<table>
<thead>
<tr>
<th><strong>Statistica Sinica Preprint No:</strong> SS-2020-0427</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Title</strong></td>
</tr>
<tr>
<td><strong>Manuscript ID</strong></td>
</tr>
<tr>
<td><strong>URL</strong></td>
</tr>
<tr>
<td><strong>DOI</strong></td>
</tr>
<tr>
<td><strong>Complete List of Authors</strong></td>
</tr>
<tr>
<td><strong>Corresponding Author</strong></td>
</tr>
<tr>
<td><strong>E-mail</strong></td>
</tr>
</tbody>
</table>

Notice: Accepted version subject to English editing.
Prior Knowledge Guided Ultra-high Dimensional Variable Screening with Application to Neuroimaging Data

Jie He and Jian Kang

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 the linear regression model. 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 high-dimensional ordinary least-square projections (HOLP). We establish the screening consistency property for PMS with different types of prior knowledge. We show that PMS is robust to prior misspecifications; and when the prior knowledge provides correct information on summarizing the true parameter settings, PMS can substantially improve the selection accuracy compared to HOLP and other existing methods. We illustrate our method with 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 with the number of variables much outweighing 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 of variable screening is sure independence screening (SIS) (Fan and Lv 2008), which has been extended to generalized linear models (Fan and Fan 2008, Fan et al. 2009, Fan and Song 2010), generalized additive models (Fan et al. 2012), quantile regression (He et al. 2013, Ma et al. 2017) and the proportional hazards model (Zhao and Li 2012, Gorst-Rasmussen and Scheike 2013). By extending screening criteria that are solely based on marginal correlations between the outcome and predictors, a variety of statistics accounting for dependence between predictors have been proposed to improve screening accuracy and robustness (Li et al. 2012a, Zhu et al. 2011, Cho and Fryzlewicz 2012, Hall and Miller 2009). In particular, 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 as-
sumptions, variable screening has also been generalized to model-free cases which means that corresponding screening statistics are applicable to a general model class without specific expression (Zhu et al. 2011, Li et al. 2012b, Cui et al. 2015, Zhu et al. 2017, Pan et al. 2019).

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 that are related to brain functions (Huettel et al. 2004, Smith and Fahrmeir 2007). There has been a growing interest in selecting important voxels with strong fMRI signals in order to identify brain regions that are highly associated with certain brain functions behaviors or psychiatric disorders. The standard brain template for fMRI images contains 200,000 spatially contiguous voxels, which can be partitioned into different groups in light of the brain anatomy. In addition, existing studies may have identified some locations or voxels where the brain activity are strongly associated with the response variable of our interest, such as cognitive behaviors or disease status. This poses an interesting question on how to incorporate such prior information—prior importance knowledge as well as prior spatial structure, into variable screening methods. Some variable screening methods have been developed to address this questions. One pioneer work is the conditional sure indepen-
dence screening (CSIS) \cite{Barut2016, Hong2018}, which directly includes the predetermined important features into the model when screening variables. Similarly, the partition-based screening (PartS) method \cite{Kang2017} incorporates the spatial-guided partition structure into generalized linear models and performs variable screening by dividing all covariates into different groups, while the covariance-insured screening (CIS) \cite{He2019} applies the prior information through inter-feature dependence. However, all these methods are developed from the frequentist perspective and only focus on incorporating one type of prior information. As straightforward approaches to integrate prior knowledge, Bayesian variable selection methods have been developed for neuroimaging applications \cite{Smith2007, Goldsmith2014, Li2015, Kang2018}. Those methods often aim to make fully Bayesian inference on model selection, i.e., selecting important predictor variables into the model, and estimate the posterior distribution of parameters along with the posterior inclusion probability of each predictor variable. Thus, they require large computational costs for high-dimensional problems, and many existing methods are not feasible for the ultra-high dimensional problem. In contrast, the variable screening methods aims to screen out the predictor variables that are not strongly associated with the response variable to achieve the efficient dimension reduction with less computational
burden. There is a need of developing Bayesian variable screening methods to systematically incorporate prior knowledge and structural information in science.

Figure 1: A simulated example for the distributions of the screening statistics by HOLP and 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 $j = 4, \ldots, 10,000$. PMS incorporates the prior selection knowledge for $\beta_1$ and $\beta_3$ and thus obtains a better screening statistics. PMS also can improve the screening statistics for $\beta_2$ compared to HOLP.

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 coefficients which can incorporate the prior information according to specific applications. Many prior models are available for linear regression, for example, the spike-slab
priors (Ishwaran et al. 2005), the non-local priors (Rossell and Telesca 2017) and the global–local shrinkage priors (Bhattacharya et al. 2016), for which the efficient posterior computation methods are available for fully Bayesian inferences on 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 the useful knowledge into the screening procedure through the normal priors. To illustrate our idea, we consider a simple example by simulating data from a linear regression model with 100 samples and 10,000 predictors being generated 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 $j = 4, \ldots, 10,000$. Suppose we have the 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 based on the empirical Bayes method. In this case, we obtain the closed form of the posterior mean of the regression coefficients, please refer to Section 3.1 for more details. That is equal
to our proposed posterior mean screening (PMS) statistics. We repeat the experiments 500 times and obtain the distributions of each screening statistics by PMS and HOLP in Figure 1. For the active predictors 1 and 3 of which the prior selection information is incorporated, the screening statistics by PMS are clearly away from zero while those by HOLP are concentrated at zero. For the active predictor 2 of which the prior selection information is not incorporated, both PMS and HOLP screening statistics are much smaller than those for predictors 1 and 3, however, the PMS screening statistic is still clearly larger than zero with almost one probability while the lower tail of the HOLP screening statistic touches zero. This result clearly indicates that PMS is slightly better. For all other predictors \((j = 4, \ldots, 10,000)\), the distributions of both PMS and HOLP screening statistics are concentrated around zero (See the third column of Figure 1 for the mixtures of all the distributions), but the PMS screening statistics has a smaller variance. In summary, the normal prior can effectively incorporate prior selection knowledge and improve the screening accuracy. In this paper, we will primarily discuss how to incorporate two types of prior knowledge: the prior selection and the prior group structure. With non-informative prior specification, PMS is equivalent to HOLP \cite{Wang and Leng2016}. We study the theoretical foundations of the proposed method. We discuss the technical conditions of formulations of 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 manuscript is organized as follows. In Section 2, we propose the unified framework of Bayesian feature screening for linear regression model and establish the theoretical properties. In Section 3, we develop constructions of the prior mean and covariance information under some specific cases with their theoretical properties. In Section 4, we propose the PMS variable screening based ensemble learning to combine different types of prior knowledge. In Section 5, we evaluate the performance of the proposed method via a series of simulation studies. We apply the PMS method to the analysis of neuroimaging data in Section 6. A brief summary and some discussions are given in Section 7. The Supplementary Material contains all technical proofs.

