Ensemble-based Ultrahigh-dimensional Variable Screening

Wei Tu[1*], **Dong Yang**[1*], **Linglong Kong**[1†], **Menglu Che**[2], **Qian Shi**[1], **Guodong Li**[3] and **Guangjian Tian**[4]

[1]Department of Mathematical and Statistical Sciences, University of Alberta
[2]Department of Statistics and Actuarial Science, University of Waterloo
[3]Department of Statistics and Actuarial Science, University of Hong Kong
[4]Huawei Noah's Ark Lab, Hong Kong, China
{wei.tu, dy2, lkong, qshi}@ualberta.ca, m3che@uwaterloo.ca, gdli@hku.hk,
tian.guangjian@huawei.com

Abstract

Since the sure independence screening (SIS) method by Fan and Lv [2008], many different variable screening methods have been proposed based on different measures under different models. However, most of these methods are designed for specific models. In practice, we often have very little information about the data generating process and different methods can result in very different sets of features. The heterogeneity presented here motivates us to combine various screening methods simultaneously. In this paper, we introduce a general ensemble-based framework to efficiently combine results from multiple variable screening methods. The consistency and sure screening property of proposed framework has been established. Extensive simulation studies confirm our intuition that the proposed ensemble-based method is more robust against model specification than using single variable screening method. The proposed ensemble-based method is used to predict attention deficit hyperactivity disorder (ADHD) status using brain function connectivity (FC).

1 Introduction

The evolution of data acquisition technologies and computing power has allowed researchers nowadays to collect and store data with high dimensionality and complex structure much more efficiently. Examples can be found in gene expression microarray data, single nucleotide polymorphism (SNP) data, magnetic resonance imaging (MRI) data, high-frequency financial data, and others. One common task is to extract useful variables from a high dimensional feature space to explain or predict a response variable. Traditional variable selection methods such as forward selection, backward elimination and best subset selection become computationally expensive or even infeasible at these conditions. To address these problems, a family of penalized least squares based methods has been developed. Examples include Lasso and Adaptive Lasso ([Tibshirani, 1996; Zou, 2006]), SCAD [Fan and Li, 2001], elastic net [Zou and Hastie, 2005], and MCP [Zhang, 2010]. However, when the dimensionality p is much larger than the sample size n or even grows exponentially with n, the aforementioned penalization methods can perform poorly or even become infeasible due to the simultaneous challenges of computational expediency, statistical accuracy and algorithm stability [Fan *et al.*, 2009]. For example, in MRI studies, images with dimension $1024 \times 1024 \times 200$ can be acquired for each subject, and due to the high cost of the MRI scanning, studies might only contain less than 100 subjects. If we treat the signal from each voxel as a feature, the dimension of feature space p is much higher than the sample size n.

A natural idea to address these challenges is to reduce the dimensionality p from a large scale to a relatively large scale d using a fast screening algorithm, and then the ultrahigh-dimensional problem can be greatly simplified into a moderately high-dimensional one. Subsequently, standard penalized variable selection methods can be applied to the remaining variables. Fan and Lv [2008] first introduced the sure independence screening (SIS) by ranking the marginal correlation of each covariate and the response. The good numerical performance and novel theoretical properties have made SIS popular in ultrahigh dimensional reduction. As a result, SIS and its extensions have been generalized to many important settings including generalized linear model [Fan *et al.*, 2010], multi-index semi-parametric models [Zhu *et al.*, 2011], non-parametric regression [Fan *et al.*, 2011], quantile regression [He *et al.*, 2013]. Other marginal screening methods based on different measure of association between predictors and response have also been studied, such as Kendall's τ [Li *et al.*, 2012a], distance correlation [Li *et al.*, 2012b]. We refer to [Liu *et al.*, 2015] for a more comprehensive list of references.

Most feature screening methods are designed for specific models, and they all enjoy good theoretical and numerical performances under a set of conditions on the data generating process. However, in practice, we rarely know the actual relationship between features and the response, and different covariates might have different relationships with the response.

[*] These authors contributed equally to this work.

[†] Correspondence Author.

The heterogeneity presented here motivates us to consider various screening methods simultaneously in practice, so the natural question here is how to aggregate the results from these screening methods to achieve more robust and accurate performances. To the authors' best knowledge, there is no existing work on addressing these problems. In this paper, we propose an ensemble-based framework for combining results from multiple variable screening methods.

It is well known that ensemble methods is usually more accurate than a single learner, and they have already achieved great success in many real-world tasks [Zhou, 2012]. In the area of variable selection, the idea of ensemble has been studied before ([Bach, 2008; Bolón-Canedo and Alonso-Betanzos, 2019]). In section 2, we introduce the details of ensemble-based variable screening. In section 3, we establish the theoretical properties of proposed method. In section 4, we conduct extensive simulation studies under different models. In section 5, we conduct a real data analysis using functional magnetic resonance imaging (fMRI) data to predict attention deficit hyperactivity disorder (ADHD) status. In summary, we have the following contributions and findings:

1. We propose an ensemble-based variable screening framework to efficiently combine results from different variable screening methods. The proposed framework is very flexible, and can be paralleled.
2. We prove that the proposed ensemble-based variable screening method inherits the nice theoretical properties (e.g. consistency, sure screening property) of the base candidates.
3. We conduct extensive simulation studies and real data analysis to illustrate the numerical performance of proposed framework. The proposed ensemble-based method more robust again model specification than using single variable screening method, that is to say, even though the proposed ensemble based screener may not outperform all base screener in a specific model setting, but it has the most consistent and robust performance across different model settings.

