Biometrics 000, 1–19 DOI: 000
000 000

Supporting Information for "A General Framework of Nonparametric Feature Selection in High-Dimensional Data" by Yu, Wang and Zeng

Hang Yu

Department of Statistics and Operation Research, University of North Carolina, Chapel Hill, NC 27599

email: hangyu@live.unc.edu

and

Yuanjia Wang

Department of Biostatistics, Columbia University, New York, NY 10032.

email: yw2016@cumc.columbia.edu

and

Donglin Zeng

Department of Biostatistics, University of North Carolina, Chapel Hill, NC 27599

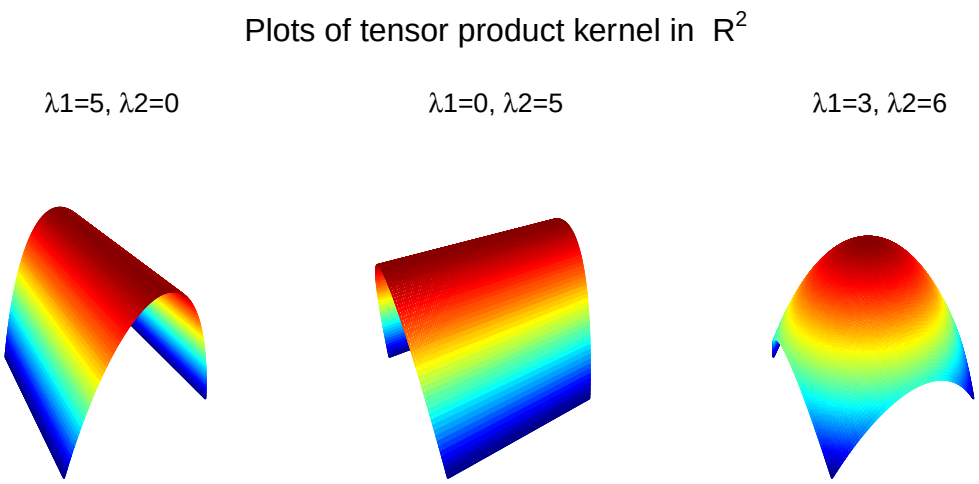
email: dzeng@email.unc.edu

This paper has been submitted for consideration for publication in *Biometrics*

Web Appendix A: Plot of The Product Kernel Function

Web figure 1 shows the product kernel function with 2-dimensional feature variables.

Web Figure 1. Tensor Product Kernel in $\mathcal{R}^2$



Note: The bandwidth $\sigma_n = 2$ and each kernel is centered at 0.

Web Appendix B: Theoretical Properties

In this Appendix, we present some theoretical properties of our proposed method. Since our proposed kernel function is new, we first provide two theorems that describe the properties for the RKHS generated by this kernel function. In the first theorem, we show that this space is dense in the $L_2(P)$ subspace consisting of all measurable functions that only depend on the feature variables for which $\lambda_m \neq 0$ in the kernel function. In the second theorem, we obtain the entropy number for the unit ball in this space. Both theorems are necessary to establish the asymptotic properties of the proposed estimator for $f(\mathbf{X})$ as given in the previous section.

To state our results, we define $f_0(\boldsymbol{X})$ as the Bayesian prediction function, which is assumed to be unique. That is, $E\{l(Y,f)\}$ attains its minimum when $f = f_0$. We assume that feature variables $X_1, X_2, \cdots, X_q$ are important in terms that $f_0(\boldsymbol{X})$ is only a function of $X_1, X_2, ..., X_q$ and for any $1 \leqslant s \leqslant q$,

$$E\left[\left[f_0(\boldsymbol{X}) - E\left\{f_0(\boldsymbol{X})\middle| X_1, X_2, X_{s-1}, X_{s+1}\cdots, X_q\right\}\right]^2\right] > 0.$$

Finally, we let $d_2(f_0, \mathcal{H}_{\boldsymbol{\lambda},\sigma_n})$ denote the $L_2(P)$-distance between f_0 and the RKHS generated by $\kappa_{\boldsymbol{\lambda},\sigma_n}$.

THEOREM 1: *We assume $\sigma_n \to 0$ as $n \to \infty$. For a vector $\boldsymbol{\lambda}_n = (\lambda_{n1}, ..., \lambda_{np_n})$ with $\lambda_{nm} \geqslant 0$ for $m = 1, ..., p_n$, the following results hold:*

(i) If $\lambda_{nm} > 0$ for $m = 1, ..., q$, i.e., λ_n's that are associated with the important features are strictly positive, then $d_2(f_0, \mathcal{H}_{\boldsymbol{\lambda}_n,\sigma_n}) \to 0$.

(ii) If for some $m \leqslant q$, $\lambda_{nm} = 0$, then $\liminf d_2(f_0, \mathcal{H}_{\boldsymbol{\lambda}_n,\sigma_n}) > 0$.

Note: The Theorem holds for $\boldsymbol{\lambda}$ whose value depends on n and denoted as $\boldsymbol{\lambda}_n$.

Proof. To prove (i), we first note that after expansion, $\kappa_{\boldsymbol{\lambda}_n,\sigma_n}(\mathbf{X}, \widetilde{\mathbf{X}})$ is the summation of a number of Gaussian kernels. In particular, one term of this summation is

$$\left\{\lambda_{n1}\lambda_{n2}\cdots\lambda_{nq}\kappa_{\sigma_n}(X_1, \widetilde{X}_1)\kappa_{\sigma_n}(X_2, \widetilde{X}_2)\cdots\kappa_{\sigma_n}(X_q, \widetilde{X}_q)\right\},$$

where $\kappa_\sigma(x,y) = \exp\{-(x-y)^2/\sigma^2\}$. Since $\lambda_{n1}, ..., \lambda_{nq} > 0$, the kernel function associated with this term is proportional to the Gaussian kernel in the space of $(X_1, \cdots, X_q)$ with bandwidth σ_n for each domain k. Therefore, the closure of the RKHS generated by $\kappa_{\boldsymbol{\lambda}_n,\sigma_n}$ includes the RKHS generated by the Gaussian kernel in the space of $(X_1, \cdots, X_q)$. The result in (i) holds since the latter is asymptotically dense in the subspace of $L_2(P)$ consisting of any functions depending on $(x_1, ..., x_q)$.

To prove (ii), if $\lambda_m = 0$, then it is clear that any function in $\mathcal{H}_{\boldsymbol{\lambda}_n,\sigma_n}$ only depends on the

feature variables except X_m. Therefore,

$$\mathcal{H}_{\boldsymbol{\lambda}_n,\sigma_n} \subset \{g(\boldsymbol{X}_{-m}) : g \in L_2(P)\},$$

where $\boldsymbol{X}_{-m}$ denotes all the feature variables excluding X_m. On the other hand, the projection of f_0 on the latter space is $E(f_0|\boldsymbol{X}_{-m})$. Therefore,

$$\liminf d(f_0, \mathcal{H}_{\boldsymbol{\lambda}_n,\sigma_n}) \geqslant d\{f_0, E(f_0|\boldsymbol{X}_{-m})\} > 0$$

since X_m is one important variable for f_0. We obtain the result.

Our next theorem studies the bracket covering number for a unit ball in $\mathcal{H}_{\boldsymbol{\lambda}_n,\sigma_n}$. We consider $\mathcal{B}_n$ as the unit ball in $\mathcal{H}_{\boldsymbol{\lambda}_n,\sigma_n}$, i.e., $\mathcal{B}_n \equiv \left\{f(\boldsymbol{x}) : \|f\|_{\mathcal{H}_{\boldsymbol{\lambda}_n,\sigma_n}} \leqslant 1\right\}$, Then the ϵ-bracket covering number for $\mathcal{B}_n$, denoted as $N_{[]}(\epsilon, \mathcal{B}_n, \|\cdot\|_{L_2(P)})$, is defined as the minimal number of pairs $[l(\boldsymbol{x}), u(\boldsymbol{x})]$ such that any function $\|u(\mathbf{X}) - l(\mathbf{X})\|_{L_2(P)} \leqslant \epsilon$ and any function f in $\mathcal{B}_n$ is between one pair, i.e., $l(\boldsymbol{x}) \leqslant f(\boldsymbol{x}) \leqslant u(\boldsymbol{x})$.