2. Posterior Mean Variable Screening

2.1 Notation and Model Specification

Let $\mathbb{R}^d$ be the $d$-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 $\text{Sym}_+^d$ the space of $d \times d$ symmetric positive definite matrices with identity matrix $\mathbf{I}_d$. 
2.2 Posterior Mean Screening Statistics

\( N_d(\mu, \Sigma) \) represents a \( d \)-dimensional normal distribution with mean \( \mu \in \mathbb{R}^d \) and covariance matrix \( \Sigma \in \text{Sym}_+^d \). 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 \( \mathcal{A} \). To sequence of number \( x_n \) and \( y_n \), \( x_n = o(y_n) \) implies that \( \lim_{n \to \infty} x_n/y_n = 0 \), while \( x_n = O(y_n) \) denotes that \( x_n/y_n \) is bounded.

Suppose the dataset includes \( n \) observations of an outcome along with \( p \) predictors. We consider the linear regression model

\[
Y = X\beta + \epsilon,
\]

where \( Y = (y_1, \ldots, y_n)^T \in \mathbb{R}^n \) denotes the outcome variable, \( X = (x_1, \ldots, x_p) \in \mathbb{R}^{n \times p} \) is a design matrix with rank \( \min\{n, p\} \), \( \epsilon \in \mathbb{R}^n \) is independent and identically distributed errors with marginal distribution \( N(0, \sigma^2) \), and \( \sigma^2 \) is an unknown nuisance parameter. In addition, \( \beta = (\beta_1, \ldots, \beta_p)^T \in \mathbb{R}^p \) is a vector of coefficients for predictors \( \{x_j\}_{j=1}^p \).

2.2 Posterior Mean Screening Statistics

We assign a multivariate normal prior to \( \beta \):

\[
\beta \sim N_p(\mu, \tau^2 \Lambda),
\]

Statistica Sinica: Newly accepted Paper
(accepted author-version subject to English editing)
where \( \mathbf{\mu} = (\mu_1, \ldots, \mu_p)^T \in \mathbb{R}^p \), \( \mathbf{\Lambda} = (\lambda_{j,k}) \in \text{Sym}_+^p \) with \( \text{Var}(\beta_j) = \tau^2 \lambda_{j,j} \) and \( \text{Cov}(\beta_j, \beta_k) = \tau^2 \lambda_{j,k} \). It is obvious that parameter \( \tau^2 > 0 \) controls the overall prior variability of \( \mathbf{\beta} \). As prior (2.2) is a conjugate prior, the posterior distribution of \( \mathbf{\beta} \) given the other parameters is also a multivariate normal distribution: \( (\mathbf{\beta} | \mathbf{\mu}, \mathbf{\Lambda}, \bullet) \sim N_p(\nu, \mathbf{K}) \), where \( \bullet = \{\sigma^2, \tau^2, \mathbf{X}, \mathbf{y}\} \), \( \nu = (\theta \mathbf{\Lambda}^{-1} + \mathbf{X}^T \mathbf{X})^{-1}(\theta \mathbf{\Lambda}^{-1} \mathbf{\mu} + \mathbf{X}^T \mathbf{y}) \) and \( \mathbf{K} = (\theta \mathbf{\Lambda}^{-1} + \mathbf{X}^T \mathbf{X})^{-1} \sigma^2 \). Here \( \theta = \sigma^2/\tau^2 > 0 \) reflects the precision of the prior knowledge on the structure of predictors.

We propose to use posterior mean \( \bar{\nu} \) as the variable screening statistics, which incorporates prior knowledge \( \mathbf{\mu} \) and \( \mathbf{\Lambda} \). For simplicity, we abbreviate this method as PMS in the rest of this paper, which is represented as:

\[
\hat{\beta}^{\text{PMS}} = (\theta \mathbf{\Lambda}^{-1} + \mathbf{X}^T \mathbf{X})^{-1}(\theta \mathbf{\Lambda}^{-1} \mathbf{\mu} + \mathbf{X}^T \mathbf{y}). \tag{2.3}
\]

However, directly compute 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 statistics is equivalent to:

\[
\hat{\beta}^{\text{PMS}} = \mathbf{\mu} + \mathbf{\Lambda} \mathbf{X}^T \mathbf{\Omega}(\mathbf{Y} - \mathbf{X} \mathbf{\mu}), \tag{2.4}
\]

where \( \mathbf{\Omega} = (\mathbf{X} \mathbf{\Lambda} \mathbf{X}^T + \theta \mathbf{I}_n)^{-1} \). As the inverse of \( \mathbf{\Omega} \) 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 HOLP when \( \mathbf{\mu} = \mathbf{0}_p \), \( \mathbf{\Lambda} = \mathbf{I}_p \) and \( \theta = 0 \). Thus HOLP is a special case of PMS with uninformative prior.
2.2 Posterior Mean Screening Statistics

In application, we compute the PMS statistics by a fast algorithm, which includes the following three steps: (1). Compute $\tilde{X} = X\Lambda$; (2). Compute $b = \Omega(y - X\mu)$ by solving $(\tilde{X}X^T + \theta I_n)b = (y - X\mu)$; (3). Compute $\hat{\beta}_{PMS} = \mu + \tilde{X}^T b$. Then we obtain output $\hat{\beta}_{PMS}$. Given a thresholding parameter $\alpha$, the selected index set can be expressed as: $\hat{M}_\alpha = \{j : |\hat{\beta}_{PMS}^j| > \alpha\}$.

