Published in final edited form as:

Wiley Interdiscip Rev Comput Stat. 2019 ; 11(2): . doi:10.1002/wics.1454.

A selective overview of feature screening methods with applications to neuroimaging data

Kevin He[1], **Han Xu**[2], **Jian Kang**[1]

[1]Department of Biostatistics, University of Michigan, Ann Arbor, Michigan

[2]Department of Statistics, University of Michigan, Ann Arbor, Michigan

Abstract

In neuroimaging studies, regression models are frequently used to identify the association of the imaging features and clinical outcome, where the number of imaging features (e.g., hundreds of thousands of voxel-level predictors) much outweighs the number of subjects in the studies. Classical best subset selection or penalized variable selection methods that perform well for low- or moderate-dimensional data do not scale to ultrahigh-dimensional neuroimaging data. To reduce the dimensionality, variable screening has emerged as a powerful tool for feature selection in neuroimaging studies. We present a selective review of the recent developments in ultrahigh-dimensional variable screening, with a focus on their practical performance on the analysis of neuroimaging data with complex spatial correlation structures and high-dimensionality. We conduct extensive simulation studies to compare the performance on selection accuracy and computational costs between the different methods. We present analyses of resting-state functional magnetic resonance imaging data in the Autism Brain Imaging Data Exchange study.

This article is categorized under:

Applications of Computational Statistics > Computational and Molecular Biology

Statistical Learning and Exploratory Methods of the Data Sciences > Image Data Mining

Statistical and Graphical Methods of Data Analysis > Analysis of High Dimensional Data

Keywords

correlated covariates; imaging data analysis; linear regression; variable screening

1 | INTRODUCTION

Recent advances in neuroimaging technology have generated high-resolution brain imaging data that can measure brain functions and structures with increasing accuracy. This provides unprecedented opportunities for researchers to precisely identify the important brain regions

Correspondence: Jian Kang, Department of Biostatistics, University of Michigan, Ann Arbor, MI 48109. jiankang@umich.edu.

CONFLICT OF INTEREST
The authors have declared no conflicts of interest for this article.

RELATED WIREs ARTICLE
Some recent statistical learning methods for longitudinal high-dimensional data

that are strongly associated with certain clinical symptoms, which will have a great impact on public health and precision medicine. To this end, a class of regression models has been widely used, where the response variable is the clinical outcome of interest and the predictors include imaging features. We refer to this model as scalar-on-image regression. Performing variable selection in scalar-on-image regression directly identifies the brain regions of interest. However, in a typical neuroimaging study, the three-dimensional brain image may involve up to millions of voxels, while the number of subjects is usually in a range of hundreds to thousands. Thus, for voxel-level selection in the scalar-on-image regression, the number of predictors is often on the exponential order of the sample size. Due to the computational infeasibility, some classical variable selection methods, such as the best subset selection, are not directly applicable to this setting. Similarly, regularization-based variable selection methods have been extensively studied during the past decades, and have many successful applications in neuroimaging studies for regression problems with a moderately high dimensionality. Once again, for a regression model with ultrahigh-dimensional predictors, many of these methods still suffer limitations of numerical instability, poor reproducibility, and heavy computational costs.

To accommodate ultrahigh-dimensional data, variable screening methods have been proposed and widely used for many applications. The key difference between variable screening and variable selection lies in their slightly different objectives. The classical variable selection methods attempt to determine the subset of predictors that are strongly associated with the response variable. In contrast, variable screening seeks to exclude a large amount of predictors that are not associated with the response variable. In practice, we can use variable screening to reduce the dimensionality first; and then apply the classical variable selection on the reduced model. Thus, variable screening can be considered as a “preprocessing” step for variable selection. As the key advantages of variable screening, it is straightforward to implement parallel computation at much lower computational cost.

The pioneer work of variable screening is the sure independence screening (SIS) by (Fan & Lv, 2008). This approach ranks predictors according to their marginal utility; namely, each feature is used independently as a predictor to determine its usefulness for predicting the response. The success of SIS relies on a fundamental assumption that the true association between the individual predictors and the response can be inferred from their marginal associations. To account for the violation of this assumption, recent researches have expressed a growing interest in conducting multivariate screenings (Cho & Fryzlewicz, 2012; Cui, Li, & Zhong, 2015; Hall & Miller, 2009; He et al., 2018; Jin, Zhang, & Zhang, 2014; Kang, Hong, & Li, 2017; Li, Peng, Zhang, & Zhu, 2012; Wang & Leng, 2016; Zhu, Li, Li, & Zhu, 2011). In particular, according to the reports by the authors, the high-dimensional ordinary-least squares projection (HOLP; Wang & Leng, 2016) substantially improves variable screening accuracy of SIS for many scenarios with theoretical supports. The partition-based screening (PartS; Kang et al., 2017) can integrate prior grouping knowledge into the variable screening with solid theoretical foundation and it achieves a better performance in spatial variable screening with neuroimaging applications. Similarly, covariance-insured screening (CIS; He et al., 2018) can effectively take advantage of the sparse block diagonal covariance structure of the ultrahigh-dimensional predictors (if any) and produce more accurate variable screening.

Many of these variable screening methods were originally proposed for general purposes, and thus can be applied to perform imaging feature screening for scalar-on-image regression in neuroimaging studies. However, it is unclear whether the important assumptions of those methods are appropriate for neuroimaging applications and whether different methods are computationally feasible for ultrahigh-dimensional imaging predictors. One of the important characteristics of brain imaging data is the complex spatial correlation structure between voxels and regions. Neighboring voxels are likely to be highly correlated and some functionally associated brain regions have long-run corrections. Therefore, it is of great interest to learn how the complex correlation structure of the imaging predictors affects variable screening accuracy for the different methods. To address these questions, we present a selective review of computationally efficient screening methods for ultrahigh-dimensional data with an emphasis on neuroimaging applications. We design simulation studies to generate complex spatially correlated imaging data for different scenarios and assess the screening accuracy and computational costs.