THEOREM 2: *For a vector $\boldsymbol{\lambda}_n = (\lambda_{n1}, ..., \lambda_{np_n})$ such that λ_{nm} is uniformly bounded by a constant M for $m = 1, ..., q$ and $\lambda_{n(q+1)} = ... = \lambda_{np_n} = 0$, it holds*

$$\log \mathcal{N}_{[]}(\epsilon, \mathcal{B}_n, \|\cdot\|_{L_2(P)}) \leqslant C\sigma_n^{-(1-v/4)q}\epsilon^{-v},$$

where v is any constant within $(0, 2)$ and C only depends on M and q.

Proof. For any $f \in \mathcal{B}_n$ with form

$$f(\boldsymbol{x}) = \sum_{i=1}^{\infty} \alpha_i \kappa_{\boldsymbol{\lambda}_n,\sigma_n}(\boldsymbol{x}, \boldsymbol{x}_i),$$

where $\boldsymbol{x}_1, \boldsymbol{x}_2, ...$ are a sequence of given points. Using the expansion of $\kappa_{\boldsymbol{\lambda}_n,\sigma_n}$, we have

$$\begin{aligned} f(\boldsymbol{x}) &= \sum_{\{k_1,...,k_s\} \subset \{1,...,q\} \cup \phi} \lambda_{nk_1} \cdots \lambda_{nk_s} \sum_{i=1}^{\infty} \alpha_i \exp\left\{-\frac{(x_{ik_1} - x_{k_1})^2 + \cdots + (x_{ik_s} - x_{k_s})^2}{\sigma_n^2}\right\} \\ &= \sum_{\{k_1,...,k_s\} \subset \{1,...,q\} \cup \phi} \sqrt{\lambda_{nk_1} \cdots \lambda_{nk_s}} f_{k_1...k_s}(\mathbf{x}), \end{aligned}$$

where x_{ik} and x_k are respectively the kth component of $\boldsymbol{x}_i$ and $\boldsymbol{x}$, and

$$f_{k_1...k_s}(\mathbf{x}) = \sum_{i=1}^{\infty} \alpha_i \sqrt{\lambda_{nk_1} \cdots \lambda_{nk_s}} \exp\left\{-\frac{(x_{ik_1} - x_{k_1})^2 + \cdots + (x_{ik_s} - x_{k_s})^2}{\sigma_n^2}\right\}.$$

Here, if the index set if empty, then the exponential part in the summation is replaced by 1.

Clearly, if we denote $\mathcal{H}_{k_1\ldots k_s}$ as the reproducing kernel Hilbert space generated by the Gaussian kernel $\exp\left[-\{(\widetilde{x}_{k_1}-x_{k_1})^2+\cdots+(\widetilde{x}_{k_s}-x_{k_s})^2\}/\sigma_n^2\right]$, then $f_{k_1\ldots k_s}(\mathbf{x})\in\mathcal{H}_{k_1\ldots k_s}$ and moreover,

$$\begin{aligned}
\|f\|^2_{\mathcal{H}_{\boldsymbol{\lambda}_n,\sigma_n}} &= \sum_{i=1}^{\infty}\sum_{j=1}^{\infty}\alpha_i\alpha_j\kappa_{\boldsymbol{\lambda}_n,\sigma_n}(\mathbf{x}_i,\mathbf{x}_j)\\
&= \sum_{i=1}^{\infty}\sum_{j=1}^{\infty}\alpha_i\alpha_j\sum_{\{k_1,\ldots,k_s\}\subset\{1,\ldots,q\}\cup\phi}\lambda_{nk_1}\cdots\lambda_{nk_s}\exp\left\{-\frac{(x_{ik_1}-x_{jk_1})^2+\cdots+(x_{ik_s}-x_{jk_s})^2}{\sigma_n^2}\right\}\\
&= \sum_{\{k_1,\ldots,k_s\}\subset\{1,\ldots,q\}\cup\phi}\sum_{i=1}^{\infty}\sum_{j=1}^{\infty}\alpha_i\alpha_j\lambda_{nk_1}\cdots\lambda_{nk_s}\exp\left\{-\frac{(x_{ik_1}-x_{jk_1})^2+\cdots+(x_{ik_s}-x_{jk_s})^2}{\sigma_n^2}\right\}\\
&= \sum_{\{k_1,\ldots,k_s\}\subset\{1,\ldots,q\}\cup\phi}\|f_{k_1\ldots k_s}\|^2_{\mathcal{H}_{k_1\ldots k_s}}.
\end{aligned}$$

Thus, $\|f\|_{\mathcal{H}_{\boldsymbol{\lambda}_n,\sigma_n}}\leqslant 1$ implies $\|f_{k_1\ldots k_s}\|_{\mathcal{H}_{k_1\ldots k_s}}\leqslant 1$ for any $k_1,\ldots,k_s$.

Consequently, since such f is dense in $\mathcal{B}_n$, we conclude

$$\mathcal{B}_n\subseteq\overline{\left\{\sum_{\{k_1,\ldots,k_s\}\subset\{1,\ldots,q\}\cup\phi}f_{k_1\ldots k_s}(\boldsymbol{x})\sqrt{\lambda_{nk_1}\cdots\lambda_{nk_s}}:\|f_{k_1\ldots k_s}\|^2_{\mathcal{H}_{k_1\ldots k_s}}\leqslant 1\right\}}.$$

Thus, there exists a constant C only depending on M and q such that

$$\log\mathcal{N}_{[]}(2^qM^{q/2}\epsilon,\mathcal{B}_n,\|\cdot\|_{L_2(P)})$$

$$\leqslant\sum_{\{k_1,\ldots,k_s\}\subset\{1,\ldots,q\}\cup\phi}\log\mathcal{N}_{[]}(\epsilon,\{f_{k_1\ldots k_s}(\boldsymbol{x}),\|f_{k_1\ldots k_s}\|^2_{\mathcal{H}_{k_1\ldots k_s}}\leqslant 1\},\|\cdot\|_{L_2(P)})$$

According to (Steinwart and Scovel (2007)), we know

$$\log\mathcal{N}_{[]}(\epsilon,\{f_{k_1\ldots k_s}(\boldsymbol{x}),\|f_{k_1\ldots k_s}\|^2_{\mathcal{H}_{k_1\ldots k_s}}\leqslant 1\},\|\cdot\|_{L_2(P)})\leqslant C\sigma_n^{-(1-v/4)s}\epsilon^{-v},$$

for any constant $v\in(0,2)$ and a constant C only depending on s. Therefore,

$$\log\mathcal{N}(\epsilon,\mathcal{B}_n,\|\cdot\|_{L_2(P)})\leqslant C(M,q)\sum_{\{k_1,\ldots,k_s\}\subset\{1,\ldots,q\}\cup\phi}\sigma_n^{-(1-v/4)s}\epsilon^{-v}\leqslant C(M,q)\sigma_n^{-(1-v/4)q}\epsilon^{-v}$$

for a constant $C(M,q)$. We have proved Theorem 2.

Our next theorem gives the main properties of the estimated prediction function. We show that the resulting prediction function from our method leads to Bayesian risk asymptotically. Moreover, with probability tending to one, the variable selection based on non-zero λ_n's is

oracle as if we knew which variables were important. Recall that $(\widehat{\boldsymbol{\lambda}_n}, \widehat{f})$ is the optimal solution of the objective function

$$L_n(\boldsymbol{\lambda}_n, f) = \mathbf{P}_n l\left\{Y, f(\mathbf{X})\right\} + \gamma_{1n}\|f\|^2_{\mathcal{H}_{\boldsymbol{\lambda}_n,\boldsymbol{\sigma}_n}} + \gamma_{2n}P(\boldsymbol{\lambda}_n),$$

where $P(\boldsymbol{\lambda}_n)$ is the truncated Lasso penalty for $\boldsymbol{\lambda}_n$. Equivalently, if we define for any $\boldsymbol{\lambda}_n$,

$$\widehat{f}_{\boldsymbol{\lambda}_n} = \arg\min_f L_n(\boldsymbol{\lambda}_n, f),$$

which exists due to the convexity of $L_n(\boldsymbol{\lambda}_n, f)$ in f, then $\widehat{\boldsymbol{\lambda}}$ minimizes $L_n(\widehat{\boldsymbol{\lambda}}_n, \widehat{f}_{\boldsymbol{\lambda}_n})$ and $\widehat{f} = \widehat{f}_{\widehat{\boldsymbol{\lambda}}_n}$.