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

2.3 Screening Consistency

In this section, we establish the theoretical properties of the proposed PMS variable screening method. Denote $\beta = (\beta_1, \ldots, \beta_p)^T$ 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, notation $x_n = o_p(y_n)$ indicates that $x_n/y_n$ converges to 0 in probability as $n \to \infty$, while $x_n = O_p(y_n)$ 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 with those introduced by HOLP (Wang and Leng 2016) and are listed in the Supplementary Material. We list a few other conditions that are related to prior specifications.

A4. For some $c_7 > 0$ and $\gamma \geq 0$, $\max_{j \in \{1, \ldots, p\}} \left| e_j^T \Lambda^{-1} (\mu - \beta) \right| \leq \frac{c_7 n^{\gamma}}{p}$.

A5. To matrix $\Lambda$, let $\lambda_{ij}$ be the element in the $i$th row and the $j$th column of matrix $\Lambda$, then there exists $\nu > 0$, such that $\max_{i \in \{1, \ldots, p\}} \lambda_{ii} = O(n^\nu)$, $\sum_{j=1}^p |\lambda_{ij}| = O\left(\frac{n^{-2\tau_3-2\tau_4-2\gamma}}{\sqrt{\log n}}\right)$, $j \neq i; i, j = 1, \ldots, p$. For some $c_8 > 1$, $\min_{j \in \mathcal{M}_0} |\beta_j| \geq \frac{c_8 n^{1-(\tau_3+\tau_4)-\gamma+\nu}}{p}$. In addition, denote the $i$th row of
matrix $\Lambda^{1/2}$ as $\tilde{\lambda}_i = (\tilde{\lambda}_{i1}, \ldots, \tilde{\lambda}_{ip})^T, i = 1, \ldots, p$, then $\sum_{u=1}^p \sum_{v \neq u} |\tilde{\lambda}_{iu}\tilde{\lambda}_{jv}| = O(n^\nu)$, $\sum_{u=1}^p |\tilde{\lambda}_{iu}\tilde{\lambda}_{ju}| = O\left(\frac{n^{1 - (\tau_3 + \tau_4) - 2\gamma + \nu}}{\sqrt{\log n}}\right)$, for $j \neq i; i, j = 1, \ldots, p$.

Here $p > cn^{1+\nu}$ for some $c > 1$.

A6. Assume $\log p = o\left[\min\left\{\frac{n^{1-\xi_1}}{\log n}, q\left(\frac{\bar{C}n^{1/2-\xi_2}}{\sqrt{\log n}}\right)\right\}\right]$, for some $0 < \xi_1 < 1$,

$0 < \xi_2 < 1/2$ and $\bar{C} > 0$ and there exist $\alpha_n$, such that $\frac{\sqrt{p}n^{\alpha_n}}{n^{1 - (\tau_3 + \tau_4) - \gamma + \nu}} \to 0$

and $\frac{\alpha_n \sqrt{p} \log n}{n^{1 - (\tau_3 + \tau_4) - \gamma + \nu}} \to \infty$.

Condition A4 provides an upper bound on the difference between the prior mean parameter and the true values of regression coefficients. This condition implies that PMS may still enjoy screening consistency even when the prior mean slightly deviates from the true settings. Condition A5 imposes some upper bound constraints on the diagonal elements and off-diagonal elements of $\Lambda$, which are related to the prior variance and prior correlations of the predictor effects $\beta$, respectively. This condition implies to ensure the screening consistency both prior variance and correlation cannot increase too fast as the sample size $n$ increases. This condition can be straightforwardly verified when $\Lambda$ is an identity matrix where the predictor effects are assumed to be a prior independent. When $\Lambda$ 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 upper bounds of the bandwidth or the block size, providing in-
2.4 Choice of Thresholding Parameters

sights on prior specifications in practice. For example, in the scalar-on-image regression for neuroimaging application, the sparse block diagonal covariance structures can be adopted in light of the brain functional and anatomical region partitions. Condition A5 provides some insights on upper bounds of the number of brain regions, the number of voxels within regions, as well as the maximum correlation within each region in the order of the sample size.

**Theorem 1** (Screening Consistency). Under Conditions A1-A6, we have

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

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. HOLP has a similar result while the order of the convergence rate is different from that of PMS.

2.4 Choice of Thresholding Parameters

The thresholding parameter \(\alpha\) is critical to the performance of the variable screening procedure. Overestimating \(\alpha\) inflates the false positive rate, while underestimating \(\alpha\) hinders sure screening. We adopt the random decoupling
method (Barut et al. 2016) to select the thresholding parameter $\alpha$.

To ensure the stability of $\alpha$, we replicate the decoupling procedure for $K$ times. Given data $\{(x_i, Y_i)\}_{i=1}^n$, we allow $(1 - \tau_r)100\%$ proportion ($\tau_r \in [0, 1]$) of inactive variables could be included into model under the case that $X$ and $Y$ are not related, which corresponds to the null model. We randomly permute the rows of design matrix $X$ and obtain the pseudo data $\{(\tilde{x}_i^{(k)}, Y_i)\}_{i=1}^n$ with $\tilde{x}_i^{(k)} = x_{\pi_k(i)}$ and $\pi_k(i)$ be a permutation of index $i$, $i = 1, \ldots, n; k = 1, \ldots, K$. Based on the above expression, we could obtain the values of PMS statistics $\{|\hat{\beta}^\text{PMS}_{j(k)}|, j = 1, \ldots, p; k = 1, \ldots, K\}$, and $\alpha^*_k$ be the $\tau_r$-quantile of $\{|\hat{\beta}^\text{PMS}_{j(k)}|, j = 1, \ldots, p\}$. Finally, we choose $\alpha^* = \max_{1 \leq k \leq K} \alpha^*_k$ as the thresholding parameter.