The remaining article is organized as follows: In Section 2, we first provide some requisite notations and then present an overview of the recent developments in marginal and multivariate variable screening, respectively, identifying strengthens and limitations to carry out such analysis in practice. In Section 3, we conduct extensive simulations to examine the performance of the discussed methods. In Section 4, we illustrate the methods using a neuroimaging data set. We conclude the article with a discussion, and make some recommendations for neuroimaging studies.

2 | METHODS

In this section, we illustrate the screening methods using a linear model for continuous outcome, though the methods can be extended to accommodate binary, count, and survival data. Consider a multiple linear regression model with n independent samples, $\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$, where $\mathbf{y} = (Y_1, \ldots, Y_n)^T$ is the response vector, $\boldsymbol{\epsilon} = (\epsilon_1, \ldots, \epsilon_\mathrm{n})^T$ is a length-n vector of independently and identically distributed random errors, $\mathbf{X}$ is an $n \times p$ design matrix, and $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_\mathrm{p})^T$ is the coefficient vector. Write $\mathbf{X} = (\mathbf{X}_1, \ldots, \mathbf{X}_n)^T = (\mathbf{x}_1, \ldots, \mathbf{x}_p)$, where $\mathbf{X}_i$ is a p-dimension covariate vector for the ith subject and $\mathbf{x}_j$ is the jth column of the design matrix, $1 \le i \le n$, $1 \le j \le p$. In the scalar-on-image regression, the index j refers to imaging feature. Without loss of generality, we assume that each covariate $\mathbf{x}_j$ is standardized to have sample mean 0 and sample standard deviation 1. In addition, we assume that the response vector is centered with sample mean 0. For any set $\mathscr{D} \subset \{1, \ldots, p\}$, we define subvectors, $\mathbf{X}_{i, \mathscr{D}} = \left\{X_{i,j}: j \in \mathscr{D}\right\}$ and $\mathbf{x}_{\mathscr{D}} = \left\{\mathbf{x}_j: j \in \mathscr{D}\right\}$. Let $\mathbf{X}_{i,-j} = \{X_{i,1}, \ldots, X_{i,p}\} \backslash \{X_{i,j}\}$ and denote by $\boldsymbol{\Sigma} = \mathbb{C}ov(\mathbf{X}_i)$. Note that we do not assume any multivariate distribution on X_i. In many existing screening methods, the largest eigenvalue of the population covariance matrix Σ is allowed to diverge as n grows with certain rate. Specifically, there exist some constant $\tau \ge 0$ and $c > 0$ such that the largest eigenvalue $\lambda_{max}(\Sigma) \le cn^{\tau}$.

When $p \gg n$, $\boldsymbol{\beta}$ is difficult to estimate without the common sparsity condition that only a small number of variables contribute to the response. For improved model interpretability and accuracy of estimation, our overarching goal is to identify the active set

$$\mathcal{S}_0 = \left\{ j : \beta_j \neq 0,\, j = 1, \ldots, p \right\}.$$

2.1 | Sure independence screening

The SIS (Fan & Lv, 2008) is the simplest approach for ultrahigh-dimensional variable screening. For continuous outcomes, the SIS selects all variables having sufficiently large absolute values of marginal sample correlation with the response. For a threshold parameter $\gamma > 0$, the selection index set by SIS is

$$\widehat{\mathcal{S}}_{\mathrm{SIS}}(\gamma) = \left\{ j : |\widehat{\mathrm{Corr}}(\mathbf{y}, \mathbf{x}_j)| > \gamma \right\}.$$

The time complexity of SIS is only $O(np)$. Under certain assumptions on the covariance Σ, SIS achieves the screening property (Fan & Lv, 2008):

$$\Pr\left\{ \mathcal{S}_0 \subseteq \widehat{\mathcal{S}}_{\mathrm{SIS}}(\gamma) \right\} \to 1 \; as \; n \to \infty.$$

SIS has inspired much subsequent research. Extensions have been proposed to accommodate generalized linear models (Fan, Samworth, & Wu, 2009; Fan & Song, 2010) and Cox proportional hazard models (Hong, Kang, & Li, 2017; Zhao & Li, 2012, 2014). Semiparametric marginal screening methods were proposed for single-index hazard models, linear transformation models and general single-index models (Fan, Feng, & Song, 2011; Li et al., 2012; Zhu et al., 2011). Nonparametric marginal screening methods were studied for linear additive model and quantile regression (Fan et al., 2011; He, Wang, & Hong, 2013). Recently, conditional SIS methods, as an alternative to marginal screening approaches, have been developed for generalized linear models (Barut, Fan, & Maathuis, 2016) and Cox proportional hazard models (Hong et al., 2017) by preincluding a set of a priori important predictors.

2.2 | High-dimensional ordinary-least squares projection

The HOLP (Wang & Leng, 2016) projects response to the row spaces of the design matrix, which may preserve the ranks of regression coefficients. Specifically, this approach utilizes the generalized inverse of the design matrix and computes the HOLP estimator $\tilde{\boldsymbol{\beta}} = \left(\tilde{\beta}_1, \ldots, \tilde{\beta}_p\right)^T = \mathbf{X}^T\left(\mathbf{X}\mathbf{X}^T\right)^{-1}\mathbf{y}$. Then for threshold parameter $\gamma > 0$, the selection index set by HOLP is

$$\widehat{\mathcal{S}}_{\mathrm{HOLP}}(\gamma) = \left\{ j : \left|\tilde{\beta}_j\right| > \gamma \right\},$$

Clearly, HOLP is straightforward to implement and the time complexity is $O(n^2p)$. The HOLP procedure (Wang & Leng, 2016) enjoys good theoretical properties as well.