For the main theorem, we assume $(Y, \mathbf{X})$ to have a bounded support and need the following conditions.

(C1). The loss function $l(y, f)$ is convex and is Lipschitz continuous with respect to f in any bounded set.

(C2). There exit $\delta > 0$ and a constant $c_1 > 0$ such that

$$E\left[l\left\{Y, f(\boldsymbol{X})\right\} - l\left\{Y, f_0(\boldsymbol{X})\right\}\right] \geqslant c_1\|f(\boldsymbol{X}) - f_0(\boldsymbol{X})\|^2_{L_2(P)}$$

whenever $E\left[l\left\{Y, f(\boldsymbol{X})\right\} - l\left\{Y, f_0(\boldsymbol{X})\right\}\right]$ is smaller than δ.

(C3). Assume $\|l_2(Y, f(\mathbf{X})) - l_2(Y, f_0(\mathbf{X}))\|_{L_2(P)} \leqslant c_2\|f(\boldsymbol{X}) - f_0(\boldsymbol{X})\|_{L_2(P)}$ for a constant c_2, where $l_2(y, x) = \partial l(y, x)/\partial x$.

(C4). For any $\widetilde{\boldsymbol{\lambda}_n} = (\lambda_{n1}, ..., \lambda_{np_n})$ such that $\lambda_{nk} = 0$ for $k > q$, let $\Lambda_{\max}(\mathbf{X}_{-q})$ and $\Lambda_{\min}(\mathbf{X}_{-q})$ be the largest and smallest eigenvalues of the matrix $\left(E\left\{K_{\widetilde{\boldsymbol{\lambda}}_n}(\mathbf{X}_j, \mathbf{X})K_{\widetilde{\boldsymbol{\lambda}}_n}(\mathbf{X}_l, \mathbf{X})|\mathbf{X}_{-q}\right\}\right)$ where $\mathbf{X}_{-q}$ denotes all unimportant variables. We assume that with probability one, there exists one constant c such that $\Lambda_{\max}(\mathbf{X}_{-q})/\Lambda_{\min}(\mathbf{X}_{-q}) \leqslant c\sigma_n^{-1/2}$ and $E\left\{\Lambda_{\min}(\mathbf{X}_{-q})\kappa_n(x, X_m)^2\right\} \leqslant c\sigma_n^{1/2}$ for any $m > q$.

(C5). Assume $\log p_n = o(n^{1-(2+q)\alpha_1-\alpha_2-\alpha_3})$. Moreover, we assume $\sigma_n = n^{-\alpha_1}$, $\gamma_{1n} = n^{-\alpha_2}$, $\gamma_{2n} = n^{-\alpha_3}$, where $\alpha_k > 0$ for $k = 1, 2, 3$ and they satisfy

(i) $1 - (2 + q)\alpha_1 - \alpha_2 > 0$

(ii) $0 < \alpha_3 < \min\left\{\frac{1}{4}(1 + \frac{\alpha_1 q}{2} + \alpha_2), 1 - (2 + q)\alpha_1 - \alpha_2, \frac{\alpha_1}{2}, \frac{\alpha_2}{2}\right\}$.

Conditions (C1)-(C3) give the assumptions for the loss functions. It can be verified that they hold for $l(y, f) = (y - f)^2$ for a continuous y and for $l(y, f) = \exp(-yf)$ for a binary y. Condition (C4) implies the equivalence between the Euclidean norm of the coefficients and the reproducing kernel Hilbert space norm, up to a scale proportional to $\sigma_n^{-1/2}$. The second half of the condition in (C4) holds automatically if the important variables are independent of the unimportant variable when $\Lambda_{\min}(\mathbf{X}_{-q})$ does not depend on $\mathbf{X}_{-q}$. We note that such a condition is analogue to the design matrix condition assumed in high dimensional linear model literature. Finally, condition (C5) allows the dimensionality of the feature variable to be ultra-high and imposes additional constraints for the choices of the bandwidth and two tuning parameters.

THEOREM 3: *Under Conditions (C1)-(C5), there exists a local minimizer* $\widehat{\boldsymbol{\lambda}}_n$ *for* $L_n(\boldsymbol{\lambda}_n, \widehat{f}_{\boldsymbol{\lambda}_n})$ *such that with probability tending to one,*

(a) $E\left\{l\big(Y, \widehat{f}_{\widehat{\boldsymbol{\lambda}}_n}\big)\right\}$ *converges to* $E\left\{l\big(Y, f_0\big)\right\}$.

(b) For $m = 1, ..., q, \widehat{\lambda}_{nm} > 0$.

(c) For $m = q+1, q+2, \cdots, p_n,\ \widehat{\lambda}_{nm} = 0$.

The first part of Theorem 3 implies that the loss of the estimated prediction function converges to the Bayes risk. The last two conclusions in Theorem 3 show that the proposed estimator has an oracle property, that the $\widehat{\lambda}_{nm}$'s associated with important feature variables should be non-zero, i.e., the estimated function does depend on important variables. More importantly, the proposed method can estimate the predicted function as if we knew which variables are important in the truth. The proof of Theorem 3(a) entails careful examination of the stochastic variability of $L_n(\boldsymbol{\lambda}_n, \widehat{f}_{\boldsymbol{\lambda}_n})$, for which we first establish a preliminary bound for $\widehat{f}_{\boldsymbol{\lambda}_n}$ and then appeal to some concentration inequalities for empirical processes with metric entropy as derived from Theorem 2. To prove Theorem 3(b) and (c) in the theorem, we examine the KKT conditions to show that the oracle estimators, i.e., λ_{nm} is known to

be zero for $m > q$, satisfies the KKT conditions with probability tending to one. Again, concentration inequalities for empirical processes are needed in technical arguments in the proof.

Proof. In the following proof, we use C to denote a constant that does not depend on n but may depend on q. We prove the main theorem based on Theorems 1 and 2 as given in Appendix. To prove Theorem 3, we consider a restricted space for $\boldsymbol{\lambda}_n$:

$$\mathcal{S} = \{\boldsymbol{\lambda}_n = (\lambda_{n1}, ..., \lambda_{nq}, \mathbf{0}) : \ M \geqslant \lambda_{nj} \geqslant 0 \text{ for } 1 \leqslant j \leqslant q\}.$$

That is, $\mathcal{S}$ is an oracle space for which we know which features are important. For any $\boldsymbol{\lambda}_n \in \mathcal{S}$, we define

$$\widehat{f}_{\boldsymbol{\lambda}_n} = \arg\min_f L_n(\boldsymbol{\lambda}_n, f),$$

where we recall

$$L_n(\boldsymbol{\lambda}_n, f) = \mathbf{P}_n \left[l \left\{ Y, f_{\boldsymbol{\lambda}_n}(\mathbf{X}) \right\} \right] + \gamma_{1n} \|f\|^2_{\mathcal{H}_{\boldsymbol{\lambda}_n, \sigma_n}} + \gamma_{2n} P(\boldsymbol{\lambda}_n).$$

Clearly, under the strictly convexity condition C1, $\widehat{f}_{\boldsymbol{\lambda}_n}$ exists and is unique. Finally, we define

$$\widetilde{\boldsymbol{\lambda}}_n = \arg\min_{\boldsymbol{\lambda}_n \in \mathcal{S}} L_n(\boldsymbol{\lambda}_n, \widehat{f}_{\boldsymbol{\lambda}_n}).$$

In many literature, $\widetilde{\boldsymbol{\lambda}}_n$ is called the oracle estimator for $\boldsymbol{\lambda}_n$ since we know which features are important.