3. Incorporating Prior Knowledge

When the specific information on $\mu$ and $\Lambda$ is available, we can straightforwardly carry out PMS by computing $\hat{\beta}^\text{PMS}$ based on (2.4). However, in many cases, the prior knowledge is relatively vague, then it is not trivial to specify the values of $\mu$ and $\Lambda$. In this section, we discuss in details on 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 the prior knowledge on which features should be selected into the PMS framework. For instance, in a functional neuroimaging study, we are interested in selecting the important brain locations that are highly associated with intelligence quotient (IQ). According to some existing studies, we may know some brain regions are more likely to be selected than 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 $S$ the indices of prior-selected features and $q = |S|$ the number of elements in $S$. Assume that $\mu_j = 0$ for feature $j \notin S$ and $\mu_S = (\mu_j : j \in S)$. Introducing a hyper-prior normal distribution to $\mu_S$, then the sampling distribution of $Y$ given the prior-selected features is $N_n(X_S \mu_S, \Omega_S^{-1})$, where $X_S = (x_j, j \in S)$, $\Omega_S = (X_S \Lambda_S X_S^T + \theta I_n)^{-1}$ and $\Lambda_S$ is the prior correlation matrix of coefficients for features in $S$. When $q \leq n$, we can assign $\mu_S$ the uninformative hyper-prior, i.e., $\pi(\mu_S) \propto 1$. When $q > n$, we assign $\mu_S$ a normal prior, i.e., $\mu_S \sim N(0_q, \tau^2 I_q)$ for $\tau^2 > 0$. We construct the PMS selection statistics, noted as $\hat{\beta}_S^{\text{PMS}}$, in Proposition 2.

**Proposition 2.** The PMS selection statistic is

$$
\hat{\beta}_S^{\text{PMS}} = \tilde{\mu} + \Lambda X^T \Omega (Y - X \tilde{\mu}),
$$

(3.5)
where \( \tilde{\mu} = (\tilde{\mu}_1, \ldots, \tilde{\mu}_p)^T \) with \( \tilde{\mu}_j = 0 \) for \( j \notin S \) and \( (\tilde{\mu}_j : j \in S) = \tilde{\mu}_S \) with

\[
\tilde{\mu}_S = \mathbb{E}(\mu_S | \Lambda, \bullet) = \begin{cases} 
(X_S^T \Omega_S X_S)^{-1} X_S^T \Omega_S Y & q \leq n \\
(X_S^T \Omega_S X_S + \tilde{\tau}^{-2} I_q)^{-1} X_S^T \Omega_S Y & q > n.
\end{cases}
\]

Given \( \Omega \), the complexity of computing \( \hat{\beta}^{\text{PMS}}_S \) 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 \( \hat{\beta}^{\text{PMS}}_S \). Let \( S_3 = S^c \cap M_0 \) be a set composed of all non-selected features. The following is an essential regularization condition for screening consistency of \( \hat{\beta}^{\text{PMS}}_S \).

**B1.** There exists a constant \( \bar{\gamma}_1 > 0 \), such that \( \|X_{S_3} \beta_{S_3}\| = O(n^{\bar{\gamma}_1}) \), where \( X_{S_3} = (X_j, j \in S_3) \) and \( \beta_{S_3} = (\beta_j, j \in S_3) \).

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 some restrictions on eigenvalues of matrix \( \Lambda_S \), while B3 makes some assumptions on the structure of a sub-matrix of \( \Lambda^{-1} \), a matrix composed of elements in \( \Lambda^{-1} \) whose rows and columns are in \( S \).

**Theorem 2** (Screening Consistency for \( \hat{\beta}^{\text{PMS}}_S \)). Under Conditions B1-B3, when \( q \leq n \), the PMS statistic \( (3.5) \) enjoys the screening consistency property.
3.2 Priors on Group-level Importance

When \( q > n \), the additional assumptions on \( \tilde{\tau}^2 \) are needed to ensure the screening consistency. For example, we may assume \( \tilde{\tau}^2 \to 0 \) as \( n \to \infty \), then \( \tilde{\mu} \to 0 \) with probability one. In this case, the prior selection information vanishes when the sample size increases, but we can verify Condition A4 by making other mild assumptions and use Theorem 1 to establish the screening consistency. For more general and weaker assumptions, the rigorous proof is non-trivial, thus we leave it to further 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 more details in 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 is hard to distinguish. For example, in the analysis of brain imaging data, the whole brain regions 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 2mm\(^3\) contains more than 180,000 voxels in AAL 116 regions. The goal is to select the voxel-level features by incorporating the region-level information.
3.2 Priors on Group-level Importance

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 $B = (b_{g,j})$ be a $m \times p$ group indicator matrix with $b_{g,j} \in \{0, 1\}$, where $b_{g,j} = 1$ indicates that feature $j$ belongs to group $g$, while $b_{g,j} = 0$, otherwise. Note that the column sum of $B$ is equal to one, i.e. $1^T_mB = 1_p$, 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 $\mu = (\bar{\mu}_1, \ldots, \bar{\mu}_m)^T$. We assume that within group features have the same levels of importance, i.e. $\mu = B^T\bar{\mu}$. Then the sampling distribution of $Y$ given $X$ and the prior selected group structure is $N_n(XB^T\bar{\mu}, \Omega_B^{-1})$, where $\Omega_B = (XB^T\Lambda_BXB + \theta I_n)^{-1}$ and $\Lambda_B$ is the prior correlation matrix for $B\beta$. Under this setting, we obtain the PMS group statistics ($\hat{\beta}_B^{PMS}$) from the following Proposition 3.

**Proposition 3.** Under the non-informative prior assumption $\pi(\bar{\mu}) \propto 1$, the posterior mean of $\bar{\mu}$ is $\hat{\mu} = E(\bar{\mu} | \Lambda, \bullet) = (BX^T\Omega_BXB)^{-1}BX^T\Omega_BY$. Then the PMS group statistics can be expressed as

$$\hat{\beta}_B^{PMS} = B^T\hat{\mu} + \Lambda X^T\Omega(Y - XB^T\bar{\mu}). \quad (3.6)$$

To establish the screening consistency for $\hat{\beta}_B^{PMS}$, we impose a couple of
regularity conditions. In particular, we need the following condition.

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

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

**Theorem 3** (Screening Consistency for $\hat{\beta}_\mathbf{B}^{\text{PMS}}$). Under Conditions C1-C4, the PMS group statistics (3.6) enjoys the screening consistency property.