2.3 | Partition-based screening

The PartS approaches (Kang et al., 2017) are proposed to leverage the prior grouping information on predictors to improve the variable screening accuracy. Suppose the predictors can be partitioned into G disjoint groups in accordance with known information. Denote by g_j the group membership of variable $\mathbf{x}_j$. Let $\widetilde{\mathbf{X}}_g = \left\{\mathbf{x}_j, g_j = g\right\}$ be the collection of predictors in group g with the intercept, where $g \in \{1, \ldots, G\}$. For each g, the partition-based variable screening statistics are constructed through the linear regression fitting of $\mathbf{y}$ on $\widetilde{\mathbf{X}}_g$, which is $\widehat{\boldsymbol{\beta}}_g^T = \left(\widetilde{\mathbf{X}}_g^T \widetilde{\mathbf{X}}_g\right)^{-1} \widetilde{\mathbf{X}}_g^T \mathbf{y}$. Then pooling $\widehat{\boldsymbol{\beta}}_g$ together, we obtain $\cup_{g=1}^{G} \widehat{\boldsymbol{\beta}}_g = \left(\widehat{\beta}_1, \ldots, \widehat{\beta}_p\right)^T$. For a threshold parameter $\gamma > 0$, the selection index set by PartS is

$$\widehat{\mathcal{S}}_{\text{PartS}} = \left\{j : \left|\widehat{\beta}_j\right| > \gamma\right\},$$

When $G = O(pn^{-1/2})$, the time complexity of PartS is $O(n^{3/2}p)$ which is faster than HOLP but slower than SIS. The performance of PartS depends on the group partition $\{g_j\}$. In the scalar-on-image regression, the group information can be naturally determined by the spatial locations of voxels. When the prior knowledge is unavailable, random group partitions can be applied. When multiple partition information is available but it is unclear which information is better, a combining rule can be adopted (Kang et al., 2017) to integrate multiple PartS statistics. In particular, suppose we have K different partition based screening statistics, denoted $\widehat{\boldsymbol{\beta}}^{(k)} = \left(\widehat{\beta}_1^{(k)}, \ldots, \widehat{\beta}_p^{(k)}\right)^T$, for $k = 1, \ldots, \boldsymbol{K}$. The combined PartS selection index set is

$$\widehat{\mathcal{S}}_{\text{CombPartS}} = \left\{j : \max_{1 \le k \le K} |\widehat{\beta}_j^{(k)}| > \gamma\right\}.$$

2.4 | Covariance-insured screening

To incorporate the correlation information, He et al. (2018) proposed compartmentalizing covariates into blocks so that variables from distinct blocks are less correlated. The algorithm starts from the idea of thresholding. Consider $\widehat{\boldsymbol{\Sigma}}$ the sample estimate of $\boldsymbol{\Sigma}$. For a threshold $\delta > 0$, let $\widehat{\boldsymbol{\Sigma}}^{\delta}$ be the regularization of $\widehat{\boldsymbol{\Sigma}}$ such that

$$\widehat{\boldsymbol{\Sigma}}_{jk}^{\delta} = \widehat{\boldsymbol{\Sigma}}_{jk} 1\left\{|\widehat{\boldsymbol{\Sigma}}_{jk}| \ge \delta\right\}.$$

The CIS procedure then partition the vector $\boldsymbol{\beta}$ into blocks, $\widehat{\mathcal{S}}_1, \ldots, \widehat{\mathcal{S}}_G$, in a way such that all off-diagonal blocks of $\widehat{\boldsymbol{\Sigma}}^{\delta}$ are zero; for example,

$$|\widehat{\boldsymbol{\Sigma}}_{jk}^{\delta}| = 0 \, for \, all \; j \in \widehat{\mathcal{S}_g}, k \in \widehat{\mathcal{S}_{g'}}, g \ne g'.$$

Here $\widehat{\mathcal{S}}_1, \ldots, \widehat{\mathcal{S}}_G$ forms a partition of the p predictors:

$$\widehat{\mathcal{S}_g} \cap \widehat{\mathcal{S}_{g'}} = \varnothing \;\; for \, g \ne g', \;\; and \;\; \widehat{\mathcal{S}_1} \cup \widehat{\mathcal{S}_2} \ldots \cup \widehat{\mathcal{S}}_G = \{1, \ldots, p\}.$$

To determine the importance of predictors, the CIS proceed by computing the partial correlation, which is defined as follows:

Partial Correlation.—The partial correlation, $\rho(\boldsymbol{Y}_i, \mathbf{X}_{i,j}|\mathbf{X}_{i,-j})$, is defined as the correlation between the residuals resulting from the linear regression of $X_{i,j}$ on $\boldsymbol{X}_{i,-j}$ and Y_{i} on $\boldsymbol{X}_{i,-j}$

$$\rho\left(Y_i, X_{i,j}|\mathbf{X}_{i,-j}\right) = \frac{\mathbb{C}ov\left[Y_i - E\left(Y_i|\mathbf{X}_{i,-j}\right), X_{i,j} - E\left(X_{i,j}|\mathbf{X}_{i,-j}\right)\right]}{\left\{\mathbb{V}\mathrm{ar}\left(Y_i - E\left(Y_i|\mathbf{X}_{i,-j}\right)\right)\mathbb{V}\mathrm{ar}\left(X_{i,j} - E\left(X_{i,j}|\mathbf{X}_{i,-j}\right)\right)\right\}^{1/2}}.$$