The whole proof can be divided into three steps. First, we show that the oracle estimator $\widetilde{\boldsymbol{\lambda}}_n$ leads to the prediction function that attains the Bayesian risk asymptotically. Second, we use the first step result to prove that $\widetilde{\lambda}_{nm}$ is strictly positive for $m = 1, ..., q$ with probability tending one. In the last step, we show that $\widetilde{\boldsymbol{\lambda}}_n$ is a local minimizer for $L_n(\boldsymbol{\lambda}_n, \widehat{f}_{\boldsymbol{\lambda}_n})$ by verifying the KKT conditions. With all these results, Theorem 3 holds if we choose $\widehat{\boldsymbol{\lambda}}_n = \widetilde{\boldsymbol{\lambda}}_n$. For convenience of notation, we use C to denote any constant depending on q.

Step 1. We first show that with probability tending to 1, the prediction loss for $\widehat{f}_{\widetilde{\boldsymbol{\lambda}}_n}$ converges to the Bayesian risk, i.e., $E\left\{ l(Y, \widehat{f}_{\widetilde{\boldsymbol{\lambda}}_n}) \right\}$ converges to $E\left\{ l(Y, f_0) \right\}$. To this end, for a fixed

$\boldsymbol{\lambda}_n^* = (\lambda_1^*, \lambda_2^*, \cdots, \lambda_q^*, \mathbf{0})$, where $\lambda_j^* > 0, \text{for } 1 \leqslant j \leqslant q$, and set to be greater than $M/2$, which results in 0 penalty for important variables, from the proof of (i) in Theorem 1, there exits $f_n^* \in \mathcal{H}_{\boldsymbol{\lambda}_n^*, \sigma_n}$ which only depends on $X_1, ..., X_q$ such that $d_2(f_0, f_n^*) \to 0$. In fact, the proof of Theorem 1 shows that this function can be obtained from the reproducing kernel Hilbert space generated by the Gaussian kernel in the space of $(X_1, ..., X_q)$. Therefore, we can choose f_n^* such that $d_2(f_0, f_n^*) \leqslant C\sigma_n{}^{\frac{q}{2}}$ using the construction in Theorem 4.26 of Steinwart and Christmann (2008). Consequently, condition C2 implies $\mathbf{P}l(Y, f_n^*) - \mathbf{P}l(Y, f_0) \leqslant C\sigma_n^q$. Moreover, according to Lemma 5.15 of Steinwart and Christmann (2008),

$$\begin{aligned} & \inf_{f \in \mathcal{H}_{\boldsymbol{\lambda}_n^*, \sigma_n}} \left\{ \mathbf{P}l(Y, f) + \gamma_{1n} \|f\|^2_{\mathcal{H}_{\boldsymbol{\lambda}_n^*, \sigma_n}} \right\} - \mathbf{P}l(Y, f_n^*) \\ \leqslant & \inf_{f \in \mathcal{H}_{\boldsymbol{\lambda}_n^*, \sigma_n}} \left\{ \mathbf{P}l(Y, f) + \gamma_{1n} \|f\|^2_{\mathcal{H}_{\boldsymbol{\lambda}_n^*, \sigma_n}} \right\} - \inf_{f \in \mathcal{H}_{\boldsymbol{\lambda}_n^*, \sigma_n}} \mathbf{P}l(Y, f) \\ \leqslant & C\gamma_{1n}. \end{aligned}$$

So we obtain

$$\inf_{f \in \mathcal{H}_{\boldsymbol{\lambda}_n^*, \sigma_n}} \left\{ \mathbf{P}l(Y, f) + \gamma_{1n} \|f\|^2_{\mathcal{H}_{\boldsymbol{\lambda}_n^*, \sigma_n}} \right\} - \mathbf{P}l(Y, f_0) \leqslant C(\gamma_{1n} + \sigma_n^q). \tag{A.1}$$

On the other hand, by the definition of $(\widetilde{\boldsymbol{\lambda}}_n, \widehat{f}_{\widetilde{\boldsymbol{\lambda}}_n})$, we have

$$\begin{aligned} & \mathbf{P}_n l(Y, \widehat{f}_{\widetilde{\boldsymbol{\lambda}}_n}) + \gamma_{1n} \|\widehat{f}_{\widetilde{\boldsymbol{\lambda}}_n}\|^2_{\mathcal{H}_{\widetilde{\boldsymbol{\lambda}}_n, \sigma_n}} + \gamma_{2n} P(\widetilde{\boldsymbol{\lambda}}_n) \\ \leqslant & \quad \mathbf{P}_n l(Y, \widetilde{f}_n^*) + \gamma_{1n} \|\widetilde{f}_n^*\|^2_{\mathcal{H}_{\boldsymbol{\lambda}_n^*, \sigma_n}} + \gamma_{2n} P(\boldsymbol{\lambda}_n^*), \end{aligned}$$

where $\widetilde{f}_n^*$ is the function in $\mathcal{H}_{\boldsymbol{\lambda}_n^*, \sigma_n}$ that attains the minimum $\left\{ \mathbf{P}l(Y, f) + \gamma_{1n} \|f\|^2_{\mathcal{H}_{\boldsymbol{\lambda}_n^*, \sigma_n}} \right\}$. Equivalently,

$$\begin{aligned} & (\mathbf{P}_n - \mathbf{P}) l(Y, \widehat{f}_{\widetilde{\boldsymbol{\lambda}}_n}) + \gamma_{1n} \|\widehat{f}\|^2_{\mathcal{H}_{\widetilde{\boldsymbol{\lambda}}_n, \sigma_n}} + \mathbf{P}l(Y, \widehat{f}_{\widetilde{\boldsymbol{\lambda}}_n}) + \gamma_{2n} P(\widetilde{\boldsymbol{\lambda}}_n) \\ \leqslant & (\mathbf{P}_n - \mathbf{P}) l(Y, \widetilde{f}_n^*) + \gamma_{1n} \|\widetilde{f}_n^*\|^2_{\mathcal{H}_{\boldsymbol{\lambda}_n^*, \sigma_n}} + \mathbf{P}l(Y, \widetilde{f}_n^*) + \gamma_{2n} P(\boldsymbol{\lambda}_n^*). \end{aligned}$$

Using (A.1), this gives

$$\begin{aligned} & (\mathbf{P}_n - \mathbf{P}) l(Y, \widehat{f}_{\widetilde{\boldsymbol{\lambda}}_n}) + \gamma_{1n} \|\widehat{f}\|^2_{\mathcal{H}_{\widetilde{\boldsymbol{\lambda}}_n, \sigma_n}} + \mathbf{P}l(Y, \widehat{f}_{\widetilde{\boldsymbol{\lambda}}_n}) + \gamma_{2n} P(\widetilde{\boldsymbol{\lambda}}_n) \\ \leqslant & (\mathbf{P}_n - \mathbf{P}) l(Y, \widetilde{f}_n^*) + \mathbf{P}l(Y, f_0) + \gamma_{2n} P(\boldsymbol{\lambda}_n^*) + C(\gamma_{1n} + \sigma_n^q). \end{aligned} \tag{A.2}$$

since $P(\boldsymbol{\lambda}_n^*) = 0$, we have

$$\mathbf{P}l(Y, \widehat{f}_{\widetilde{\boldsymbol{\lambda}}_n}) - \mathbf{P}l(Y, f_0) \leqslant (\mathbf{P}_n - \mathbf{P}) \left\{ l(Y, \widetilde{f}_n^*) - l(Y, \widehat{f}_{\widetilde{\boldsymbol{\lambda}}_n}) \right\} + C(\gamma_{1n} + \sigma_n^q).$$