Theorems 1, 2 and 3 indicate that our proposed PMS method is robust to the prior knowledge. Even though the prior knowledge is not exactly correct, as long as it is not too far away from the truth, screening results are still consistent. We also demonstrate this property via 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 others. We propose to tackle this issue by combining the PMS screening (CPMS), the screening statistics which inte-
grates prior knowledge from different sources simultaneously. Assume that we have \( K (K < \infty) \) types of prior knowledge, and denote by \( \hat{\beta}^{(1)}, \ldots, \hat{\beta}^{(K)} \) the corresponding PMS statistics, respectively. We introduce each entry of the CPMS statistics \( \hat{\beta}_{\text{CPMS}} \) as \( \hat{\beta}_{\text{CPMS}}^j = \max\{|\hat{\beta}^{(1)}_j|, \ldots, |\hat{\beta}^{(K)}_j|\}, j = 1, \ldots, p \).

Given a thresholding parameter \( \alpha \), the selected feature set can be expressed as \( \tilde{\mathcal{M}}_\alpha = \{j : \hat{\beta}_{\text{CPMS}}^j > \alpha\} \). For CPMS, the thresholding parameter \( \alpha \) could also be selected by the random decoupling method. Similar with 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 \( \alpha_n^K, C' \) and \( \tilde{C}' > 0 \), such that

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

to some \( 0 < \xi_1 < 1 \) and \( 0 < \xi_2 < 1/2 \).

5. **Simulation Studies**

We compared the performance of the proposed PMS method with existing variable screening methods, such as the SIS, HOLP, CIS and PartS methods, through a series of simulation studies. We assigned two different settings to a \( p \)-dimensional linear regression model \([2.1]\): group structure and image regression case. All simulation results were based on 200 replicates and signal-
to-noise ratio $R^2 = 0.5$ and $0.9$. We evaluate the screening accuracy using three criteria: the false positive rates (FPR) when the power of detecting the true signals is $80\%$, the false negative rates (FNR) while the FPR is controlled at $10\%$, and the median number of variables that are needed to include all true signals (Model Size). In all the tables, we report the FPR and the FNR multiplied by $1000$ with the standard deviations in the parentheses, and the model size with the corresponding $25\%$ and $75\%$ quantile intervals in the parentheses. The R package PMS (https://github.com/kangjian2016/PMS.git) is available.

5.1 Group Structure Case

Similar to a scenario in Zou and Hastie (2005), we introduced a group structure to all covariates in this setting. The specific distribution information was summarized as $x_{j+3m} = z_j + N(0, \delta^2)$, where $z_j \sim N(0,1), j = 1,2,3$ were independent, $\delta^2 = 0.01$ and $m = 0,\ldots,4$. In addition, $x_j \sim N(0,\delta^2), j = 16,\ldots,p$ were independent. Thus, the first 15 features were divided into 3 groups denoted as $G_0^k = \{x_{k+3m}, m = 0,\ldots,4\}, k = 1,2,3$ and the regression coefficients were set as $\beta_{G_0^1} = 0.5$, $\beta_{G_0^2} = 3$, $\beta_{G_0^3} = 5$ and $\beta_j = 0, j = 16,\ldots,p$.

Assuming we have another group series $G_k = \{x_{k+3m} : m = 0,\ldots,9\}, k = 1,2,3$. Obviously, half of elements in $G_k$ were active features while the others were inactive. For each $G_k$, we exchanged two active features with another two
5.1 Group Structure Case

Table 1: Screening accuracy for predictors with group structure.

<table>
<thead>
<tr>
<th>(n, p)</th>
<th>Method</th>
<th>$R^2 = 0.5$</th>
<th></th>
<th>$R^2 = 0.9$</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>FPR</td>
<td>FNR</td>
<td>Model Size</td>
<td>FPR</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>(100, 10, 000)</td>
<td>SIS</td>
<td>430(286)</td>
<td>294(128)</td>
<td>5108(2270,7931)</td>
<td>429(300)</td>
</tr>
<tr>
<td></td>
<td>HOLP</td>
<td>434(283)</td>
<td>298(120)</td>
<td>5043(2550,8003)</td>
<td>430(298)</td>
</tr>
<tr>
<td></td>
<td>CIS</td>
<td>38(6)</td>
<td>0(0)</td>
<td>412(385,429)</td>
<td>24(7)</td>
</tr>
<tr>
<td></td>
<td>PartS-I</td>
<td>296(236)</td>
<td>273(110)</td>
<td>8209(5809,9421)</td>
<td>242(206)</td>
</tr>
<tr>
<td></td>
<td>PartS-II</td>
<td>345(208)</td>
<td>314(106)</td>
<td>8212(5917,9422)</td>
<td>330(0.216)</td>
</tr>
<tr>
<td></td>
<td>PMS-selection</td>
<td>0(0)</td>
<td>0(0)</td>
<td>24(20,28)</td>
<td>0(0)</td>
</tr>
<tr>
<td></td>
<td>PMS-group</td>
<td>11(58)</td>
<td>22(63)</td>
<td>30(29,30)</td>
<td>2(13)</td>
</tr>
<tr>
<td>(100, 20, 000)</td>
<td>SIS</td>
<td>442(280)</td>
<td>287(120)</td>
<td>10813(5088,16144)</td>
<td>421(291)</td>
</tr>
<tr>
<td></td>
<td>HOLP</td>
<td>370(285)</td>
<td>256(133)</td>
<td>8074(3082,14809)</td>
<td>424(287)</td>
</tr>
<tr>
<td></td>
<td>CIS</td>
<td>22(3)</td>
<td>0(0)</td>
<td>458(434,478)</td>
<td>15(3)</td>
</tr>
<tr>
<td></td>
<td>PartS-I</td>
<td>290(186)</td>
<td>328(97)</td>
<td>17266(14564,18544)</td>
<td>249(183)</td>
</tr>
<tr>
<td></td>
<td>PartS-II</td>
<td>347(161)</td>
<td>368(96)</td>
<td>17395(14569,18546)</td>
<td>310(177)</td>
</tr>
<tr>
<td></td>
<td>PMS-selection</td>
<td>0(0)</td>
<td>0(5)</td>
<td>24(20,28)</td>
<td>0(0)</td>
</tr>
<tr>
<td></td>
<td>PMS-group</td>
<td>6(38)</td>
<td>10(45)</td>
<td>30(29,30)</td>
<td>2(20)</td>
</tr>
<tr>
<td>(200, 20, 000)</td>
<td>SIS</td>
<td>370(285)</td>
<td>256(133)</td>
<td>8074(3082,14809)</td>
<td>305(284)</td>
</tr>
<tr>
<td></td>
<td>HOLP</td>
<td>295(213)</td>
<td>230(119)</td>
<td>15588(8152,18236)</td>
<td>183(201)</td>
</tr>
<tr>
<td></td>
<td>CIS</td>
<td>29(5)</td>
<td>0(0)</td>
<td>636(615,654)</td>
<td>13(7)</td>
</tr>
<tr>
<td></td>
<td>PartS-I</td>
<td>219(213)</td>
<td>230(119)</td>
<td>15588(8152,18236)</td>
<td>183(201)</td>
</tr>
<tr>
<td></td>
<td>PartS-II</td>
<td>305(224)</td>
<td>269(97)</td>
<td>15611(8159,18239)</td>
<td>249(215)</td>
</tr>
<tr>
<td></td>
<td>PMS-selection</td>
<td>0(0)</td>
<td>2(10)</td>
<td>25(19,28)</td>
<td>0(0)</td>
</tr>
<tr>
<td></td>
<td>PMS-group</td>
<td>5(34)</td>
<td>22(58)</td>
<td>30(30,30)</td>
<td>1(0)</td>
</tr>
</tbody>
</table>