The direct linkage between $\boldsymbol{\beta}$ and the partial correlations has been well established in the literature (see Peng, Wang, Zhou, & Zhu, 2009; Whittaker, 1990). The CIS approach can then be summarized as follows:

- Identify the disconnected blocks by thresholding the sample covariance matrix.
- Compute the block-wise sample partial correlations $\hat{\rho}\left(Y_i, X_{i,j}|\mathbf{X}_{i,\hat{\mathcal{S}}_g\backslash\{j\}}\right)$.
- Compute $\hat{\mathcal{S}}_{CIS} = \left\{j \in \hat{\mathcal{S}}_g, 1 \le g \le G: \left|\hat{\rho}\left(Y_i, X_{i,j}|\mathbf{X}_{i,\hat{\mathcal{S}}_g\backslash\{j\}}\right)\right| > \nu\right\}$, where ν is a predefined threshold.

3 | SIMULATION STUDY

We perform simulation studies to compare the performance of SIS, HOLP, CIS, and PartS for variable screening in scalar-onimage regression. The two-dimensional image predictors are simulated from Gaussian processes on equal space grid points in $[-1, 1]^2$. We vary the number of patients from $n = 500$ to 1,000, and the number of predictors from $p = 10{,}000$ to 40,000. The data and parameters are simulated under three scenarios:

1. *Setting 1.* There is only one region generated by one Gaussian process. We define $\mathcal{S} = \left\{s_j\right\}_{j=1}^{p}$ as a collection of equally spaced $\sqrt{p} \times \sqrt{p}$ grid points in $[-1, 1]^2$. The covariance structure of the covariates $\mathbf{X}$ is set to decay exponentially over space; that is, $\mathbb{C}ov\left(\mathbf{x}_j, \mathbf{x}_{j'}\right) = \exp\left(-0.1s_j^2 - 0.1s_{j'}^2 - 0.5\left(s_j - s_{j'}\right)^2\right)$ for any $j \ne j'$. Figure 1a (Setting 1) shows the graphic representation of the covariance structure. The parameters with nonzero effects are in a circle of the graph, with a radius 0.1 for the true parameters (see Figure 1b, Setting 1). The values of the nonzero effects follow a [0.5,1] uniform distribution.

2. *Setting 2.* There are 16 different regions with the same numbers of predictors within each region (see Figure 1a, Setting 2). The covariance structure within each region is the same as in setting 1. Different regions are correlated with correlation 0.9. The parameters with nonzero effects are in two equal-sized circles in the graph, with a radius 0.1. The locations of the two circles are randomly placed (see Figure 1b, Setting 2). The values of the two circles follow [0.5,1] and [−1,−0.5] uniform distribution, respectively.

3. *Setting 3*. In this scenario, each circle of nonzero parameters is at the center of a specific covariate region (see Figure 1b, Setting 3). All other set-ups are the same as Setting 2.

Given covariates and coefficients that are generated from each of the above settings, we generate the response y from linear regression. We set the variance of random errors such that the model-explained variance ratio $R^2 = 0.5$ or 0.9. We replicate our simulation 50 times and compare the selection performance. All results based on CIS and PartS are obtained by an *R* package. Sure independent screening is obtained by the *R* package SIS, while HOLP is implemented by the *R* package *screening*, Two criteria are reported in Table 1: the false-positive rates (FPR) when covering 80% of the truth, and the false-negative rates (FNR) while keeping false-positive rate = 0.1. Figure 2 shows the selection performances of each of the methods.

In Setting 1, for which all the predictors belong to one region, the SIS has the best performance among all the methods. The performances of HOLP and CIS are similar. In Settings 2 and 3, CIS is competitive and performs better than SIS when $R^2 = 0.9$, and provides comparable performance as HOLP for $R^2 = 0.5$. Interestingly, PartS, which utilizes the spatial information and partitions predictors into small groups, produces better results than CIS and HOLP for most of the scenarios with $R^2 = 0.5$.

4 | APPLICATION

We applied the aforementioned four methods: SIS, HOLP, PartS, and CIS to analyze the resting state fMRI data in the Autism Brain Imaging Data Exchange (ABIDE) study (Di Martino et al., 2014). The fMRI measures the blood oxygen level signal that is linked to the neural activities, whereas the resting-state fMRI only measures the brain activity at resting state without performing any task. The ABIDE study aggregated 20 R-fMRI datasets from 17 experiment sites involving a total of 1,112 subjects. For each subject, the R-fMRI signal was recorded for each voxel in the brain over multiple time points, and demographic information such as age, gender, and Intelligence quotient (IQ) were also collected. Several standard imaging preprocessing steps (Di Martino et al., 2014) including motion corrections, slice-timing correction, and spatial smoothing were performed. All the brains were registered into the 3 mm standard Montreal Neurological Institute (MNI) space consisting of 38,547 voxels in the 90 brain regions that are defined by the Automated Anatomical Labeling (Tzourio-Mazoyer et al., 2002) system. After removing the missing values, the complete datasets include 414 health subjects. Our analysis focused on identifying the important brain regions that are strongly associated with IQ among the health subjects. The IQ ranges from 73.0 to 146.0 with a mean of 111.3 and a standard deviation of 12.5. To select the imaging biomarkers for IQ prediction, we compare four types of the imaging statistics derived from the R-fMRI data: Fractional amplitude of low-frequency fluctuations (fALFF), Regional Homogeneity (ReHo), the weighted degree centrality (WDC), and local functional connectivity density (LFCD), where fALFF measures the spontaneous fluctuations in the fMRI signal intensity and reflects the local brain activity; ReHo evaluates the similarity or synchronization between the fMRI time series of a given voxel and its nearest neighbors; WDC is a measure of local brain network connectivity and identifies the most connected voxels by counting the number of direct connections (edges) to all other voxels; and LFCD mapping finds the given seed's neighbors and neighbor's neighbors until edges become weaker than the given threshold value.