Clearly, $\|\widehat{f}_{\widetilde{\boldsymbol{\lambda}}_n}\|_{\mathcal{H}_{\widetilde{\boldsymbol{\lambda}}_n, \sigma_n}} \leqslant C\gamma_{1n}^{-1/2}$, $\|\widetilde{f}_n^*\|_{\mathcal{H}_{\boldsymbol{\lambda}_n^*, \sigma_n}} \leqslant C\gamma_{1n}^{-1/2}$ and by Vert and Vert (2006), we have $\|\widehat{f}_{\widetilde{\boldsymbol{\lambda}}_n}\|_\infty \leqslant C\sigma_n^{-q/2}\gamma_{1n}^{-1/2}$, $\|\widetilde{f}_n^*\|_\infty \leqslant C\sigma_n^{-q/2}\gamma_{1n}^{-1/2}$. We finally conclude

$$\mathbf{P}l(Y, \widehat{f}_{\widetilde{\boldsymbol{\lambda}}_n}) - \mathbf{P}l(Y, f_0) \leqslant \sup_{f \in \mathcal{F}_n, g \in \mathcal{G}_n} (\mathbf{P}_n - \mathbf{P}) \{l(Y, g) - l(Y, f)\} + C(\gamma_{1n} + \sigma_n^q), \qquad (A.3)$$

where

$$\mathcal{F}_n \equiv \left\{ f(\mathbf{x}) = \sum_{i=1}^n \alpha_i \kappa_{\widetilde{\boldsymbol{\lambda}}_n, \sigma_n}(\mathbf{x}_i, \mathbf{x}) : \|f\|_{\mathcal{H}_{\widetilde{\boldsymbol{\lambda}}_n, \sigma_n}} \leqslant C\gamma_{1n}^{-1/2}, \quad \|f\|_\infty \leqslant C\sigma_n^{-q/2}\gamma_{1n}^{-1/2} \right\}$$

and $\mathcal{G}_n$ is defined the same way except that $\widetilde{\boldsymbol{\lambda}}_n$ is $\boldsymbol{\lambda}_n^*$ and $\mathcal{H}_{\widetilde{\boldsymbol{\lambda}}_n, \sigma_n}$ is $\mathcal{H}_{\boldsymbol{\lambda}_n^*, \sigma_n}$.

From Theorem 2, we have

$$\log \mathcal{N}_{[]} \left(\epsilon, \mathcal{F}_n, \|\cdot\|_{L_2(P)} \right) \leqslant C\sigma_n^{-(1-v/4)q} \epsilon^{-v} \gamma_{1n}^{-v/2}$$

for any constant $v \in (0, 2)$. By Condition C1, it holds that for any $f_1, f_2 \in \mathcal{F}_n$,

$$\|l(Y, f_1) - l(Y, f_2)\|_{L_2(P)} \leqslant C\|f_1 - f_2\|_{L_2(P)}$$

We obtain

$$\log \mathcal{N}_{[]} \left(\epsilon, \{l(Y, f) : f \in \mathcal{F}_n\}, \|\cdot\|_{L_2(P)} \right) \leqslant C\sigma_n^{-(1-v/4)q} \epsilon^{-v} \gamma_{1n}^{-v/2}.$$

On the other hand, from condition C1, $\|l(Y, \widehat{f}_{\widetilde{\lambda}_n})\|_{L_2(P)} \leqslant C\sigma_n^{-q/2}\gamma_{1n}^{-1/2}$ so using Theorem 2.14.2 in van der Vaart and Wellne (1996), it gives

$$E\Big\{ \sup_{f \in \mathcal{F}_n} |(\mathbf{P}_n - \mathbf{P}) l(Y, f)| \Big\} \leqslant Cn^{-1/2} \int_0^{C\sigma_n^{-q/2}\gamma_{1n}^{-1/2}} \sqrt{1 + \log \mathcal{N}_{[]}(\epsilon, \{l(Y, f) : f \in \mathcal{F}_n, \|\cdot\|_{L^2(P))}\}} d\epsilon$$

$$\leqslant Cn^{-1/2}\sigma_n^{-(1-3v/8)q}\gamma_{1n}^{-1/2}. \qquad (A.4)$$

From the Talagrand inequality (Wainwright (2006)) and Lipschitz continuity of $l(y, f)$ in f, we have

$$\Pr\left[\sup_{f \in \mathcal{F}_n} \{|(\mathbf{P}_n - \mathbf{P}) l(Y, f)|\} - E\left\{ \sup_{f \in \mathcal{F}_n} |(\mathbf{P}_n - \mathbf{P}) l(Y, f)| \right\} > \frac{t\sigma_n^{-q/2}\gamma_{1n}^{-1/2}}{\sqrt{n}} \right] \leqslant 2e^{-\frac{1}{2}\frac{t^2}{w_n + \frac{1}{3}t/\sqrt{n}}} \qquad (A.5)$$

where

$$w_n = \sup_{f\in\mathcal{F}_n} var\left\{l(Y,f)\right\}\sigma_n^q\gamma_{1n} + 2|E\sup_{f\in\mathcal{F}_n}(\mathbf{P}_n-\mathbf{P})l(Y,f)|\sigma_n^{q/2}\gamma_{1n}^{1/2}.$$

We particularly choose $t=\sqrt{C}a_n^{-1}$ and $a_n=\sigma_n^{(1-3v/8)q}\gamma_{1n}^{1/2}$. Note $\sup_{f\in\mathcal{F}_n} var\left\{l(Y,f)\right\}\leqslant C\sigma_n^{-q}\gamma_{1n}^{-1}$, then (A.4) and (A.5) yield

$$\Pr\left[\left\{\sup_{f\in\mathcal{F}_n}|(\mathbf{P}_n-\mathbf{P})l(Y,f)|\right\} > \sqrt{C}n^{-1/2}a_n^{-1}\right] \leqslant 2\exp(-\frac{C}{a_n^2+n^{-1/2}a_n}).$$

The same inequality holds for $\Pr\left[\left\{\sup_{f\in\mathcal{G}_n}|(\mathbf{P}_n-\mathbf{P})l(Y,f)|\right\} > \sqrt{C}n^{-1/2}a_n^{-1}\right]$.

Hence, from equation (A.3), we conclude that with probability at least $1-4\exp(-\frac{C}{a_n^2+n^{-1/2}a_n})$, it holds

$$\mathbf{P}l(Y,\widehat{f}) - \mathbf{P}l(Y,f_0) \leqslant C\left(n^{-\frac{1}{2}}a_n^{-1}+\sigma_n^q+\gamma_{1n}\right).$$

$n^{-1/2}a_n = n^{-1/2+\alpha_1(1-3v/8)q+\alpha_2/2}\to 0$, so with probability at least $1-4\exp(-Cn^{1/2-\alpha_1(1-3v/8)q-\alpha_2/2})$,

$$\mathbf{P}l(Y,\widehat{f}) - \mathbf{P}l(Y,f_0) \leqslant Cn^{-\xi_1},$$

where $\xi_1=\min(1/2+\alpha_1(1-3v/8)q+\alpha_2/2, \alpha_1 q, \alpha_2)$. This implies that with probability tending to 1, $\limsup_n E\left\{l(Y,\widehat{f}_{\widetilde{\boldsymbol{\lambda}}_n})\right\} \leqslant E\left\{l(Y,f_0)\right\}$. Since $El(Y,f_0)$ is the minimal risk, it yields

$$\lim_{n\to\infty} E\left\{l(Y,\widehat{f}_{\widetilde{\boldsymbol{\lambda}}_n})\right\} = E\left\{l(Y,f_0)\right\}$$

with probability 1. Further from Condition C5, we obtain that with probability at least $1-4\exp\{-Cn^{1/2-\alpha_1(1-3v/8)q-\alpha_2/2}\}$,

$$d_2(\widehat{f}_{\widetilde{\boldsymbol{\lambda}}_n}, f_0) \leqslant Cn^{-\xi_1/2}. \tag{A.6}$$

As a note, the main advantage of using truncated penalty in this Step 1 (Equation A.1, comparing to the objective function under the oracle, i.e., we know which features are truly important) is that there would be no penalty for λ's for the oracle estimators. As a result, the derived convergence rate at this step does not depend on the regularization parameter for the feature selection, i.e., γ_{2n}. This is necessary for proving the oracle property for the

proposed method, since γ_{2n} has to vanish slower than other tuning parameters, in order to make the penalty for the non-oracle estimator relative large.