2 Ensemble-based Variable Screening

2.1 A General Framework

We consider the problem of variable screening in ultrahigh-dimensional feature space, where we observe response variable Y and the associated covariate vector $\mathbf{X} = (X_1, \dots, X_p)^T$. Consider the conditional distribution function of Y given $\mathbf{x}$, denoted by $F(y|\mathbf{X}) = P(Y < y|\mathbf{X})$. Define two sets of variables:

$$\begin{aligned}\mathcal{A} =& \{j : F(y|\mathbf{X}) = P(Y < y|\mathbf{X}) \\ & \text{functionally depends on } X_j\},\\ \mathcal{I} =& \{j : F(y|\mathbf{X}) = P(Y < y|\mathbf{X}) \\ & \text{does not functionally depends on } X_j\}.\end{aligned}$$

If $j \in \mathcal{A}$, X_j is referred to as an active feature, otherwise an inactive feature. The goal is to reduce dimensionality p from a large scale to a moderate scale by a fast and efficient method, while including all the variables in set $\mathcal{A}$.

The key idea of the marginal screening procedure is to rank all predictors by using a utility measure between the response and each predictor and then to retain the top variables for further investigation. It usually involves three steps:

1. Calculate the screening statistic vector $F^\top = [f_1, \dots, f_p]$. For example, $f_j = \text{cor}(X_j, Y)$ is used in SIS.
2. Obtain the rank vector $R^\top = [r_1, \dots, r_p]$ by ranking F from largest to smallest (usually a large f_j indicates a stronger relationship between X_j and Y).
3. Choose the $\lfloor \gamma p \rfloor$ top ranked features as the active set $\widehat{\mathcal{A}}$, and $\gamma \in (0, 1)$ is predetermined. To be specific, the vote vector $V^\top = [v_1, \dots, v_p]$ with $v_j = I(r_j \leq \lfloor \gamma p \rfloor)$.

Different screening methods usually result in different ranking, and in the next section we introduce the concept of variable screening ensembles to combine the results from different methods.

2.2 Constructing Variable Screening Ensembles

Generally, an ensemble is constructed in two steps: generating the base learners and combining them. To get a good ensemble, it is generally believed that the base learners should be as accurate as possible, and as diverse as possible. Consider K base variable screening candidates, each one of them returns screening utilities of all p variables. Now we introduce the variable screening ensemble matrix:

$$\mathbf{W} = \begin{bmatrix} \mathbf{W}_1^\top \\ \mathbf{W}_2^\top \\ \vdots \\ \mathbf{W}_K^\top \end{bmatrix} = \begin{bmatrix} w_{11} & w_{12} & w_{13} & \dots & w_{1p} \\ w_{21} & w_{22} & w_{23} & \dots & w_{2p} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ w_{K1} & w_{K2} & w_{K3} & \dots & w_{Kp} \end{bmatrix}$$

where $\mathbf{W}_i^\top = [w_{i1}, \dots, w_{ip}], i = 1, \dots, K$ denotes the screening utilities obtained using screening candidate i. $\mathbf{W}_i$ can be the vector of the screening statistics vector F_i of all p features, for example, the correlation between $\mathbf{X}$ and Y in SIS.

If a single base variable screening candidate has been used K times through bootstrapping or parameter tuning, we obtain homogeneous ensembles. On the other hand, if multiple screening candidates have been used, this leads to heterogeneous ensembles. Since the screening statistics produced by different base candidates can have different scale and/or range, $\mathbf{W}_i$ and $\mathbf{W}_j$ may not be comparable. In this case, we can choose the rank of all p features R_i for $\mathbf{W}_i$ obtained using base candidate i. Another choice is to use the binary vote V_i introduced in the last section. Intuitively, using the rank R will be more robust compared to directly using screening statistic F, and it will be more efficient compared to using binary vote V since V can be sensitive to the choice of threshold parameter γ. In section 4, we further use numerical studies to confirm this intuition. After obtaining the ensemble matrix, we discuss how to efficiently combine the columns of the matrix.

2.3 Aggregation Functions

After constructing the ensemble matrix, the next step would be finding appropriate combination functions to aggregate the results. A general combination function is a multivariate function that projects the j-th column of the ensemble matrix to a real number, that is,

$$f(w_{1j}, \ldots, w_{Kj}) \to \mathbb{R}.$$

There are many combination functions available in the literature [Zhou, 2012]. In our ensemble screening framework, we consider two major types of combination functions.

Mean Combination: Taking means of prediction results is a commonly adopted approach in machine learning literature [Nguyen *et al.*, 2018]. In our approach, mean combination function combines the output of the screening algorithms by taking the mean of each column in the ensemble matrix. Provided the screening utilities associated with the j-th column as $w_{1j}, \ldots, w_{Kj}$, the mean ensemble function is defined as

$$f(w_{1j}, \ldots, w_{Kj}) = \frac{1}{K}\sum_{i=1}^{K} w_{ij}, j = 1, \ldots, p.$$

Median Combination: Instead of choosing the mean of $w_{1j}, \ldots, w_{Kj}$, we use the median, which is more robust when those screening utilities are skewed or with outliers.