To compare the performance of the different methods on IQ prediction using fALFF, WDC, ReHo, and LFCD, we adopted a 10-fold cross-validation approach. Specifically, we randomly split the data into 10 subsets with the approximately equal size. Each time we pick one subset as the testing dataset and consider the rest as the training dataset. We applied the four different variable screening methods to fit the training dataset and obtain a set of selected voxels, where for PartS we used AAL 90 regions as a group partition. Then we made a prediction on the IQ value in the testing dataset using linear models with the elastic net penalty (implemented by R package glmnet) and random forest models with 500 trees. We consider two measures to evaluate the prediction performance: predicted mean square error (PMSE) and predicted R2 (PR2). Figure 2 shows box plots of the PMSE and PR2 over 10-fold cross-validations for four different screening methods on four different imaging statistics using linear models and random forest models. Table 2 summaries the average values of those measures. For all the imaging statistics, there are no significant difference in PMSE among different measures. For ReHo and WDC, PartS and CIS achieve a better PR2 compared to other screening methods using both linear models and random forest models.

5 | DISCUSSION

Fast predictor screening methods for dimension reduction are crucial to the analysis of big neuroimaging data. A particularly attractive feature of SIS is that it is computationally fast. However, the validity of the the SIS hinges upon assumptions such as partial faithfulness, which is defined as follows:

Partial Faithfulness.

The distribution of ($\boldsymbol{X}_i$, $\boldsymbol{Y}_i$) is said to be ($\boldsymbol{X}_i$, $\boldsymbol{Y}_i$)-partially faithful if for every j ∈*{1, …, p}:*

$$\rho\left(Y_i, X_{i,j} \middle| \mathbf{X}_{i,\mathscr{C}}\right) = 0 \,\text{for some}\, \mathscr{C} \subseteq \left\{1, \dots, p\right\} \backslash \left\{j\right\} \Rightarrow \rho\left(Y_i, X_{i,j} \middle| \mathbf{X}_{i,-j}\right) = 0.$$

For instance, when $\mathscr{C}$ is empty, $\rho\left(Y_i, X_{ij} | \mathbf{X}_{i,C}\right) = \mathrm{Corr}\left(Y_i, X_{i,j}\right) = 0 \Rightarrow \beta_j = 0$.

While successful in many applications, this assumption can be violated. To illustrate, we examine a sample version of partial correlation

$$\hat{\rho}\left(Y_i, X_{i,j} \middle| \mathbf{X}_{i,\mathscr{C}}\right) = \frac{\mathbf{x}_j^T(\mathbf{I}_n - \Pi_{\mathscr{C}})\mathbf{y}}{\sqrt{\mathbf{x}_j^T(\mathbf{I}_n - \Pi_{\mathscr{C}})\mathbf{x}_j}\sqrt{\mathbf{y}^T(\mathbf{I}_n - \Pi_{\mathscr{C}})\mathbf{y}}},$$

where $\mathbf{I}_n$ is the identity matrix and $\Pi_{\mathscr{C}} = \mathbf{x}_{\mathscr{C}}\left(\mathbf{x}_{\mathscr{C}}^T\mathbf{x}_{\mathscr{C}}\right)^{-1}\mathbf{x}_{\mathscr{C}}^T$ is the projection matrix onto the space spanned by $\mathbf{x}_{\mathscr{C}}$. The numerator $\hat{\rho}\left(Y_i, X_{i,j} | \mathbf{X}_{i,c}\right)$ can be decomposed as

$$\mathbf{x}_j^T(\mathbf{I}_n - \Pi_{\mathscr{C}})\mathbf{y} = \beta_j \mathbf{x}_j^T(\mathbf{I}_n - \Pi_{\mathscr{C}})\mathbf{x}_j + \sum_{k \in S_0 \backslash (\mathscr{C} \cup \{j\})} \beta_k \mathbf{x}_j^T(\mathbf{I}_n - \Pi_{\mathscr{C}})\mathbf{x}_k + \mathbf{x}_j^T(\mathbf{I}_n - \Pi_{\mathscr{C}})\epsilon\,. \tag{1}$$

Equation (1) indicates that only when the last two terms in (1) are negligible compared with the first one, the partial faithfulness is valid. In practice, however, this assumption may be violated and the marginal effects can be quite different from the joint effects. As discussed by Fan and Lv (2008)) and Fan and Song (2010), the marginal screening methods do not consider the correlation between the predictors. Thus, they may select irrelevant variables that are highly correlated with important variables (false-positive) and fail to select relevant variables that are marginally unimportant but jointly informative (false-negative).

In contrast, the CIS leverages the group information including correlation by compartmentalizing covariates into correlated blocks. This approach may bypass the difficulty encountered in traditional multivariate screening procedures and render improved computational feasibility, better screening efficiency and weaker theoretical conditions. However, the computation burden of CIS increases when the maximal number of variables in the disconnected blocks is large. Moreover, the issue of collinearity among the predictors adds difficulty to the estimation of partial correlation, which may impact the performance of variable screening. Evidenced by our simulations as well as the analysis of the functional magnetic resonance imaging data, the partition based screening method provides a useful toolkit for variable screening. In particular, when the spatial information is available, we show that screening accuracy can be improved by using partition based screening compared with other screening methods such as SIS, high-dimensional ordinary least-squares projection and covariance-guided screening. Finally, as pointed out by one of the reviewers, the implementation of the CIS approach requires an estimate of the sample covariance matrix, which can be computationally challenging for brain image studies with ultrahigh-dimensional predictors. Alternatively, prior knowledge for brain regions can be incorporated to improve the estimation and computational efficiency. For example, suppose the predictors can be partitioned into disjoint groups in accordance with known information such as the spatial location. The sample covariance matrix can then be estimated for each group separately to further refine the identification of block structure.