Step 2. We show that with probability tending to one, $\widetilde{\lambda}_{nm} > 0$ for $m = 1, .., q$. First, from Theorem 2, there exits some positive number ϵ such that

$$d_2(f_0, \mathcal{H}_{\boldsymbol{\lambda}_n, \sigma_n}) \geqslant \epsilon$$

whenever $\lambda_{nm} = 0$ for some $m \leqslant q$. By Condition C1, it gives $\inf_{f \in \mathcal{H}_{\boldsymbol{\lambda}_n, \sigma_n}} E\{l(Y, f)\} - E\{l(Y, f_0)\} \geqslant \widetilde{\epsilon}$ for some $\widetilde{\epsilon} > 0$. Therefore,

$$\begin{aligned}
1 - \Pr(\widetilde{\lambda}_{nm} > 0, m = 1, ..., q) &\leqslant \sum_{m=1}^{q} \Pr(\widetilde{\lambda}_{nm} = 0) \\
&\leqslant \sum_{m=1}^{q} \Pr\left[\inf_{f \in \mathcal{H}_{\widetilde{\boldsymbol{\lambda}}_n, \sigma_n}} E\{l(Y, f)\} - E\{l(Y, f_0)\} \geqslant \widetilde{\epsilon}\right] \\
&\leqslant \sum_{m=1}^{q} \Pr\left[E\left\{l(Y, \widehat{f}_{\widetilde{\boldsymbol{\lambda}}_n})\right\} - E\{l(Y, f_0)\} \geqslant \widetilde{\epsilon}\right] \\
&\leqslant 4q \exp\left\{-Cn^{1/2-\alpha_1(1-3v/8)q-\alpha_2/2}\right\},
\end{aligned}$$

where the last step is from the result in Step 1. Therefore, we conclude that with probability at least

$$\theta_{1n} = 1 - 4q \exp\{-Cn^{1/2-\alpha_1(1-3v/8)q-\alpha_2/2}\},$$

$\widetilde{\lambda}_{nm} > 0$ for all $m = 1, ..., q$.

Step 3. We show that $\widetilde{\boldsymbol{\lambda}}_n$ is a local minimizer. We prove it by verifying the following KKT conditions:

$$\frac{\partial}{\partial \lambda_{nm}}\bigg|_{\widetilde{\lambda}_{nm}} \left[\mathbf{P}_n\left\{l(Y, \widehat{f}_{\boldsymbol{\lambda}_n})\right\} + \gamma_{1n}\|\widehat{f}_{\boldsymbol{\lambda}_n}\|^2_{\mathcal{H}_{\boldsymbol{\lambda}_n, \sigma_n}} + \gamma_{2n} P(\boldsymbol{\lambda}_n)\right] = 0$$

for $m = 1, ..., q$, and

$$\frac{\partial}{\partial \lambda_{nm}}\bigg|_{\widetilde{\lambda}_{nm}=0+} \left[\mathbf{P}_n\left\{l(Y, \widehat{f}_{\boldsymbol{\lambda}_n})\right\} + \gamma_{1n}\|\widehat{f}_{\boldsymbol{\lambda}_n}\|^2_{\mathcal{H}_{\boldsymbol{\lambda}_n, \sigma_n}} + \gamma_{2n} P(\boldsymbol{\lambda}_n)\right] > 0$$

for $m = q + 1, ..., p_n$. First, from Step 2, we know that the first q equations hold with probability at least θ_{1n}. It remains to verify the last $(p_n - q)$ KKT conditions. To this end,

we define

$$g(\boldsymbol{\lambda}_n) = \mathbf{P}_n \left\{ l(Y, \widehat{f}_{\boldsymbol{\lambda}_n}) \right\} + \gamma_{1n} \|\widehat{f}_{\boldsymbol{\lambda}_n}\|^2_{\mathcal{H}_{\boldsymbol{\lambda}_n,\sigma_n}},$$

where $\widehat{f}_{\boldsymbol{\lambda}_n}$ is the optimal solution of $L_n(\boldsymbol{\lambda}_n, f)$ so takes form

$$\widehat{f}_{\boldsymbol{\lambda}_n}(\mathbf{X}) = \sum_{j=1}^{n} \widehat{\alpha}_j(\boldsymbol{\lambda}_n) \kappa_{\boldsymbol{\lambda}_n,\sigma_n}(\mathbf{X}_j, \mathbf{X}) = \widehat{\boldsymbol{\alpha}}(\boldsymbol{\lambda}_n)^\intercal \mathbf{K}_{\boldsymbol{\lambda}_n}(\mathbf{X}, \mathbf{X}).$$

Here, $\mathbf{K}_{\boldsymbol{\lambda}_n}$ is the kernel matrix $(\kappa_{\boldsymbol{\lambda}_n,\sigma_n}(\mathbf{X}_i, \mathbf{X}_j))$ and $\widehat{\boldsymbol{\alpha}}(\boldsymbol{\lambda}_n)$ is the vector of $(\widehat{\alpha}_i(\boldsymbol{\lambda}_n))$. Furthermore, by performing a functional differentiation for the objective function with respect to f, $\widehat{f}_{\boldsymbol{\lambda}_n}$ satisfies the functional equation

$$\mathbf{P}_n l_2(Y, \widehat{f}_{\boldsymbol{\lambda}_n}) h(\mathbf{X}) + 2\gamma_{1n} \langle h(\mathbf{X}), \widehat{f}_{\boldsymbol{\lambda}_n}(\mathbf{X}) \rangle = 0 \tag{A.7}$$

for any $h(\mathbf{X}) = \sum_{j=1}^{n} \xi_j \kappa_{\boldsymbol{\lambda}_n,\sigma_n}(\mathbf{X}_j, \mathbf{X})$.

After differentiating $g(\boldsymbol{\lambda}_n)$ with respect to λ_{nm} and evaluating at $\lambda_{nm} = 0$, we have

$$\frac{\partial g}{\partial \lambda_{nm}} = \mathbf{P}_n l_2(Y, \widehat{f}_{\boldsymbol{\lambda}_n}) \frac{\partial \widehat{f}_{\boldsymbol{\lambda}_n}}{\partial \lambda_{nm}}(\mathbf{X}) + \gamma_{1n} \frac{\partial}{\partial \lambda_{nm}} (\widehat{\boldsymbol{\alpha}}(\boldsymbol{\lambda}_n)^\intercal \mathbf{K}_{\boldsymbol{\lambda}_n} \widehat{\boldsymbol{\alpha}}(\boldsymbol{\lambda}_n))$$

where

$$\begin{aligned}
\frac{\partial \widehat{f}_{\boldsymbol{\lambda}_n}}{\partial \lambda_{nm}}(\mathbf{X}) &= \sum_{j=1}^{n} \frac{\partial \widehat{\alpha}_j(\boldsymbol{\lambda}_n)}{\partial \lambda_{nm}} \kappa_{\boldsymbol{\lambda}_n,\sigma_n}(\mathbf{X}_j, \mathbf{X}) + \sum_{j=1}^{n} \widehat{\alpha}_j(\boldsymbol{\lambda}_n) \frac{\partial \kappa_{\boldsymbol{\lambda}_n,\sigma_n}(\mathbf{X}_j, \mathbf{X})}{\partial \lambda_{nm}} \\
&= \sum_{j=1}^{n} \frac{\partial \widehat{\alpha}_j(\boldsymbol{\lambda}_n)}{\partial \lambda_{nm}} \kappa_{\boldsymbol{\lambda}_n,\sigma_n}(\mathbf{X}_j, \mathbf{X}) + \sum_{j=1}^{n} \widehat{\alpha}_j(\boldsymbol{\lambda}_n) \kappa_{\boldsymbol{\lambda}_n,\sigma_n}(\mathbf{X}_j, \mathbf{X}) \kappa(X_{jm}, X_m).
\end{aligned}$$

As a result of (A.7),

$$\frac{\partial g}{\partial \lambda_{nm}} = \mathbf{P}_n l_2(Y, \widehat{f}_{\boldsymbol{\lambda}_n}) \left\{ \sum_{j=1}^{n} \widehat{\alpha}_j(\boldsymbol{\lambda}_n) \kappa_{\boldsymbol{\lambda}_n,\sigma_n}(\mathbf{X}_j, \mathbf{X}) \kappa_n(X_{jm}, X_m) \right\} + \gamma_{1n} \widehat{\boldsymbol{\alpha}}(\boldsymbol{\lambda}_n)^\intercal \frac{\partial}{\partial \lambda_{nm}} \mathbf{K}_{\boldsymbol{\lambda}_n} \widehat{\boldsymbol{\alpha}}(\boldsymbol{\lambda}_n).$$