As an alternative to mean, median provides a more robust result when the distribution of w_{ij} is skewed. For example, we use the ranking vector $R_i^\top$ obtained from base candidate i as each row of the ensemble matrix, $\mathbf{W}_i^\top$. For active variable X_j, if most base candidates return a small rank except for one or two "outliers", then the mean-aggregated result will be largely influenced by these "outliers", while the median-aggregated result will still be robust. Formally we have

$$f(w_{1j}, \ldots, w_{Kj}) = \text{Median}\left\{w_{1j}, \ldots, w_{Kj}\right\}, j = 1, \ldots, p.$$

2.4 Variables Selected

Denote the combined information of the ensemble matrix by applying mean or median aggregation function as $\{E_1, \ldots, E_p\}$. For any given threshold $\gamma \in (0, 1)$, the $\lfloor \gamma p \rfloor$ top ranked predictors are selected as the active set:

$$\widehat{\mathcal{A}} = \{1 \leq j \leq p : |E_j| \text{is among the first } \lfloor \gamma p \rfloor \text{ largest of all}\}.$$

3 Technical Results

As variable screening only serves as the first step in high dimensional data analysis, the most important property as far as practical application is concerned the sure screening property [Fan and Lv, 2008]. That is, with probability approaching 1, the screening algorithm keeps all of the true active variables.

$$P(\mathcal{A} \subseteq \widehat{\mathcal{A}}) \to 1.$$

We require that the sure screening property holds for each screening algorithms. For the quantile based variable screening algorithms, *e.g.* QaSIS , we require the sure screening property holds at each quantile level τ_k which means the selected variables contain the true active set $\mathcal{A}_k$ with a probability tending to one.

Regarding the screening utilities, we require all the screening algorithms in the ensemble have consistency of the screening utilities, that is for $1 \leq i \leq K$,

$$P(\max_{1\leq j\leq p} |\widehat{w}_{ij} - w_{ij}| > \delta_n) \to 0,$$

where w_{ij} is the corresponding screening utility for screening algorithm i and variable j, $\widehat{w}_{ij}$ is an estimator for w_{ij}, and δ_n is some threshold number that is usually related to n. This requirement is reasonable as it is shown in most of the variable screening literatures [Fan and Lv, 2008; Fan *et al.*, 2011; He *et al.*, 2013; Zhu *et al.*, 2011].

Lemma 3.1 *(Consistency of aggregated screening utilities) Given number of K screening algorithms which are based on screening utilities $w_{ij}, i = 1, \ldots, K$ and $j = 1, \ldots, p$. Denote f as the combination function. Assume the following:*

$$P(\max_{1\leq j\leq p} |\widehat{w}_{ij} - w_{ij}| > \delta_n) \to 0$$

$$|f(\widehat{w}_{1j}, \ldots, \widehat{w}_{Kj}) - f(w_{1j}, \ldots, w_{Kj})| \leq \max_{1\leq i\leq K} |\widehat{w}_{ij} - w_{ij}|,$$

in which δ_n is some threshold constant related to n. We have the consistency of the aggregated screening utility which is:

$$P(\max_{1\leq j\leq p} |f(\widehat{w}_{1j}, \ldots, \widehat{w}_{Kj}) - f(w_{1j}, \ldots, w_{Kj})| > \delta_n) \to 0.$$

Theorem 3.2 *(Sure screening property) Denote w_j as the j-th aggregated screening utility, and w_j^* as the sample estimate. Denote $\mathcal{A}$ as the active variable set. Assume $\min_{j\in\mathcal{A}} w_j > 2\delta_n$ and our screening method select the variables with $w_j^* > \delta_n$. Given the assumptions of the previous lemma, we have the sure screening property for our ensemble screening approach:*

$$P(\mathcal{A} \subseteq \widehat{\mathcal{A}}) \to 1,$$

in which δ_n is some threshold number related to n.

Remark: To guarantee the sure screening property holds for our ensemble procedure, we just need δ_n to be the smallest among all the thresholds we combined. As long as all of the candidate methods show sure screening property, given the assumptions in Lemma 3.1 hold, our ensemble approach also enjoys sure screening property.

Lemma 3.3 *(Lower bound for simultaneous screening probability) Denote $\widehat{\mathcal{T}}_1$ and $\widehat{\mathcal{T}}_2$ as two selected sets by applying two different screening algorithms. Define $\widehat{\mathcal{T}}^{simu} = \widehat{\mathcal{T}}_1 \cap \widehat{\mathcal{T}}_2$ as the variable set selected by both of the screening algorithms. Define $\widehat{\Pi}_K^1$, $\widehat{\Pi}_K^2$, $\widehat{\Pi}_K^{simu}$ as the probabilities of selected sets of screening algorithm 1, screening algorithm 2 and the simultaneous set containing a variable set K, $K \subseteq \{1, \ldots, p\}$. Then we have:*

$$\widehat{\Pi}_K^{simu} \geq 2\min\left\{\widehat{\Pi}_K^1, \widehat{\Pi}_K^2\right\} - 1.$$

