Supplementary Materials of "An Ensemble Method for Feature Screening"

Xi Wu[1], Shifeng Xiong[2], and Weiyan Mu[3]

[1] National Cancer Center/National Clinical Research Center for Cancer/Cancer Hospital, Chinese Academy of Medical Sciences and Peking Union Medical College, Beijing 100021, China

[2] NCMIS, KLSC, Academy of Mathematics and Systems Science
Chinese Academy of Sciences, Beijing 100190, China

[3] School of Science,
Beijing University of Civil Engineering and Architecture, Beijing 100044, China

Abstract The Supplementary Materials are organized as follows. Section A presents simulation results with analytic functions for regression and classification. Section B provides all proofs.

A Simulations for regression and classification

A.1 Regression

Here we also consider the test functions in Section 5.1, but assign random errors $\varepsilon_i \sim N(0, \sigma^2)$ in the responses. The data of predictors are generated from the uniform distribution (UD), $\mathbf{X}_1, \ldots, \mathbf{X}_n$ identically and independently distributed from $U(0,1)^p$, or the normal distribution (ND), $\mathbf{X}_1, \ldots, \mathbf{X}_n$ identically and independently distributed from $N(\mathbf{1}_p/2, \mathbf{\Sigma})$, where $\mathbf{\Sigma} = (\sigma_{ij})_{i,j=1,\ldots,p}$ with $\sigma_{ii} = 1$ for $i = 1, \ldots, p$ and $\sigma_{ij} = \rho$ for $i \neq j$. For each test function, the selections of n, p, p_0, σ, ρ, and the distribution of the predictors are displayed in Table S1.

The coverage rates that the selected subset include the true submodel over 1000 repetitions are given in Table S1. We also conduct a simulation for equation (3) in Exam-

Table S1: Coverage rates in regression

function (I), $\sigma = 1$, UD

	$n = 100, p = 150$		$n = 200, p = 500$		$n = 400, p = 2000$	
	$p_0 = 5$	$p_0 = 10$	$p_0 = 5$	$p_0 = 10$	$p_0 = 5$	$p_0 = 10$
DC-SIS	0.234	0.004	0.242	0.016	0.471	0.060
MDC-SIS	0.260	0.006	0.263	0.039	0.537	0.081
LSLB	0.730	0.295	0.876	0.551	0.993	0.811
LSQB	0.833	0.742	0.955	0.949	0.998	0.963
ensemble ($\delta = 0.6$)	0.807	0.742	0.954	0.949	0.998	0.963
ensemble ($\delta = 0.7$)	0.807	0.742	0.953	0.945	0.998	0.961
ensemble ($\delta = 0.8$)	0.807	0.742	0.950	0.945	0.990	0.955

function (II), $\sigma = 1/4$, UD

	$n = 100, p = 150$		$n = 200, p = 500$		$n = 400, p = 2000$	
	$p_0 = 5$	$p_0 = 10$	$p_0 = 5$	$p_0 = 10$	$p_0 = 5$	$p_0 = 10$
DC-SIS	0.884	0.018	0.999	0.287	1.000	0.882
MDC-SIS	0.915	0.024	1.000	0.389	1.000	0.902
LSLB	0.975	0.086	1.000	0.613	1.000	0.988
LSQB	0.655	0.004	0.974	0.108	1.000	0.634
ensemble ($\delta = 0.6$)	0.979	0.087	1.000	0.610	1.000	0.988
ensemble ($\delta = 0.7$)	0.973	0.086	1.000	0.589	1.000	0.985
ensemble ($\delta = 0.8$)	0.919	0.030	1.000	0.324	1.000	0.890

function (III), $\sigma = 1/10$, UD

	$n = 100, p = 150$		$n = 200, p = 500$		$n = 400, p = 2000$	
	$p_0 = 5$	$p_0 = 10$	$p_0 = 5$	$p_0 = 10$	$p_0 = 5$	$p_0 = 10$
DC-SIS	0.895	0.062	0.995	0.582	1.000	0.966
MDC-SIS	0.924	0.100	0.995	0.687	1.000	0.970
LSLB	0.980	0.494	1.000	0.998	1.000	0.997
LSQB	0.998	0.668	1.000	1.000	1.000	1.000
ensemble ($\delta = 0.6$)	0.995	0.629	1.000	0.999	1.000	1.000
ensemble ($\delta = 0.7$)	0.991	0.628	1.000	0.998	1.000	1.000
ensemble ($\delta = 0.8$)	0.945	0.497	1.000	0.865	1.000	0.991