Therefore, we obtain

$$\begin{aligned}
&\frac{\partial L_n(\boldsymbol{\lambda}_n, \widehat{f}_{\boldsymbol{\lambda}_n})}{\partial \lambda_{nm}}|_{\widetilde{\lambda}_{nm}=0+} \\
=&(\mathbf{P}_n - \mathbf{P}) l_2(Y, \widehat{f}_{\widetilde{\boldsymbol{\lambda}}_n}) \left\{ \sum_{j=1}^{n} \widehat{\alpha}_j(\widetilde{\boldsymbol{\lambda}}_n) K_{\widetilde{\boldsymbol{\lambda}}_n}(\mathbf{X}_j, \mathbf{X}) \kappa_n(X_{jm}, X_m) \right\} + \gamma_{1n} \widehat{\boldsymbol{\alpha}}(\widetilde{\boldsymbol{\lambda}}_n)^\intercal \frac{\partial}{\partial \lambda_{nm}} \mathbf{K}_{\widetilde{\boldsymbol{\lambda}}_n} \widehat{\boldsymbol{\alpha}}(\widetilde{\boldsymbol{\lambda}}_n) \\
&+ \mathbf{P} l_2(Y, \widehat{f}_{\widetilde{\boldsymbol{\lambda}}_n}) \left\{ \sum_{j=1}^{n} \widehat{\alpha}_j(\widetilde{\boldsymbol{\lambda}}_n) K_{\widetilde{\boldsymbol{\lambda}}_n}(\mathbf{X}_j, \mathbf{X}) \kappa_n(X_{jm}, X_m) \right\} + \gamma_{2n} \frac{\partial P_m(\widetilde{\lambda}_{nm})}{\partial \lambda_{nm}}.
\end{aligned}$$

As a note, since $\widetilde{\boldsymbol{\lambda}}_n$ takes value zero at its jth component when $j > q$, any term in the above expression depends on $\mathbf{X}$ only through $\mathbf{X}_q$, the first q components of $\mathbf{X}$.

On the other hand, we have

$$\begin{aligned}
&\mathbf{P} l_2(Y, \widehat{f}_{\widetilde{\boldsymbol{\lambda}}_n}) \left\{ \sum_{j=1}^{n} \widehat{\alpha}_j(\widetilde{\boldsymbol{\lambda}}_n) K_{\widetilde{\boldsymbol{\lambda}}_n}(\mathbf{X}_j, \mathbf{X}) \kappa_n(X_{jm}, X_m) \right\} \\
= \;& \mathbf{P} \left\{ l_2(Y, \widehat{f}_{\widetilde{\boldsymbol{\lambda}}_n}) - l_2(Y, f_0) \right\} \left\{ \sum_{j=1}^{n} \widehat{\alpha}_j(\widetilde{\boldsymbol{\lambda}}_n) K_{\widetilde{\boldsymbol{\lambda}}_n}(\mathbf{X}_{jq}, \mathbf{X}_q) \kappa_n(X_{jm}, X_m) \right\},
\end{aligned}$$

since any direction derivative of the expected loss function at f_0 is zero. By Condition C3, since $l_2(Y, f)$ is locally Lipschitz continuous with respect to f at f_0 in $L_2(P)$, it holds

$$\begin{aligned}
&|\mathbf{P} l_2(Y, \widehat{f}_{\widetilde{\boldsymbol{\lambda}}_n}) \left\{ \sum_{j=1}^{n} \widehat{\alpha}_j(\widetilde{\boldsymbol{\lambda}}_n) K_{\widetilde{\boldsymbol{\lambda}}_n}(\mathbf{X}_j, \mathbf{X}) \kappa_n(X_{jm}, X_m) \right\}| \\
\leqslant \;& C \| \widehat{f}_{\widetilde{\boldsymbol{\lambda}}_n} - f_0 \|_{L_2(P)} \| \left\{ \sum_{j=1}^{n} \widehat{\alpha}_j(\widetilde{\boldsymbol{\lambda}}_n) K_{\widetilde{\boldsymbol{\lambda}}_n}(\mathbf{X}_{jq}, \mathbf{X}_q) \kappa_n(X_{jm}, X_m) \right\} \|_{L_2(P)} \\
= \;& C \| \widehat{f}_{\widetilde{\boldsymbol{\lambda}}_n} - f_0 \|_{L_2(P)} \\
& \times E \Big[\sum_{i,j=1}^{n} \widehat{\alpha}_j(\widetilde{\boldsymbol{\lambda}}_n) \widehat{\alpha}_l(\widetilde{\boldsymbol{\lambda}}_n) \kappa_n(X_{jm}, X_m) \kappa_n(X_{lm}, X_m) E \left\{ K_{\widetilde{\boldsymbol{\lambda}}_n}(\mathbf{X}_{jq}, \mathbf{X}_q) K_{\widetilde{\boldsymbol{\lambda}}_n}(\mathbf{X}_{lq}, \mathbf{X}_q) | X_m \right\} \Big]^{1/2}.
\end{aligned}$$

According to Condition (C.4),

$$\sum_{i,j=1}^{n} \widehat{\alpha}_j(\widetilde{\boldsymbol{\lambda}}_n) \widehat{\alpha}_l(\widetilde{\boldsymbol{\lambda}}_n) \kappa_n(X_{jm}, X_m) \kappa_n(X_{lm}, X_m) E \left\{ K_{\widetilde{\boldsymbol{\lambda}}_n}(\mathbf{X}_{jq}, \mathbf{X}_q) K_{\widetilde{\boldsymbol{\lambda}}_n}(\mathbf{X}_{lq}, \mathbf{X}_q) | X_m \right\}$$

$$\leqslant \Lambda_{\max}(\mathbf{X}_{-q}) \sum_{j=1}^{n} \left\{ \widehat{\alpha}_j(\widetilde{\boldsymbol{\lambda}}_n) \kappa_n(X_{jm}, X_m) \right\}^2 \leqslant c \sigma_n^{-1/2} \Lambda_{\min}(\mathbf{X}_{-q}) \sum_{j=1}^{n} \left\{ \widehat{\alpha}_j(\widetilde{\boldsymbol{\lambda}}_n) \kappa_n(X_{jm}, X_m) \right\}^2,$$

so from the second half of condition (C.4),

$$\begin{aligned}
& E \left[\sum_{i,j=1}^{n} \widehat{\alpha}_j(\widetilde{\boldsymbol{\lambda}}_n) \widehat{\alpha}_l(\widetilde{\boldsymbol{\lambda}}_n) \kappa_n(X_{jm}, X_m) \kappa_n(X_{lm}, X_m) E \left\{ K_{\widetilde{\boldsymbol{\lambda}}_n}(\mathbf{X}_{jq}, \mathbf{X}_q) K_{\widetilde{\boldsymbol{\lambda}}_n}(\mathbf{X}_{lq}, \mathbf{X}_q) | X_m \right\} \right] \\
\leqslant \;& c \sigma_n^{-1/2} E \left[\Lambda_{\min}(\mathbf{X}_{-q}) \sum_{j=1}^{n} \left\{ \widehat{\alpha}_j(\widetilde{\boldsymbol{\lambda}}_n) \kappa_n(X_{jm}, X_m) \right\}^2 \right] \\
\leqslant \;& c E \left[E \left\{ \Lambda_{\min}(\mathbf{X}_{-q}) | \mathbf{X}_q \right\} \sum_{j=1}^{n} \widehat{\alpha}_j(\widetilde{\boldsymbol{\lambda}}_n)^2 \right] \\
\leqslant \;& c E \left[\sum_{i,j=1}^{n} \widehat{\alpha}_j(\widetilde{\boldsymbol{\lambda}}_n) \widehat{\alpha}_l(\widetilde{\boldsymbol{\lambda}}_n) E \left\{ K_{\widetilde{\boldsymbol{\lambda}}_n}(\mathbf{X}_{jq}, \mathbf{X}_q) K_{\widetilde{\boldsymbol{\lambda}}_n}(\mathbf{X}_{lq}, \mathbf{X}_q) \right\} \right] = c \| \widehat{f}_{\widetilde{\boldsymbol{\lambda}}_n}(\mathbf{X}) \|_{L_2(P)}^2.
\end{aligned}$$

Thus, we conclude

$$\begin{aligned}
&|\mathbf{P} l_2(Y, \widehat{f}_{\widetilde{\boldsymbol{\lambda}}_n}) \left\{ \sum_{j=1}^{n} \widehat{\alpha}_j(\widetilde{\boldsymbol{\lambda}}_n) K_{\widetilde{\boldsymbol{\lambda}}_n}(\mathbf{X}_j, \mathbf{X}) \kappa_n(X_{jm}, X_m) \right\}| \\
\leqslant\ & C\|\widehat{f}_{\widetilde{\boldsymbol{\lambda}}_n} - f_0\|_{L_2(P)} \|\widehat{f}_{\widetilde{\boldsymbol{\lambda}}_n}\|_{L_2(P)} \\
\leqslant\ & C\|\widehat{f}_{\widetilde{\boldsymbol{\lambda}}_n} - f_0\|_{L_2(P)} \Big(\|\widehat{f}_{\widetilde{\boldsymbol{\lambda}}_n} - f_0\|_{L_2(P)} + \|f_0\|_{L_2(P)} \Big) \\
\leqslant\ & C n^{-\xi_1/2}.
\end{aligned}$$

where the last step uses the result from (A.6) and the boundedness of $\|f_0\|_{L_2(P)}$.