We constructed the prior covariance information of PMS based on $\tilde{G}_k^{3=k}$. Letting $\tilde{G}_4 = \{x_j, j = 31, \ldots, p\}$ and $L$ be the normalized graph Laplacian matrix for $\{\tilde{G}_k^{1=k}\}$, we introduced a network structure to the prior covariance matrix, which took the form of $\Lambda = (L + \varepsilon I_p)^{-1}$ with $\varepsilon = 10^{-3}$. As $\tilde{G}_k^{3=k}$ were inconsistent with $G_k^{3=k}$, matrix $\Lambda$ was misspecified. For PMS with prior
5.1 Group Structure Case

selection, termed as “PMS-selection”, the pre-selected set was \( S = \bigcup_{k=1}^{3} \tilde{G}_k \).

For PMS with prior group information, termed as “PMS-group”, \( \{\tilde{G}_k\}^4_{k=1} \) was chosen as the group partition. We considered two PartS methods with different partition structures. One was based on the true group structure \( \{G^0_k\}^4_{k=1} \), termed as “PartS-I.” The other structure, termed as “PartS-II”, was obtained by randomly partitioning all features into 357 groups and assigned features in \( \{\tilde{G}_k\}^3_{k=1} \) 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 was summarized in Table[1], indicating that PMS-selection and PMS-group outperformed all other methods in terms of all the criteria.

We reported computation time of different methods with varied sample sizes and dimensions in Figure[2] where both PMS methods were comparable to SIS and HOLP, but much faster than PartS and CIS.

![Figure 2: Comparisons of computational time with varied sample sizes and dimensions. The computations were performed on Macbook Pro with 2.3 GHz Quad-Core CPU and 16 GB memory.](image-url)
5.2 Scalar-on-Image Regression

We applied PMS to scalar-on-image regression model. Three 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 be 99, 107 and 90. We generated effects of region “Postcentral.L” from uniform distribution $U[4,5]$, effects of region “Precuneus.R” from $U[-1,-0.5]$ and effects of region “Cuneus.L” from $U[0,0.5]$. We chose the union set of a small neighborhood of each feature as pre-selected feature set and finish the variable screening procedure by the proposed PMS method. The screening accuracy results for different methods were summarized in Table 2.

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

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

To further understand the performance of different methods, we conducted...
Table 3: Screening accuracy for scalar-on-image regression with the simulated data from different activation shapes.

<table>
<thead>
<tr>
<th>Method</th>
<th>FPR</th>
<th>FNR</th>
<th>Model Size</th>
<th>FPR</th>
<th>FNR</th>
<th>Model Size</th>
</tr>
</thead>
<tbody>
<tr>
<td>SIS</td>
<td>345(53)</td>
<td>951(76)</td>
<td>4529(4920,5104)</td>
<td>313(23)</td>
<td>985(38)</td>
<td>4394(4153,4651)</td>
</tr>
<tr>
<td>HOLP</td>
<td>312(67)</td>
<td>951(82)</td>
<td>4091(3718,4817)</td>
<td>269(37)</td>
<td>919(45)</td>
<td>3785(3548,4075)</td>
</tr>
<tr>
<td>CIS</td>
<td>784(9)</td>
<td>994(28)</td>
<td>8313(8270,8377)</td>
<td>786(5)</td>
<td>1(0)</td>
<td>8343(8315,8372)</td>
</tr>
<tr>
<td>PartS</td>
<td>752(0)</td>
<td>976(11)</td>
<td>8029(8028,8029)</td>
<td>752(0)</td>
<td>976(11)</td>
<td>8029(8028,8029)</td>
</tr>
<tr>
<td>PMS-spatial</td>
<td>189(62)</td>
<td>189(134)</td>
<td>3080(2483,3766)</td>
<td>137(23)</td>
<td>135(97)</td>
<td>2356(2103,2652)</td>
</tr>
<tr>
<td>PMS-group</td>
<td>24(114)</td>
<td>16(87)</td>
<td>268(263,270)</td>
<td>10(57)</td>
<td>7(59)</td>
<td>269(266,271)</td>
</tr>
</tbody>
</table>