function (IV), $\sigma = 1/2$, $p_0 = 3$, ND

	$n = 100, p = 100$		$n = 200, p = 300$		$n = 400, p = 800$	
	$\rho = 0$	$\rho = 0.5$	$\rho = 0$	$\rho = 0.5$	$\rho = 0$	$\rho = 0.5$
DC-SIS	0.719	0.620	0.879	0.709	0.996	0.883
MDC-SIS	0.603	0.514	0.786	0.592	0.933	0.788
LSLB	0.593	0.177	0.826	0.354	0.952	0.476
LSQB	0.499	0.660	0.742	0.772	0.887	0.859
ensemble ($\delta = 0.6$)	0.763	0.706	0.896	0.772	0.983	0.879
ensemble ($\delta = 0.7$)	0.749	0.706	0.879	0.772	0.991	0.881
ensemble ($\delta = 0.8$)	0.736	0.706	0.879	0.772	0.999	0.883

function (V), $\sigma = 1/10$, $p_0 = 3$, ND

	$n = 100, p = 100$		$n = 200, p = 300$		$n = 400, p = 800$	
	$\rho = 0$	$\rho = 0.5$	$\rho = 0$	$\rho = 0.5$	$\rho = 0$	$\rho = 0.5$
DC-SIS	0.927	0.952	1.000	0.995	1.000	1.000
MDC-SIS	0.125	0.870	0.083	0.981	0.040	0.998
LSLB	0.024	0.656	0.010	0.906	0.012	1.000
LSQB	0.023	0.094	0.010	0.064	0.001	0.081
ensemble ($\delta = 0.6$)	0.903	0.800	0.986	0.829	1.000	0.973
ensemble ($\delta = 0.7$)	0.923	0.814	1.000	0.932	1.000	0.993
ensemble ($\delta = 0.8$)	0.927	0.887	1.000	0.986	1.000	1.000