Lemma 3.4 *Let $K \subset \{1, \ldots, p\}$ be a set of variables and $\widehat{\mathcal{T}}_i$ be the set of selected variables by applying a variable screening algorithm i. If $\max\left\{P(K \subseteq \widehat{\mathcal{T}}_1), P(K \subseteq \widehat{\mathcal{T}}_2)\right\} \leq \varepsilon$, then*

$$P(\widehat{\Pi}_K^{simu} \geq \xi) \leq \varepsilon^2/\xi.$$

Theorem 3.5 *(Error control) Denote $S = |\mathcal{T}|$ and $N = |\mathcal{F}|$ as the number of underlying true important and unimportant variables. Correspondingly, denote $\widehat{S} = |\mathcal{T} \cap \widehat{\mathcal{T}}^{simu}|$*

and $\widehat{N} = |\mathcal{F} \cap \widehat{\mathcal{F}}^{simu}|$ as the number of estimated important and unimportant variables. In addition denote $V = \mathrm{E}(|\mathcal{F} \cap \widehat{\mathcal{T}}^{simu}|)$ as the expected number of falsely selected variables in $\widehat{\mathcal{T}}^{simu}$. Assume exchangeability which is $P(k \in \widehat{\mathcal{T}}) = \mathrm{E}(\widehat{N})/N$, where $k \in (1, \ldots, p)$. Also assume that the candidate variable screening process is not worse than random guessing. Given the screening threshold T_n and the threshold of selection probability π_{thr}, we have:

$$\mathrm{E}(V) \leq \frac{1}{2\pi_{thr} - 1} \frac{T_n^2}{p}.$$

Remark: If a variable is important, the underlying true ranking of the variable is higher than the other unimportant variables. With a threshold T_n, the true important variables are supposed to rank within T_n. If T_n is a relative small number say 30, and the probability threshold π_{thr} is decently greater than 50%, given $p = 1000$ the expectation number of falsely discovered variables could be controlled within a small number. Intuitively, when each candidate screening method is good enough and the threshold T_n is small, the number of falsely selected variables could be controlled at a very low level.

4 Numerical Studies

4.1 Simulation Settings

In this section, we demonstrate the numerical performance of our proposed approach. For the base screening candidates, we consider three popular methods, the SIS in [Fan and Lv, 2008], the sure independence ranking and screening (SIRS) in [Zhu *et al.*, 2011], and the quantile adaptive sure independence screening (QaSIS) in [He *et al.*, 2013]. The SIS is based on linear model, and both SIRS and QaSIS are model free. These three methods all have their advantages and disadvantages, and none of them uniformly performs better than the others in all models. For QaSIS, different quantile levels can be used for the screening, we first consider the ensemble of QaSIS with different quantile levels, and there will be a more homogeneous scenario. Secondly we consider a more heterogeneous case by combining all these three different methods.

We consider two evaluation metrics: the first one is $\mathcal{R}$, that is the smallest number of features that we need to include to ensure that all the active variables are selected; the second one (denoted by $\mathcal{S}$) is the proportion of the active predictors selected when the threshold $T_n = \lfloor n/\log(n) \rfloor$ is adopted. An effective variable screening procedure is expected to have the value of $\mathcal{R}$ reasonably small comparing to the number of active variables and the value of $\mathcal{S}$ close to one. All the experiments have been repeated 100 times, and the median and interquartile range (IQR) have been reported for both $\mathcal{R}$ and $\mathcal{S}$.

Depending on the choice of $\mathbf{W}_i$ in the ensemble matrix and the aggregation function, different variable screening ensembles can be obtained. We use VSrE (Variable Screening Ensemble), VSrE-R and VSrE-V to denote the ensembles obtained by choosing $\mathbf{W}$ as the original screening statistic F, the rank R and the binary vote V respectively (defined in section 2.1). For the aggregation function, mean and median are used. For example, mean-based ensemble is denoted as VSrE(mean). Note that the median aggregation function for the vote based VSrE may not be a good choice since median is not a good summary statistics for binary data.

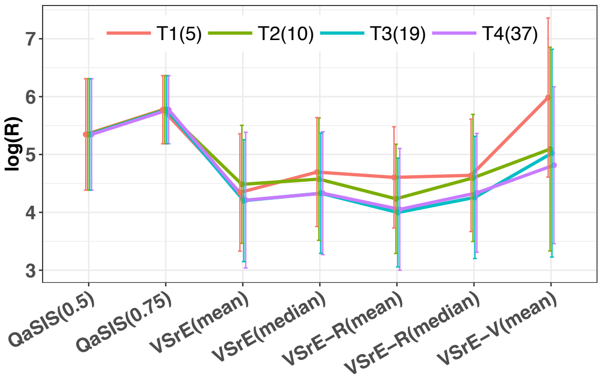


Figure 1: Comparisons of different ensemble-based methods via combining QaSIS with different quantile levels under Model 1. Each line here represents different choices of quantile set with the number in the bracket represents the number of quantile levels used. For each method under each model, median $\log(\mathcal{R})$ and its IQR (error bar) are presented.