Another simulation study for the scalar-on-image regression. We generated $p$-dimensional features from Gaussian processes on equally spaced grids $\{s_j\}_{j=1}^p$ in $[-1,1]^2$. The covariance function of $\mathbf{x}$ took the form of $\text{Cov}(x_i, x_j) = \exp(-0.5\|s_i - s_j\|^2)$, $i, j \in \{1, \ldots, p\}, i \neq j$. We assumed that the true signal was concentrated at one circle and one equal-sized triangle, as it was indicated in Case II in Figure 4. The true signal from the circle region were generated from $\text{U}[-0.5,0]$, while the true signal in the triangle region followed $\text{U}[0,0.5]$. The true size is 217. To apply the proposed PMS method, we introduced sparsity spatial structure $\Lambda_1(i,j) = \exp(-\rho\|s_i - s_j\|_2)I\{\|s_i - s_j\|_2 \leq 0.4\}$ to the prior covariance matrix with $\rho = 0.3$. To implement PMS with prior group information, i.e. PMS-group, we designed a circle region and a triangle region, each of which covered 125% of corresponding true region. See Case III in Figure 4. We considered all features in the union of the two regions as one group, while the others as another group. We also considered PMS only incor-
5.3 Sensitivity Analysis

We performed sensitivity analysis for PMS. The true signal was of the same with that in Section 5.2. We varied prior specifications and evaluated the screening accuracy. To the prior covariance matrix $\mathbf{\Lambda}$, on the one hand, we attempted different norms—both $\mathbf{\Lambda}_1$ and $\mathbf{\Lambda}_2(i,j) = \exp(-\rho\|\mathbf{s}_i - \mathbf{s}_j\|_2^2)I\{\|\mathbf{s}_i - \mathbf{s}_j\|_2 \leq 0.4\}$. On the other hand, we also considered $\mathbf{\Lambda}$ with different values of $\rho$, as the results were quite similar with each other, only results of $\rho = 0.3$ were reported. For prior selection $\mathcal{S}$, we designed three different cases that were 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: prior true positive (PTP), prior true negative (PTN), prior false positive (PFP) and prior false negative (PFN). In Case I, the prior selection set had both false positives and false negatives where the PTP rate was $\frac{|\mathcal{M}_0 \cap \mathcal{S}|}{|\mathcal{M}_0|} \approx 75\%$ and the PFP rate was $\frac{|\mathcal{M}_0^c \cap \mathcal{S}|}{|\mathcal{M}_0|} \approx 25\%$. Case II was an ideal case with prior information being consistent with the true signal, i.e., $\mathcal{M}_0 = \mathcal{S}$. In Case III, $\mathcal{M}_0 \subset \mathcal{S}$ with $\frac{|\mathcal{M}_0^c \cap \mathcal{S}|}{|\mathcal{M}_0|} \approx 25\%$. With the prior selected set, we used the same method in Section 5.2 to specify prior group information for PMS-group. All the sensitivity analysis results were summarized in Table 4. “PMS-spatial” refers to PMS incorporating prior spatial covariance without using prior selection and prior group information.

![Figure 4: Prior selection regions and the true signal regions for three cases in sensitivity analysis.](image)

Results listed in the Tables indicate that 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
5.3 Sensitivity Analysis

Table 4: Sensitivity analysis on prior specifications for PMS.

<table>
<thead>
<tr>
<th>Λ</th>
<th>Method</th>
<th>S</th>
<th>FPR</th>
<th>FNR</th>
<th>Model Size</th>
<th>FPR</th>
<th>FNR</th>
<th>Model Size</th>
</tr>
</thead>
<tbody>
<tr>
<td>A₁</td>
<td>PMS-spatial</td>
<td>N/A</td>
<td>189(62)</td>
<td>189(134)</td>
<td>3080(2483,3766)</td>
<td>137(23)</td>
<td>135(97)</td>
<td>2356(2103,2652)</td>
</tr>
<tr>
<td></td>
<td>PMS-selection</td>
<td>I</td>
<td>20(4)</td>
<td>122(8)</td>
<td>3018(2608,3625)</td>
<td>19(1)</td>
<td>125(2)</td>
<td>3090(2902,3349)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>II</td>
<td>0(0)</td>
<td>1(2)</td>
<td>217(217,217)</td>
<td>0(0)</td>
<td>1(2)</td>
<td>217(217,217)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>III</td>
<td>5(0)</td>
<td>1(2)</td>
<td>272(272,272)</td>
<td>5(0)</td>
<td>2(4)</td>
<td>272(272,1180)</td>
</tr>
<tr>
<td></td>
<td>PMS-group</td>
<td>I</td>
<td>76(166)</td>
<td>155(85)</td>
<td>5533(4021,6970)</td>
<td>29(64)</td>
<td>131(51)</td>
<td>4933(3788,6602)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>II</td>
<td>0(0)</td>
<td>0(0)</td>
<td>217(217,217)</td>
<td>0(0)</td>
<td>0(0)</td>
<td>217(217,217)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>III</td>
<td>24(114)</td>
<td>16(87)</td>
<td>268(263,270)</td>
<td>10(57)</td>
<td>7(59)</td>
<td>269(266,271)</td>
</tr>
<tr>
<td>A₂</td>
<td>PMS-spatial</td>
<td>N/A</td>
<td>188(63)</td>
<td>188(135)</td>
<td>3068(2479,3762)</td>
<td>134(23)</td>
<td>134(99)</td>
<td>2351(2097,2642)</td>
</tr>
<tr>
<td></td>
<td>PMS-selection</td>
<td>I</td>
<td>20(6)</td>
<td>122(10)</td>
<td>3061(2773,3378)</td>
<td>19(1)</td>
<td>125(3)</td>
<td>3049(2911,3216)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>II</td>
<td>0(0)</td>
<td>1(2)</td>
<td>217(217,217)</td>
<td>0(0)</td>
<td>1(2)</td>
<td>217(217,217)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>III</td>
<td>5(0)</td>
<td>1(3)</td>
<td>272(272,272)</td>
<td>5(1)</td>
<td>2(4)</td>
<td>272(272,1286)</td>
</tr>
<tr>
<td></td>
<td>PMS-group</td>
<td>I</td>
<td>76(166)</td>
<td>155(84)</td>
<td>5515(4006,6943)</td>
<td>29(65)</td>
<td>130(51)</td>
<td>4916(3771,6581)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>II</td>
<td>0(0)</td>
<td>0(0)</td>
<td>217(217,217)</td>
<td>0(0)</td>
<td>0(0)</td>
<td>217(217,217)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>III</td>
<td>24(114)</td>
<td>16(87)</td>
<td>270(265,272)</td>
<td>10(57)</td>
<td>7(59)</td>
<td>269(266,271)</td>
</tr>
</tbody>
</table>

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 also robust to prior mis-specification, as it is mentioned above in Section 2. Even though prior selection information was inconsistent with the true signals, PMS still performed very well. Moreover, even the block structure for the covariance is mis-specified, PMS still outperformed other methods. In contrast, the performance of PartS highly depended on the accuracy of the prior partition structure, as shown in Figure 3 and additional results in the Supplementary Material. From the sensitivity analysis, the PMS results were robust to mild change of the prior knowledge.
All these results indicate that incorporating appropriate prior knowledge can substantially improve screening accuracy.