REFERENCES

Barut E, Fan J, & Maathuis A (2016). Conditional sure independence screening. Journal of the American Statistical Association, 111(515), 1266–1277. [PubMed: 28360436]

Cho H, & Fryzlewicz P (2012). High dimensional variable selection via tilting. Journal of the Royal Statistical Society: Series B, 74(3), 593–622.

Cui H, Li R, & Zhong W (2015). Model-free feature screening for ultrahigh dimensional discriminant analysis. Journal of the American Statistical Association, 110, 630641.

Di Martino A, Yan C-G, Li Q, Denio E, Castellanos FX, Alaerts K, … Milham MP (2014). The autism brain imaging data exchange: Towards a large-scale evaluation of the intrinsic brain architecture in autism. Molecular Psychiatry, 19(6), 659–667. [PubMed: 23774715]

Fan J, Feng Y, & Song R (2011). Nonparametric independence screening in sparse ultrahigh-dimensional additive models. Journal of the American Statistical Association, 106(494), 544–557. [PubMed: 22279246]

Fan J, & Lv J (2008). Sure independence screening for ultrahigh dimensional feature space. Journal of the Royal Statistical Society: Series B, 70(5), 849–911.

Fan J, Samworth R, & Wu Y (2009). Ultrahigh dimensional feature selection: Beyond the linear model. Journal of Machine Learning Research, 10, 2013–2038. [PubMed: 21603590]

Fan J, & Song R (2010). Sure independence screening in generalized linear models and np-dimensionality. Annals of Statistics, 38(6), 3567–3604.

Hall P, & Miller R (2009). Using generalized correlation to effect variable selection in very high dimensional. Journal of Computational and Graphical Statistics, 18, 533550.

He K, Kang J, Hong HG, Zhu J, Li Y, Lin H, …, Li Y (2018). Covariance-insured screening. arXiv:1805.06595.

He X, Wang L, & Hong H (2013). Quantile-adaptive model-free variable screening for high-dimensional heterogeneous data. Annals of Statistics, 41(1), 342–369.

Hong G, Kang J, & Li Y (2017). Conditional screening for ultra-high dimensional covariates with survival outcomes. Lifetime Data Analysis, 24, 45–71.

Jin J, Zhang CH, & Zhang Q (2014). Optimality of graphlet screening in high dimensional variable selection. Journal of Machine Learning Research, 15, 2723–2772.

Kang J, Hong G, & Li Y (2017). Partition-based ultrahigh-dimensional variable screening. Biometrika, 104(4), 785–800. [PubMed: 29643546]

Li G, Peng H, Zhang J, & Zhu L (2012). Robust rank correlation based screening. Annals of Statistics, 40(3), 1846–1877.

Peng J, Wang P, Zhou N, & Zhu J (2009). Partial correlation estimation by joint sparse regression models. Journal of the American Statistical Association, 104(486), 735–746. [PubMed: 19881892]

Tzourio-Mazoyer N, Landeau B, Papathanassiou D, Crivello F, Etard O, Delcroix N, … Joliot M (2002). Automated anatomical labeling of activations in SPM using a macroscopic anatomical parcellation of the MNI MRI single-subject brain. NeuroImage, 15(1), 273{289.

Wang X, & Leng C (2016). High dimensional ordinary least squares projection for screening variables. Journal of the Royal Statistical Society: Series B, 78(3), 589–611.

Whittaker J (1990). Graphical models in applied multivariate statistics Wiley Series in Probability and Mathematical Statistics. Chichester: John Wiley & Sons.

Zhao DS, & Li Y (2012). Principled sure independence screening for cox models with ultra-high-dimensional covariates. Journal of Multivariate Analysis, 105(1), 397–411. [PubMed: 22408278]

Zhao DS, & Li Y (2014). Score test variable screening. Biometrics, 70(4), 862–871. [PubMed: 25124197]

Zhu L, Li L, Li R, & Zhu L (2011). Model-free feature screening for ultrahighdimensional data. Journal of the American Statistical Association, 106(496), 1464–1475. [PubMed: 22754050]