Table S2: Coverage rates in classification

	$n = 100, p = 100$		$n = 200, p = 400$		$n = 400, p = 1000$	
	$p_0 = 3$	$p_0 = 5$	$p_0 = 3$	$p_0 = 5$	$p_0 = 3$	$p_0 = 5$
			Case (A)			
DC-SIS	0.883	0.301	0.973	0.491	1.000	0.920
MDC-SIS	0.883	0.301	0.973	0.492	1.000	0.920
MV-SIS	0.870	0.276	0.968	0.473	1.000	0.899
LSLB	0.935	0.562	0.998	0.913	1.000	0.995
LSQB	0.429	0.013	0.583	0.017	0.855	0.065
ensemble ($\delta = 0.6$)	0.918	0.557	0.987	0.913	1.000	0.991
ensemble ($\delta = 0.7$)	0.919	0.557	0.987	0.913	1.000	0.991
ensemble ($\delta = 0.8$)	0.917	0.557	0.987	0.911	1.000	0.988
			Case (B)			
DC-SIS	0.779	0.107	0.884	0.184	0.987	0.461
MDC-SIS	0.779	0.107	0.884	0.184	0.987	0.461
MV-SIS	0.735	0.122	0.828	0.192	0.965	0.427
LSLB	0.331	0.003	0.436	0.001	0.728	0.033
LSQB	0.981	0.261	1.000	0.532	1.000	0.933
ensemble ($\delta = 0.6$)	0.964	0.226	0.988	0.428	1.000	0.866
ensemble ($\delta = 0.7$)	0.964	0.226	0.988	0.428	1.000	0.866
ensemble ($\delta = 0.8$)	0.962	0.226	0.987	0.426	1.000	0.866
			Case (C)			
DC-SIS	0.814	0.073	0.924	0.228	0.994	0.468
MDC-SIS	0.814	0.073	0.924	0.228	0.994	0.468
MV-SIS	0.767	0.083	0.865	0.227	0.981	0.437
LSLB	0.395	0.000	0.527	0.005	0.820	0.014
LSQB	0.994	0.158	0.996	0.608	1.000	0.923
ensemble ($\delta = 0.6$)	0.978	0.128	0.993	0.472	0.997	0.822
ensemble ($\delta = 0.7$)	0.978	0.128	0.993	0.472	0.997	0.822
ensemble ($\delta = 0.8$)	0.978	0.128	0.993	0.471	0.995	0.822
			Case (D)			
DC-SIS	0.750	0.078	0.971	0.205	1.000	0.661
MDC-SIS	0.750	0.078	0.971	0.205	1.000	0.661
MV-SIS	0.743	0.081	0.966	0.203	1.000	0.615
LSLB	0.400	0.022	0.668	0.052	0.983	0.303
LSQB	0.664	0.063	0.944	0.225	1.000	0.597
ensemble ($\delta = 0.6$)	0.746	0.081	0.944	0.260	1.000	0.657
ensemble ($\delta = 0.7$)	0.754	0.081	0.952	0.256	1.000	0.696
ensemble ($\delta = 0.8$)	0.755	0.084	0.952	0.256	1.000	0.701
			Case (E)			
DC-SIS	0.778	0.130	0.916	0.184	0.993	0.317
MDC-SIS	0.778	0.130	0.916	0.184	0.993	0.317
MV-SIS	0.767	0.127	0.912	0.156	0.989	0.306
LSLB	0.992	0.321	1.000	0.407	1.000	0.661
LSQB	0.853	0.057	0.927	0.048	1.000	0.139
ensemble ($\delta = 0.6$)	0.989	0.319	0.992	0.401	1.000	0.658
ensemble ($\delta = 0.7$)	0.989	0.319	0.992	0.401	1.000	0.658
ensemble ($\delta = 0.8$)	0.989	0.319	0.992	0.401	1.000	0.657

ple 2 in Section 3.1 where the SIS-type methods fail. In this example of $p_0 = 2$, let $(n, p) = (100, 200)$ with 198 noisy predictors from $U(0,1)^{198}$, and the coverage rates of correct screening over 1000 repetitions of DC-SIS, MDC-SIS, LSLB, LSQB, and the ensemble methods with $\delta = 0.6, 0.7$, and 0.8 are 0.094, 0.101, 0.995, 0.922, 0.995, 0.995, and 0.892, respectively. From these simulation results, main findings are similar to those in the interpolation cases. The performance of ensemble screening is satisfactory even when the SIS-type methods or the linear screening methods fail. A different finding from the interpolation cases is that the selection of δ influences the performance of ensemble screening. As shown in Section 3.2, the linear screening methods may perform poorly when the predictors are correlated, and thus we should select relative large δ in ensemble screening. From Table S1 we can see that $\delta = 0.8$ is a good choice for such cases.

A.2 Classification

We now consider binary response $y \in \{0, 1\}$ and the following five cases.

Case (A): Generate $\mathbf{X}_1, \ldots, \mathbf{X}_n$ as the ND in Section A.1. $\rho = 0.5$. $y_1, \ldots, y_n$ are independently generated from the Bernoulli distribution $B(1, \mathrm{pr}(\mathbf{X}_i))$, where $\mathrm{pr}(\mathbf{X}_i) = \exp(-4 + 2\sum_{j=1}^{p_0} X_{ij}) / \big[1 + \exp(-4 + 2\sum_{j=1}^{p_0} X_{ij})\big]$

Case (B): Generate $\mathbf{X}_1, \ldots, \mathbf{X}_n$ as the ND in Case (A) but let $\rho = 0.1$. $y_1, \ldots, y_n$ are independently generated from the Bernoulli distribution $B(1, \mathrm{pr}(\mathbf{X}_i))$, where $\mathrm{pr}(\mathbf{X}_i) = \exp(f(X_{i1}, \ldots, X_{ip_0}) - 5) / \big[1 + \exp(f(X_{i1}, \ldots, X_{ip_0}) - 5)\big]$, where f is the weighted sphere function in Section 5.1.