4.2 Models Settings

We consider the following four models, from linear, nonlinear to heteroscedastic cases. The random error term ε follows a standard normal distribution in all four models and is independent of $\mathbf{X}$.

Model 1$(n = 200, p = 2000)$ [Fan and Lv, 2008]: This is a model of the form $Y = \boldsymbol{X\beta} + \varepsilon$, in which all the columns of $\boldsymbol{X}$ are generated from a standard normal distribution. There are in total eight true active predictors which are generated as follows: set $a = 4\log(n)/n^{\frac{1}{2}}$ and the coefficients corresponding to the active predictors are derived by $(-1)^u(a + |z|)$, where u follows a Bernoulli distribution with $p = 0.4$ and z is drawn from a standard normal distribution.

Model 2$(n = 200, p = 2000)$ [Zhu *et al.*, 2011]: The random data is generated from $Y = 2(X_1 + 0.8X_2 + 0.6X_3 + 0.4X_4 + 0.2X_5) + \exp(X_{20} + X_{21} + X_{22}) \cdot \varepsilon$, where $\boldsymbol{X} = (X_1, X_2, \ldots, X_{2000})$ follows a multivariate normal distribution with mean $\mathbf{0}$ and covariance $\boldsymbol{\Sigma} = (\sigma_{ij})_{p \times p}$ with $\sigma_{ij} = \rho^{|i-j|}$. We consider $\rho = 0.8$ here. The number of active variables here is 8.

Model 3$(n = 400, p = 1000)$ [Fan *et al.*, 2011]: First, define the following functions $g_1(x) = x$; $g_2(x) = (2x-1)^2$; $g_3(x) = \sin(2\pi x)/(2 - \sin(2\pi x))$; $g_4(x) = 0.1\sin(2\pi x) + 0.2\cos(2\pi x) + 0.3\sin(2\pi x)^2 + 0.4\cos(2\pi x)^3 + 0.5\sin(2\pi x)^3$. The random data are generated from: $Y = 5g_1(X_1) + 3g_2(X_2) + 4g_3(X_3) + 6g_4(X_4) + \sqrt{1.74}\varepsilon$, where $\boldsymbol{X} = (X_1, X_2, \ldots, X_{1000})$ has the similar structure as in Model 2 with $\rho = 0.4$.

Model 4$(n = 400, p = 5000)$ [He *et al.*, 2013]: $Y = 2(X_1^2 + X_2^2) + [10^{-1}\exp(X_1 + X_2 + X_{18} + X_{19} + \ldots + X_{30})] \cdot \varepsilon$., where $\mathbf{X} = (X_1, X_2, \ldots, X_{5000})$ has the similar structure as Model 2 with $\rho = 0.8$. The number of active variables here is 15.

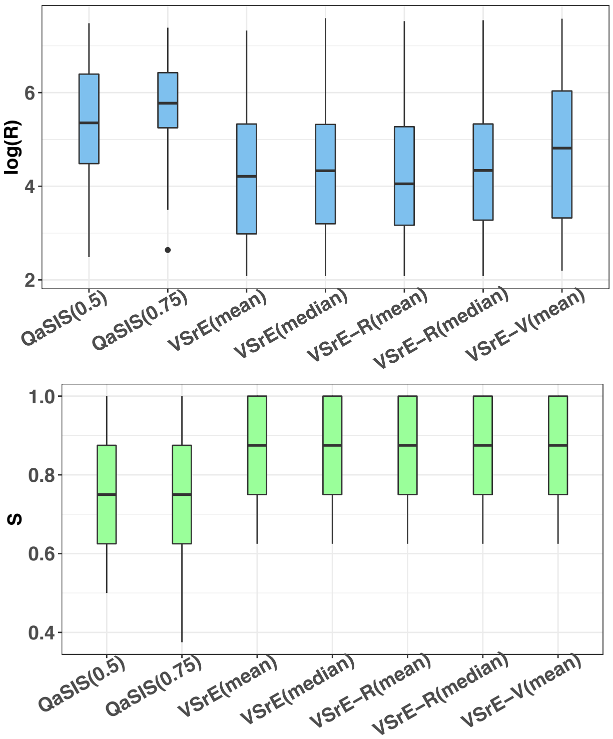


Figure 2: Boxplots of $\log(\mathcal{R})$ (upper) and $\mathcal{S}$ (lower) for ensemble-based screening methods by combining QaSIS with different quantile levels under Model 1. Quantile level set T_2 with $K = 10$ is used here.

4.3 Combining Quantile Levels

We first consider VSrE constructed by applying QaSIS separately on different quantile levels. In order to show the effect of different choices of quantile sets, we show the results of applying four different quantile sets which have 5, 10, 19 and 37 equally spaced quantile levels respectively as follows: $T_1 = \{0.1, 0.3, \ldots, 0.9\}$, $T_2 = \{0.05, 0.15, \ldots, 0.95\}$, $T_3 = \{0.05, 0.1, \ldots, 0.95\}$, $T_4 = \{0.05, 0.075, \ldots, 0.95\}$. As an illustration of the choice of VSrEs and combination functions, all five types of methods are applied. Model 1 is used for this simulation. The line chart in Figure 1 shows the comparison of single screening method (QaSIS(0.5) and QaSIS(0.75)) and the proposed ensemble-based methods using 4 different sets of quantile levels, while the boxplots in Figure 2 focus on the case when quantile level set T_2 with $K = 10$ is used. From Figure 1 and Figure 2, we have the following observations:

1. All proposed ensemble-based methods outperform the single base screening method, with regard to both $\mathcal{R}$ and $\mathcal{S}$.
2. The performance of ensemble-based method improves with the increase of the number of base screening methods K. But the improvement becomes substantial as K reaches a certain level, while the computation time increases linearly with K.
3. When the screening statistic is comparable across ensemble candidates, both mean- and median- based VSrE and VSrE-R provides reasonably good results.