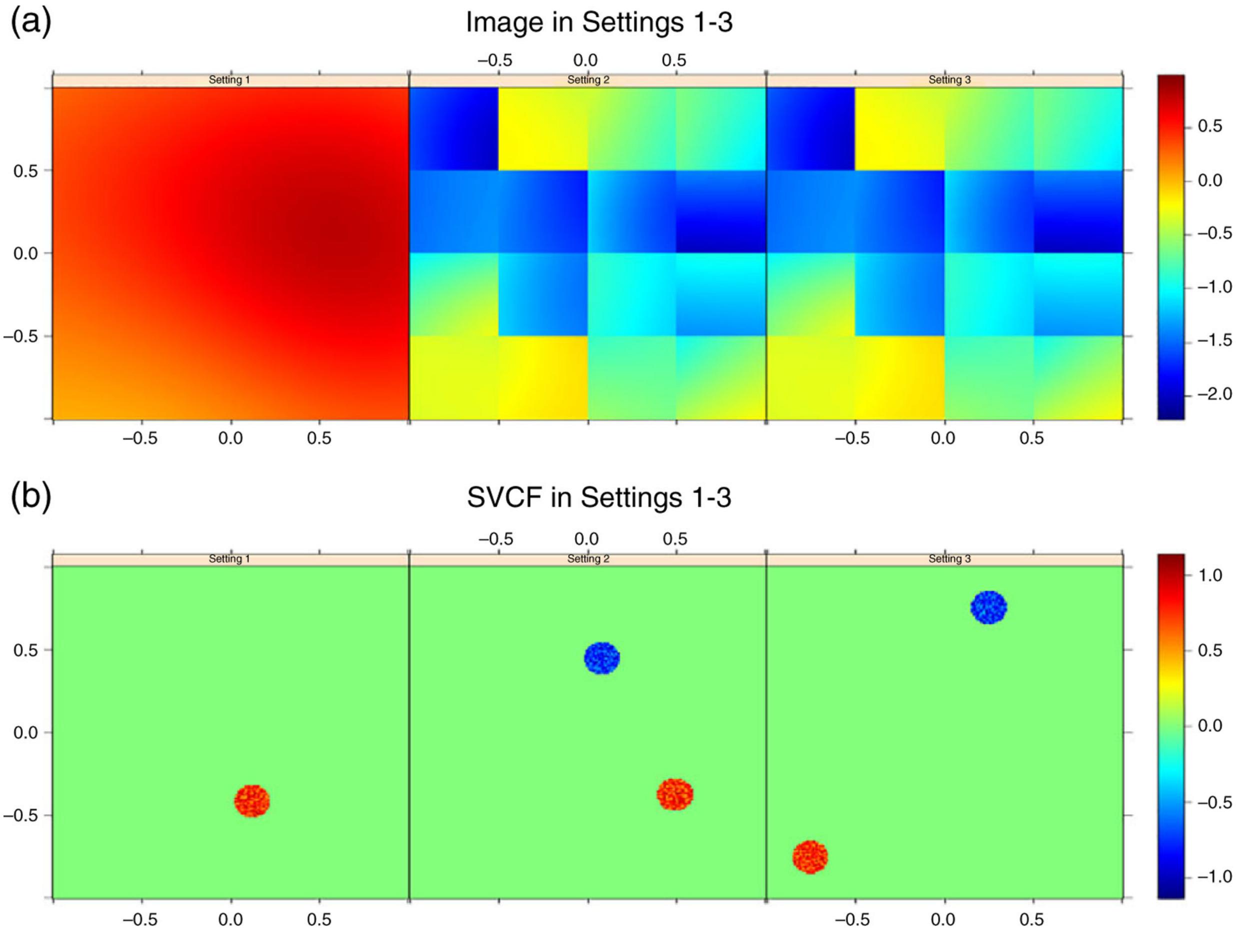
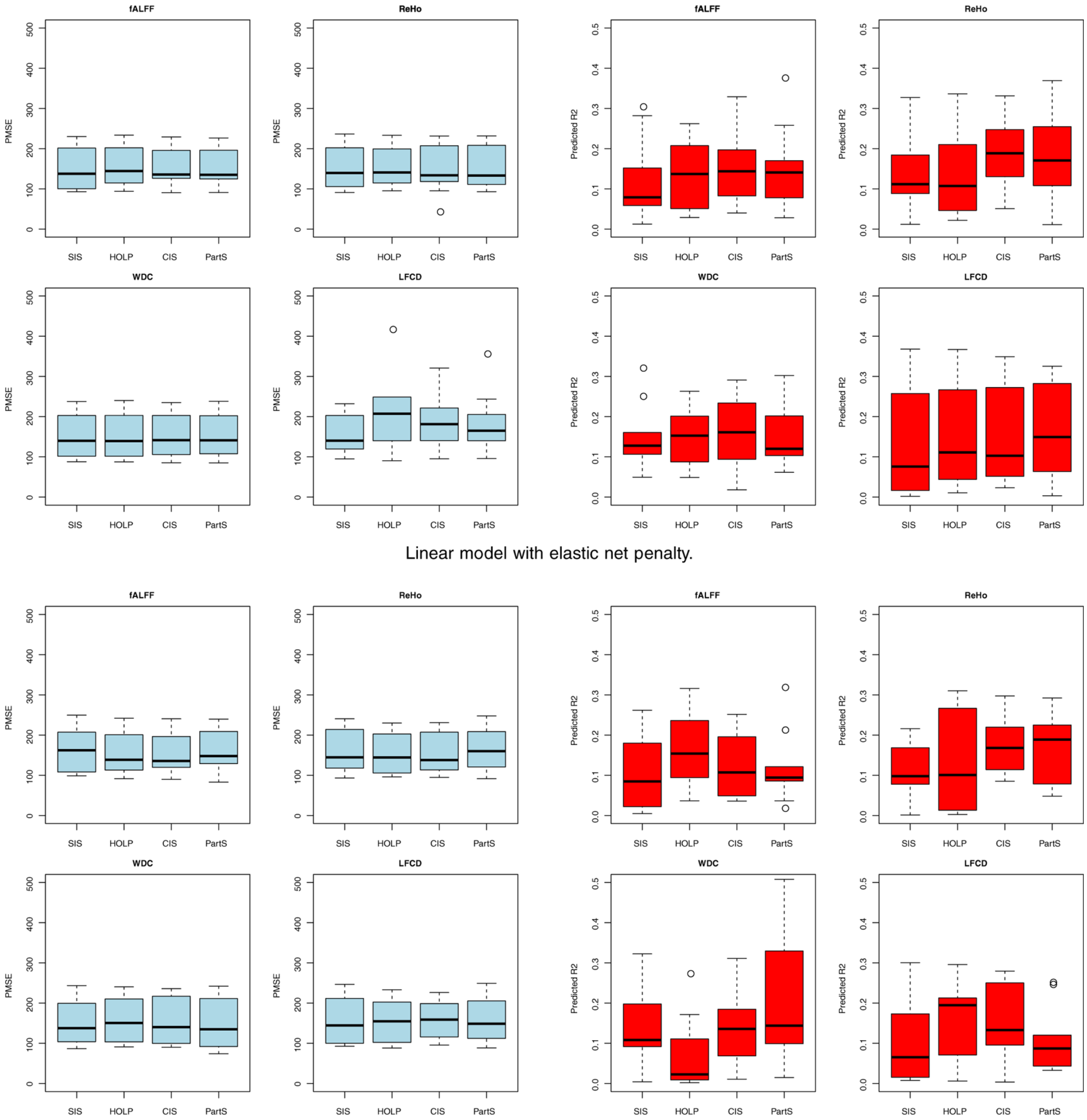


FIGURE 1.