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

6. Data Application

We applied the proposed PMS method to the resting-state functional magnetic resonance imaging data from the Autism Brain Imaging Data Exchange Study (Di Martino et al., 2014). The fMRI measures the blood oxygen level signal that is linked to the neural activities, while the resting–state fMRI measures the brain activity at resting state. The ABIDE study was composed of 20 resting–state functional magnetic resonance datasets from 17 experiment sites. The human brain was registered into the 3mm standard Montreal Neurological Institute space composed of 38,547 voxels, which can be partitioned into 90 regions according to the automated anatomical labeling (AAL) brain atlas. Removing all individuals with missing values, there was a total of 441 healthy subjects. For each subject, the R–fMRI signal was recorded for each voxel over some time points. The intelligence quotient (IQ) as well as some other demographic information such as age and gender were 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: fractional amplitude of low–frequency fluctuations (fALFF) and local functional connectivity density (LFCD). In particular, fALFF measures the spontaneous fluctuations in the fMRI signal intensity and reflects the local brain activity, while LFCD mapping finds the given neighbors and neighbors’ neighbors until edges become weaker than the given threshold value.

Figure 5: Selected 1000 features that are shown on five axial slices.

According to the AAL partition criterion, we constructed a block diagonal structure to the prior correlation matrix $\Lambda$. Each block corresponds to one region with a sparsity spatial structure $\exp(-0.5\|s_i - s_j\|_2^2)$, where $s_j$ represented the 3-dimensional standardized coordinate of voxel $j$, $j = 1, \ldots, 38547$. To add prior selected information, we chose features in brain regions in Ta-
ble 7 of Li et al. (2009) as pre-selected features. To make further comparison, we also applied CPMS, which was obtained by combing all prior selected information provided by SIS, HOLP, PartS and the pre-selected features. As all parameter estimations by the CIS method were zero to these two datasets, we did not consider this method in this study. Table 5 summarizes the regions with more than 40 voxels when choosing the first 1000 features by CPMS. The corresponding selected features are also shown on five axial slices in Figure 5.

![Graphs showing predicted MSE for different methods in the fALFF and the LFCD dataset.](image)

Figure 6: Predicted MSE for different methods in the fALFF and the LFCD dataset.

We also adopted 10-fold cross-validation to compare the performance of different methods on IQ prediction. We randomly split all data into 10 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 group partition. We made prediction by the ridge regression; and the corresponding results are summarized in Figure 6. From Figure 6, we found that the prior knowledge provided by Li et al. (2009) seems not consistent with the fALFF measure case; thus, the predicted MSE
of PMS is relatively higher than PartS and SIS. However, we improved the prediction accuracy by the CPMS method. Using all four sources of prior information simultaneously, the predicted MSE can be decreased to a large degree. In addition, under the LFCD measure, the predicted MSE of PMS is significantly smaller than other methods, indicating that PMS can select important features with higher accuracy if the prior knowledge is reasonable.

Table 5: Automated anatomical labelling regions selected in the fALFF and the LFCD dataset with more than 40 voxels selected by the CPMS selection method when choosing the first 1000 features.

<table>
<thead>
<tr>
<th>Dataset</th>
<th>Selected region</th>
<th>Voxel counts</th>
<th>Median rank</th>
<th>Voxel counts</th>
<th>Median rank</th>
</tr>
</thead>
<tbody>
<tr>
<td>fALFF</td>
<td>Frontal_Mid_R</td>
<td>72</td>
<td>454</td>
<td>Parietal_Inf_L</td>
<td>48</td>
</tr>
<tr>
<td></td>
<td>Precentral_L</td>
<td>69</td>
<td>405</td>
<td>ParaHippocampal_L</td>
<td>45</td>
</tr>
<tr>
<td></td>
<td>Frontal_Mid_L</td>
<td>63</td>
<td>551</td>
<td>Occipital_Mid_L</td>
<td>42</td>
</tr>
<tr>
<td></td>
<td>Temporal_Sup_L</td>
<td>49</td>
<td>336</td>
<td>Fusiform_R</td>
<td>42</td>
</tr>
<tr>
<td></td>
<td>Supp_Motor_Area_R</td>
<td>48</td>
<td>140</td>
<td></td>
<td></td>
</tr>
<tr>
<td>LFCD</td>
<td>Frontal_Mid_R</td>
<td>88</td>
<td>526</td>
<td>Frontal_Inf_Tri_R</td>
<td>45</td>
</tr>
<tr>
<td></td>
<td>Occipital_Mid_R</td>
<td>60</td>
<td>474</td>
<td>Temporal_Mid_R</td>
<td>41</td>
</tr>
<tr>
<td></td>
<td>Insula_L</td>
<td>59</td>
<td>203</td>
<td>Occipital_Sup_R</td>
<td>40</td>
</tr>
<tr>
<td></td>
<td>Precuneus_R</td>
<td>52</td>
<td>337</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

7. Discussion

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

Acknowledgement

The authors thank the Associate Editor and two referees sincerely for their constructive and valuable comments, which helped improve this paper substantially. This research was partially supported by the NIH grants R01DA048993, R01MH105561, R01GM124061 and R01CA249096.

References


Jie He
Department of Biostatistics, University of Michigan, Ann Arbor
E-mail: jiehe@umich.edu

Jian Kang
Department of Biostatistics, University of Michigan, Ann Arbor
E-mail: jiankang@umich.edu