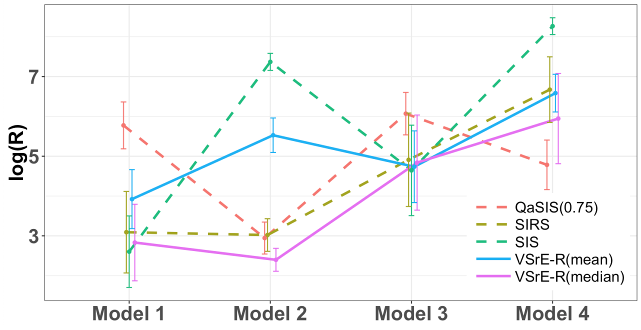


Figure 3: Comparison of proposed ensemble-based methods (solid lines) and single ensemble candidates (dash lines) under different model settings. For each method under each model, median $\log(\mathcal{R})$ and its IQR (error bar) are presented. Note that the ensemble-based VSrE-R here is obtained by combining QaSIS(0.75), SIRS and SIS.

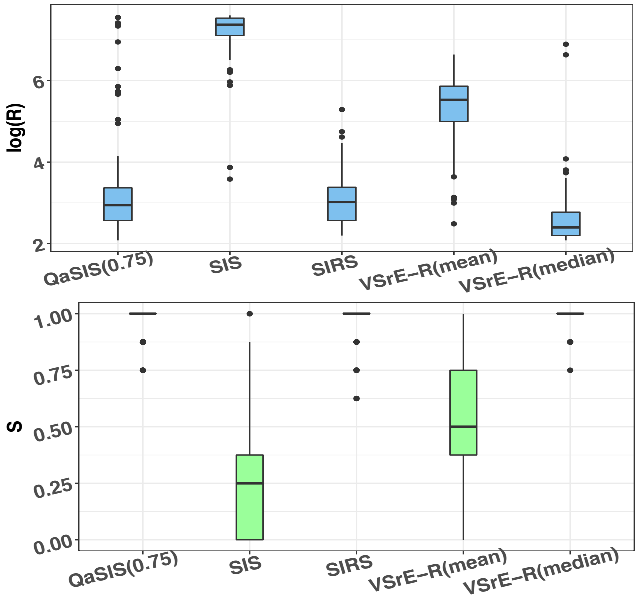


Figure 4: Boxplots of $\log(\mathcal{R})$ (upper) and $\mathcal{S}$ (lower) for different variable screening methods under Model 2. Note that the ensemble-based VSrE-R here is obtained by combining QaSIS(0.75), SIRS and SIS.

4.4 Combining Different Variable Screening Methods

Instead of constructing a VSrE by applying QaSIS on different quantile levels, we could apply different screening methods to the sample and combine the results. In the following, we show the power of simply combining three commonly used variable screening methods which are SIS, SIRS and QaSIS. Quantile level 0.75 is used for QaSIS. As the screening utilities of different methods may have very different magnitudes, it is more appropriate to use VSrE-R in order to avoid the situation that some screening methods with relative large screening utilities might dominate the combined result.

Figure 3 compares the ensemble-based VSrE-R(mean) and VSrE-R(median) with the single screening candidates under Model 1 to Model 4. Figure 4 provides more detailed results on Model 2 about both $\mathcal{R}$ and $\mathcal{S}$. Based on Figure 3 and Figure 4, we have the following observations:

1. The proposed VSrE-R(median) has the most robust performance across all four models.
2. Each ensemble candidate has its own advantage: SIS performs the best under Model 1, SIRS under Model 3 and QaSIS under Model 4. The ensemble-based methods inherit the good performance of the base screening candidates, while still robust to model misspecification as long as some of the base learners perform reasonably well.
3. For the aggregation function, median is a more robust choice.