Typical simulated images (200 × 200) and the true spatially varying coefficient function (SVCF) for different settings



Linear model with elastic net penalty.

Random forest model with 500 trees.

FIGURE 2.

Box plots of PMSE (light blue) and PR2 (red) over 10 cross-validations for four different imaging measures (fALFF, WDC, ReHo, and LFCD) and four different feature screening methods (SIS, HOLP, CIS, and PartS) using linear models and random forest models. CIS, covariance-insured screening; fALFF, fractional amplitude of low-frequency fluctuations; HOLP, high-dimensional ordinary-least squares projection; LFCD, local functional connectivity density; PartS, partition-based screening; PMSE, predicted mean square error; PR2, predicted R2; ReHo, Regional Homogeneity; SIS, sure independence screening; WDC, weighted degree centrality

TABLE 1

Screening accuracy of different methods

					SIS		**HOLP**		**CIS**		**PartS**	
Setting	***n***	***P***	***P*0**	$\boldsymbol{R^2}$	**FPR**	**FNR**	**FPR**	**FNR**	**FPR**	**FNR**	**FPR**	**FNR**
1	500	10,000	77	0.5	0.009	0.000	0.051	0.043	0.041	0.029	0.078	0.160
				0.9	0.001	0.000	0.043	0.009	0.006	0.000	0.072	0.145
	500	40,000	311	0.5	0.007	0.000	0.048	0.028	0.104	0.128	0.107	0.250
				0.9	0.001	0.000	0.043	0.003	0.004	0.002	0.089	0.184
	1,000	40,000	311	0.5	0.004	0.000	0.047	0.027	0.045	0.057	0.133	0.333
				0.9	0.000	0.000	0.043	0.001	0.003	0.000	0.103	0.247
2	500	10,000	154	0.5	0.161	0.190	0.695	0.856	0.274	0.232	0.164	0.243
				0.9	0.134	0.180	0.098	0.173	0.024	0.034	0.130	0.190
	500	40,000	622	0.5	0.156	0.187	0.722	0.888	0.551	0.491	0.148	0.248
				0.9	0.121	0.151	0.373	0.558	0.034	0.051	0.113	0.196
	1,000	40,000	622	0.5	0.094	0.129	0.729	0.874	0.379	0.338	0.110	0.203
				0.9	0.081	0.132	0.456	0.645	0.040	0.042	0.098	0.195
3	500	10,000	135	0.5	0.073	0.000	0.745	0.968	0.356	0.378	0.097	0.187
				0.9	0.071	0.000	0.120	0.207	0.008	0.000	0.090	0.081
	500	40,000	620	0.5	0.071	0.000	0.823	0.990	0.557	0.611	0.094	0.146
				0.9	0.069	0.000	0.488	0.829	0.003	0.000	0.090	0.118
	1,000	40,000	620	0.5	0.068	0.000	0.863	0.996	0.379	0.404	0.090	0.141
				0.9	0.066	0.000	0.534	0.933	0.002	0.000	0.087	0.103

Notes. n is the number of patients; p is the number of predictors; p_0 is the average number of true signals; R^2 is the model-explained variance ratio; FPR is the false-positive rates when covering 80% of the truth; FNR is the false-negative rates while keeping false-positive rate = 0.1.

TABLE 2

Average of PMSE and PR2 over 10 cross-validations for four different imaging measures (fALFF, WDC, ReHo, and LFCD) and four different feature screening methods (SIS, HOLP, CIS, and PartS) using linear models and random forest model

Method	**Data**	**Measure**	**SIS**	**HOLP**	**CIS**	**PartS**
Elastic-net	fALFF	PMSE	137.428	144.179	135.743	134.912
		PR2	0.079	0.141	0.144	0.141
	WDC	PMSE	139.892	139.328	141.585	141.105
		PR2	0.128	0.153	0.161	0.120
	ReHo	PMSE	139.627	140.790	133.741	133.169
		PR2	0.112	0.109	0.188	0.171
	LFCD	PMSE	140.206	207.365	181.210	164.875
		PR2	0.076	0.113	0.103	0.165
Random-forest	fALFF	PMSE	162.259	138.460	135.681	147.705
		PR2	0.085	0.154	0.109	0.094
	WDC	PMSE	137.816	150.593	140.195	134.822
		PR2	0.109	0.023	0.136	0.144
	ReHo	PMSE	144.759	144.312	137.984	159.988
		PR2	0.098	0.106	0.168	0.189
	LFCD	PMSE	144.484	154.812	158.927	148.611
		PR2	0.065	0.195	0.133	0.087

CIS, covariance-insured screening; fALFF, fractional amplitude of low-frequency fluctuations; HOLP, high-dimensional ordinary-least squares projection; LFCD, local functional connectivity density; PartS, partition-based screening; PMSE, predicted mean square error; PR2, predicted R2; ReHo, Regional Homogeneity; SIS, sure independence screening; WDC, weighted degree centrality.