariate Analysis* **105**, 397–411.

Zhu, L., Xu, K., Li, R. and Zhong, W. (2017). Projection correlation between two random vectors. *Biometrika* **104**, 829–843.

Zhu, L.-P., Li, L., Li, R. and Zhu, L.-X. (2011). Model-free feature screening for ultrahigh-dimensional data. *Journal of the American Statistical Association* **106**, 1464–1475.

Zou, H. and Hastie, T. (2005). Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **67**, 301–320.

Jie He
Department of Biostatistics, University of Michigan, Ann Arbor, MI 48109-2029, USA.
E-mail: jiehe@umich.edu
Jian Kang
Department of Biostatistics, University of Michigan, Ann Arbor, MI 48109-2029, USA.
E-mail: jiankang@umich.edu

(Received January 2020; accepted March 2021)