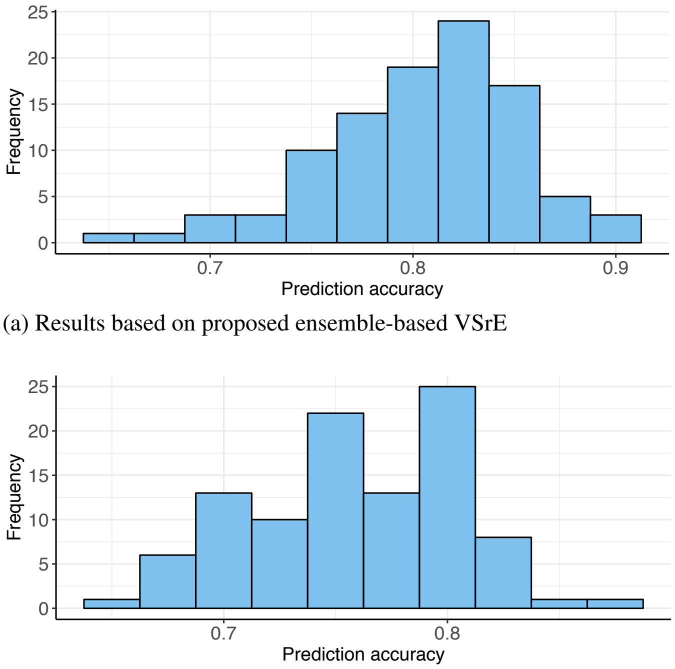
5 Real Data Analysis

Attention deficit hyperactivity disorder (ADHD) is a brain disorder marked by an ongoing pattern of inattention and hyperactivity-impulsivity that interferes with functioning or development. Recent development of medical imaging such as functional magnetic resonance imaging (fMRI) and diffusion tensor imaging (DTI) shows promising potential in predicting patients outcomes and understanding the underlying pathophysiology of diseases [Greicius *et al.*, 2007].

We use the ADHD-200 Consortium data which is a publicly available resting-state fMRI (rs-fMRI) data [Milham *et al.*, 2012] in this study. fMRI measures brain activity by detecting changes associated with blood flow [Huettel *et al.*, 2004]. This data set contains 120 subjects ($n = 120$) from the NYU site (New York University Child Study Center) of the ADHD-200 Consortium, 42 are typically developing children showing ADHD negative and 78 are diagnosed as ADHD. The Anatomical Automatic Labeling (AAL) atlas [Tzourio-Mazoyer *et al.*, 2002] was used for the parcellation.

In this study, our goal is to find which brain connectivity pair of ROIs (region of interest) is contributing to the ADHD express levels hence each brain connectivity is considered to be a variable. For each subject, there are 172 time points and the AAL has 116 ROIs. We obtain the mean time series for each of the 116 regions by averaging the fMRI time series over all voxels in the region, hence initially we have $p = (116 \times 116 - 116)/2 = 6670$ predictors. Because of the large p small n scenario ($n = 120$, $p = 6670$), a variable screening procedure is necessary to remove some noise brain connectivities in order to apply lower dimensional variable selection approaches. We apply VSrE-R(median) by combining SIS, SIRS and QaSIS with quantile level 0.75 and use QaSIS at quantile level 0.75 as a comparison. In the variable selection step, we employ both LASSO and SCAD. The tuning parameters of LASSO and SCAD are selected. In the next step, we choose those ROI connectivities that are simultaneously selected by both LASSO and SCAD. For the classification, we adopt the support vector machine classifier (SVM) [Hearst *et al.*, 1998] with linear kernel. All tuning parameter are selected by 10 folds cross validation.

Unlike the simulation settings, we do not know the true active variables, so we can not use metrics like $\mathcal{R}$ and $\mathcal{S}$. Instead we measure the usefulness of the selected features by their predictive power. Figure 5 shows the prediction accuracy using 100 bootstrapped samples. The top panel is the histogram of accuracy using ensemble-based method while the bottom panel is the one using QaSIS with quantile level 0.75. We can see a substantial improvement using the proposed method.



(a) Results based on proposed ensemble-based VSrE

(b) Results based on QaSIS

Figure 5: Prediction accuracy of ADHD status using single screening candidate (QaSIS (0.75)) and proposed ensemble-based method

Also, some of the partial correlations screened out by the proposed ensemble-based method include regions associated with a network called the default-mode network, which is a large and robustly replicable network of brain regions that is associated with task-irrelevant mental processes and mind-wandering. Similar findings are reported in [Uddin *et al.*, 2008; Tian *et al.*, 2006].

6 Discussion

In this paper, we introduce a general ensemble-based framework to combine results from different variable screening methods. The simulation studies confirm our intuition that the ensemble-based methods indeed inherit the good performance of the candidates, that is to say, as long as some candidates have decent performance, the ensemble-based method will work reasonably well. This is very important in practice, since when we do not have much information about the actual data generating process, the ensemble-based methods will be a more robust and safer choice than relying on a single method. Also, we find that the median rank-based aggregation method has the most robust performance. It is also worthy to point out that the proposed framework can be easily implemented in a parallel way so it will not add too much computation burden.

References

[Bach, 2008] Francis R Bach. Bolasso: model consistent lasso estimation through the bootstrap. In *Proceedings of the 25th international conference on Machine learning*, pages 33–40. ACM, 2008.

[Bolón-Canedo and Alonso-Betanzos, 2019] Verónica Bolón-Canedo and Amparo Alonso-Betanzos. Ensembles for feature selection: A review and future trends. *Information Fusion*, 52:1–12, 2019.

[Fan and Li, 2001] Jianqing Fan and Runze Li. Variable selection via nonconcave penalized likelihood and its oracle properties. *Journal of the American Statistical Association*, 96(456):1348–1360, 2001.

[Fan and Lv, 2008] Jianqing Fan and Jinchi Lv. Sure independence screening for ultrahigh dimensional feature space. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 70(5):849–911, 2008.