Case (C): Generate the data in the same manner as in Case (B) except let $\mathrm{pr}(\mathbf{X}_i) = \Phi(f(X_{i1}, \ldots, X_{ip_0}) - 5)$.

Case (D): Generate $\mathbf{X}_1, \ldots, \mathbf{X}_n$ i.i.d. $\sim 0.5N(-\mathbf{1}_p, \mathbf{\Sigma}) + 0.5N(\mathbf{1}_p, \mathbf{I}_p)$ with $\rho = 0.2$ in $\mathbf{\Sigma}$. $y_1, \ldots, y_n$ are independently generated from the Bernoulli distribution $B(1, \mathrm{pr}(\mathbf{X}_i))$, where $\mathrm{pr}(\mathbf{X}_i) = \exp(f(X_{i1}, \ldots, X_{ip_0}) - 5) / \big[1 + \exp(f(X_{i1}, \ldots, X_{ip_0}) - 5)\big]$ and f is Ackley's model.

Case (E): Generate the data in the same manner as in Case (D) except let $\text{pr}(\mathbf{X}_i) = \Phi(f(X_{i1}, \ldots, X_{ip_0}) - 100)$ and f is Zakharov's model.

Combinations of (n, p, p_0) used in the simulations are displayed in Table S2. Besides the methods compared in the previous two subsections, we add Cui, Li, and Zhong (2015)'s MV-SIS method which is a model-free feature screening approach for high dimensional discriminant analysis. The simulation results based on 1000 repetitions are presented in Table S2. We can see that, similar to the previous two subsections, the coverage rates of the proposed methods are close to the best one from the existing methods in most cases. In particular, they have much better overall performance than the SIS-type methods.

B Proofs

Lemma B.1. *Under Assumption 1, for $\mathcal{A} \subset \mathbb{Z}_p$, $f_{\mathcal{A}} = f_{\mathcal{A}\cap\mathcal{A}_0}$.*

Proof. To simplify the notation, let $p = 3$, $p_0 = 2$, and $\mathcal{A} = \{1, 3\}$. The proof of the general case is almost the same. Let X_1, X_2, X_3 be independent random variables distributed from $U(0, 1)$. By Assumption 1, we have $f_{\mathcal{A}}(x_1, x_3) = E(f(X_1, X_2, X_3)|X_1 = x_1, X_3 = x_3) = E(f_{\mathcal{A}_0}(X_1, X_2)|X_1 = x_1, X_3 = x_3) = E(f_{\mathcal{A}_0}(X_1, X_2)|X_1 = x_1) = E(f(X_1, X_2, X_3)|X_1 = x_1) = f_{\mathcal{A}\cap\mathcal{A}_0}(x_1)$. □

Proof. of Proposition 1. By Lemma 1 and Assumption 2, $f_{\mathcal{A}} = f_{\mathcal{A}\cap\mathcal{A}_0} = f_{\mathcal{A}_0}$, which implies $\mathcal{A} = \mathcal{A}_0$. □

It is obvious to obtain the following lemma.

Lemma B.2. *Under Assumption 3, for $f, g \in L^2(\mathcal{I}^p) \cap L^2_P(\mathcal{I}^p)$, $f =_P g$ if and only if $f = g$.*

Proof. of Proposition 2. (i)⇒(ii): Note that $f_{\mathcal{A}_0} \in L^2_P(\mathcal{I}^{\mathcal{A}_0})$ and $f = f_{\mathcal{A}_0}$. It follows that $f_{\mathcal{A}_0} = f_{\mathcal{A}_0,P}$, which implies $f =_P f_{\mathcal{A}_0,P}$ by Lemma B.2. Similarly we can prove $f_{\mathcal{A}} = f_{\mathcal{A},P}$ for $\mathcal{A} \subsetneqq \mathcal{A}_0$. Therefore, $f_{\mathcal{A},P} \neq f_{\mathcal{A}_0,P}$, which implies $f_{\mathcal{A},P} \neq_P f_{\mathcal{A}_0,P}$ by Lemma B.2.
(ii)⇒(i): Since $f_{\mathcal{A}_0,P} \in L^2(\mathcal{I}^{\mathcal{A}_0})$, we have $f_{\mathcal{A}_0,P} = f_{\mathcal{A}_0}$, which implies $f = f_{\mathcal{A}_0}$. Similarly we can prove $f_{\mathcal{A},P} = f_{\mathcal{A}}$ for $\mathcal{A} \subsetneqq \mathcal{A}_0$, and thus $f_{\mathcal{A}_0} \neq f_{\mathcal{A}}$. □

Lemma B.3. *Under Assumptions 1-4,*

$$\beta_j(P) = \left(\int_{\mathcal{I}^p} b(x_j) f_{\mathcal{A}_0}(\mathbf{x}_{\mathcal{A}_0}) \psi(\mathbf{x}) d\mathbf{x} - \zeta \int_{\mathcal{I}^p} f_{\mathcal{A}_0}(\mathbf{x}_{\mathcal{A}_0}) \psi(\mathbf{x}) d\mathbf{x} \right) / (\eta - \zeta^2) \quad \textit{for } j = 1, \ldots, p_0,$$
$$\beta_j(P) = 0 \quad \textit{for } j = p_0 + 1, \ldots, p,$$

where $\zeta = \int_{\mathcal{I}} b(x)\varphi(x)dx$ *and* $\eta = \int_{\mathcal{I}} b(x)^2 \varphi(x)dx$.

Proof. Under Assumption 4, $\mathbf{u} = \zeta \mathbf{1}_p$ and $\mathbf{\Sigma} = (\sigma_{ij})_{p \times p}$ with $\sigma_{ii} = \eta$ for $i = 1, \ldots, p$ and $\sigma_{ij} = \zeta^2$ for $i \neq j$. Some algebra gives

$$\begin{pmatrix} 1 & \mathbf{u}' \\ \mathbf{u} & \mathbf{\Sigma} \end{pmatrix}^{-1} = \begin{pmatrix} 1 + p\zeta^2/(\eta - \zeta^2) & -\zeta \mathbf{1}_p'/(\eta - \zeta^2) \\ -\zeta \mathbf{1}_p/(\eta - \zeta^2) & \mathbf{I}_p/(\eta - \zeta^2) \end{pmatrix},$$

where $\mathbf{I}_p$ denotes the $p \times p$ identity matrix. By (5),

$$\boldsymbol{\beta}(P) = -\frac{\zeta}{\eta - \zeta^2} \int_{\mathcal{I}^p} f(\mathbf{x}) \psi(\mathbf{x}) d\mathbf{x} + \frac{\mathbf{v}}{\eta - \zeta^2}. \qquad \text{(B.1)}$$

The result of the lemma for $j = 1, \ldots, p_0$ follows from (B.1) and Assumption 1. For $j = p_0 + 1, \ldots, p$,

$$\int_{\mathcal{I}^{p_0+1}} b(x_j) f_{\mathcal{A}_0}(x_1, \ldots, x_{p_0}) \varphi(x_1) \cdots \varphi(x_{p_0}) \varphi(x_j) dx_1 \cdots dx_{p_0} dx_j$$
$$= \int_{\mathcal{I}} b(x_j)\varphi(x_j)dx_j \int_{\mathcal{I}^{p_0}} f_{\mathcal{A}_0}(x_1, \ldots, x_{p_0}) \varphi(x_1) \cdots \varphi(x_{p_0}) dx_1 \cdots dx_{p_0} = \zeta \int_{\mathcal{I}^{p_0}} f_{\mathcal{A}_0}(\mathbf{x}) \psi(\mathbf{x}) d\mathbf{x},$$

which implies $\beta_j(P) = 0$. □

Proof. of Proposition 4. This proposition follows from Lemma B.3 and Assumption 4. □

References

Cui, H., Li, R., and Zhong, W. (2015). Model-free feature screening for ultrahigh dimensional discriminant analysis, *Journal of the American Statistical Association*, 110: 630–641.