[Fan *et al.*, 2009] Jianqing Fan, Richard Samworth, and Yichao Wu. Ultrahigh dimensional feature selection: beyond the linear model. *Journal of Machine Learning Research*, 10:2013–2038, December 2009.

[Fan *et al.*, 2010] Jianqing Fan, Rui Song, et al. Sure independence screening in generalized linear models with np-dimensionality. *The Annals of Statistics*, 38(6):3567–3604, 2010.

[Fan *et al.*, 2011] Jianqing Fan, Yang Feng, and Rui Song. Nonparametric independence screening in sparse ultra-high-dimensional additive models. *Journal of the American Statistical Association*, 106(494):544–557, 2011.

[Greicius *et al.*, 2007] Michael D Greicius, Benjamin H Flores, Vinod Menon, Gary H Glover, Hugh B Solvason, Heather Kenna, Allan L Reiss, and Alan F Schatzberg. Resting-state functional connectivity in major depression: abnormally increased contributions from subgenual cingulate cortex and thalamus. *Biological Psychiatry*, 62(5):429–437, 2007.

[He *et al.*, 2013] Xuming He, Lan Wang, and Hyokyoung Grace Hong. Quantile-adaptive model-free variable screening for high-dimensional heterogeneous data. *The Annals of Statistics*, 41(1):342–369, 2013.

[Hearst *et al.*, 1998] Marti Hearst, Susan T. Dumais, E Osman, John Platt, and Bernhard Scholkopf. Support vector machines. *IEEE Intelligent Systems*, 13(4):18–28, July 1998.

[Huettel *et al.*, 2004] Scott A Huettel, Allen W Song, and Gregory McCarthy. *Functional magnetic resonance imaging*, volume 1. Sinauer Associates Sunderland, MA, 2004.

[Li *et al.*, 2012a] Gaorong Li, Heng Peng, Jun Zhang, and Lixing Zhu. Robust rank correlation based screening. *The Annals of Statistics*, pages 1846–1877, 2012.

[Li *et al.*, 2012b] Runze Li, Wei Zhong, and Liping Zhu. Feature screening via distance correlation learning. *Journal of the American Statistical Association*, 107(499):1129–1139, 2012.

[Liu *et al.*, 2015] JingYuan Liu, Wei Zhong, and RunZe Li. A selective overview of feature screening for ultrahigh-dimensional data. *Science China Mathematics*, 58(10):1–22, 2015.

[Milham *et al.*, 2012] Michael P Milham, Damien Fair, Maarten Mennes, and Stewart Mostofsky. The ADHD-200 consortium: a model to advance the translational potential of neuroimaging in clinical neuroscience. *Frontiers in Systems Neuroscience*, 6:62, 2012.

[Nguyen *et al.*, 2018] Tien Thanh Nguyen, Xuan Cuong Pham, Alan Wee-Chung Liew, and Witold Pedrycz. Aggregation of classifiers: a justifiable information granularity approach. *IEEE Transactions on Cybernetics*, pages 1–10, 2018.

[Tian *et al.*, 2006] Lixia Tian, Tianzi Jiang, Yufeng Wang, Yufeng Zang, Yong He, Meng Liang, Manqiu Sui, Qingjiu Cao, Siyuan Hu, Miao Peng, et al. Altered resting-state functional connectivity patterns of anterior cingulate cortex in adolescents with attention deficit hyperactivity disorder. *Neuroscience letters*, 400(1-2):39–43, 2006.

[Tibshirani, 1996] Robert Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society. Series B (Methodological)*, (1):267–288, 1996.

[Tzourio-Mazoyer *et al.*, 2002] Nathalie Tzourio-Mazoyer, Brigitte Landeau, Dimitri Papathanassiou, Fabrice Crivello, Olivier Etard, Nicolas Delcroix, Bernard Mazoyer, and Marc Joliot. Automated anatomical labeling of activations in SPM using a macroscopic anatomical parcellation of the MNI MRI single-subject brain. *Neuroimage*, 15(1):273–289, 2002.

[Uddin *et al.*, 2008] Lucina Q Uddin, AM Clare Kelly, Bharat B Biswal, Daniel S Margulies, Zarrar Shehzad, David Shaw, Manely Ghaffari, John Rotrosen, Lenard A Adler, F Xavier Castellanos, et al. Network homogeneity reveals decreased integrity of default-mode network in adhd. *Journal of neuroscience methods*, 169(1):249–254, 2008.

[Zhang, 2010] Cun-Hui Zhang. Nearly unbiased variable selection under minimax concave penalty. *The Annals of Statistics*, 38(2):894–942, 2010.

[Zhou, 2012] Zhi-Hua Zhou. *Ensemble methods: foundations and algorithms*. Chapman and Hall/CRC, 2012.

[Zhu *et al.*, 2011] Li-Ping Zhu, Lexin Li, Runze Li, and Li-Xing Zhu. Model-free feature screening for ultrahigh-dimensional data. *Journal of the American Statistical Association*, 106(496):1464–1475, 2011.

[Zou and Hastie, 2005] Hui Zou and Trevor Hastie. Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 67(2):301–320, 2005.

[Zou, 2006] Hui Zou. The adaptive lasso and its oracle properties. *Journal of the American statistical association*, 101(476):1418–1429, 2006.