Hence, for $m > q$, since $\sum_{i,j=1}^{n} \widehat{\alpha}_i(\widetilde{\boldsymbol{\lambda}}_n) K_{\widetilde{\boldsymbol{\lambda}}_n}(\mathbf{X}_i, \mathbf{X}_j) \kappa_n(X_{im}, X_{jm}) \widehat{\alpha}_j(\widetilde{\boldsymbol{\lambda}}_n) \geqslant 0$, we have

$$\begin{aligned}
&\mathbf{P}\Big(\frac{\partial}{\partial \lambda_{nm}}\Big|_{\widetilde{\lambda}_{nm}=0+} \Big[\mathbf{P}_n \left\{ l(Y, f_{\boldsymbol{\lambda}_n}) \right\} + \gamma_{1n} \|f\|^2_{\mathcal{H}_{\boldsymbol{\lambda}_n, \sigma_n}} + \gamma_{2n} P(\boldsymbol{\lambda}_n) \Big] \leqslant 0 \Big) \\
=&\mathbf{P}\Big(\frac{\partial}{\partial \lambda_{nm}}\Big|_{\widetilde{\lambda}_{m}=0+} \{ g(\boldsymbol{\lambda}_n) + \gamma_{2n} P(\boldsymbol{\lambda}_n) \} \leqslant 0 \Big) \\
=&\mathbf{P}\Bigg[(\mathbf{P}_n - \mathbf{P}) l_2(Y, \widehat{f}_{\widetilde{\boldsymbol{\lambda}}_n}) \left\{ \sum_{j=1}^{n} \widehat{\alpha}_j(\widetilde{\boldsymbol{\lambda}}_n) K_{\widetilde{\boldsymbol{\lambda}}_n}(\mathbf{X}_j, \mathbf{X}) \kappa_n(X_{jm}, X_m) \right\} \\
&\quad \leqslant -\gamma_{1n} \sum_{i,j=1}^{n} \widehat{\alpha}_i(\widetilde{\boldsymbol{\lambda}}_n) K_{\widetilde{\boldsymbol{\lambda}}_n}(\mathbf{X}_i, \mathbf{X}_j) \kappa_n(X_{im}, X_{jm}) \widehat{\alpha}_j(\widetilde{\boldsymbol{\lambda}}_n) \\
&\quad - \mathbf{P} l_2(Y, \widehat{f}_{\widetilde{\boldsymbol{\lambda}}_n}) \left\{ \sum_{j=1}^{n} \widehat{\alpha}_j(\widetilde{\boldsymbol{\lambda}}_n) K_{\widetilde{\boldsymbol{\lambda}}_n}(\mathbf{X}_j, \mathbf{X}) \kappa_n(X_{jm}, X_m) \right\} - \gamma_{2n} \frac{\partial P(\boldsymbol{\lambda}_n)}{\partial \lambda_{nm}}\Big|_{\widetilde{\lambda}_{nm}=0+} \Bigg] \\
&\quad \leqslant \mathbf{P}\Bigg[(\mathbf{P}_n - \mathbf{P}) l_2(Y, \widehat{f}_{\widetilde{\boldsymbol{\lambda}}_n}) \left\{ \sum_{j=1}^{n} \widehat{\alpha}_j(\widetilde{\boldsymbol{\lambda}}_n) K_{\widetilde{\boldsymbol{\lambda}}_n}(\mathbf{X}_j, \mathbf{X}) \kappa_n(X_{jm}, X_m) \right\} \\
&\qquad \leqslant n^{-\xi_1/2} - \gamma_{2n} \frac{\partial P_m(\lambda_{nm})}{\partial \lambda_{nm}}\Big|_{\widetilde{\lambda}_{nm}=0+} \Bigg].
\end{aligned}$$

This gives

$$\begin{aligned}
&\mathbf{P}\Bigg[\frac{\partial}{\partial \lambda_{nm}}\Big|_{\widetilde{\lambda}_{nm}=0+} (\mathbf{P}_n \left\{ l(Y, f_{\boldsymbol{\lambda}_n}) \right\} + \gamma_{1n} \|f\|^2_{\mathcal{H}_{\boldsymbol{\lambda}_n, \sigma_n}} + \gamma_{2n} P(\boldsymbol{\lambda}_n)) \leqslant 0 \Bigg] \\
\leqslant &\mathbf{P}\Bigg[|(\mathbf{P}_n - \mathbf{P}) l_2(Y, \widehat{f}_{\widetilde{\boldsymbol{\lambda}}_n}) \left\{ \sum_{j=1}^{n} \widehat{\alpha}_j(\widetilde{\boldsymbol{\lambda}}_n) K_{\widetilde{\boldsymbol{\lambda}}_n}(\mathbf{X}_j, \mathbf{X}) \kappa_{n_k}(X_{jm}, X_m) \right\}| \geqslant -n^{-\xi_1/2} + c_0 \gamma_{2n} \Bigg] \\
\leqslant &\mathbf{P}\Bigg[|(\mathbf{P}_n - \mathbf{P}) l_2(Y, \widehat{f}_{\widetilde{\boldsymbol{\lambda}}_n}) \left\{ \sum_{j=1}^{n} \widehat{\alpha}_j(\widetilde{\boldsymbol{\lambda}}_n) K_{\widetilde{\boldsymbol{\lambda}}_n}(\mathbf{X}_j, \mathbf{X}) \kappa_{n_k}(X_{jm}, X_m) \right\}| \geqslant c_0 \gamma_{2n}/2 \Bigg]. \qquad (A.8)
\end{aligned}$$

The last two steps use the condition that $\partial / \partial P(\boldsymbol{\lambda}_n)\Big|_{\lambda_{nm}=0+} \geqslant c_0 > 0$ and that from Condition

C5, $\alpha_3 < \xi_1/2$. To obtain an upper bound for (A.8), we need to estimate

$$(\mathbf{P}_n - \mathbf{P}) l_2(Y, \widehat{f}_{\widetilde{\boldsymbol{\lambda}}_n}) \left\{ \sum_{j=1}^{n} \widehat{\alpha}_j(\widetilde{\boldsymbol{\lambda}}_n) K_{\widetilde{\boldsymbol{\lambda}}_n}(\mathbf{X}_j, \mathbf{X}) \kappa_n(X_{jm}, X_m) \right\}.$$

Note $l_2(Y, \widehat{f}_{\widetilde{\boldsymbol{\lambda}}_n}) \left\{ \sum_{j=1}^{n} \widehat{\alpha}_j(\widetilde{\boldsymbol{\lambda}}_n) K_{\widetilde{\boldsymbol{\lambda}}_n}(\mathbf{X}_j, \mathbf{X}) \kappa_n(X_{jm}, X_m) \right\}$ belongs to class

$$\mathcal{L}_n = \Big\{ l_2\left\{Y, f_1(\boldsymbol{X})\right\} f_2(\boldsymbol{X}) : \|f_1\|_{\mathcal{H}_{\widetilde{\boldsymbol{\lambda}}_n, \sigma_n}} \leqslant C\gamma_{1n}^{-1/2}, \quad \|f_2\|_{\mathcal{H}_{\widetilde{\boldsymbol{\lambda}}_n, \sigma_n}} \leqslant C\gamma_{1n}^{-1/2},$$

$$f_1 \text{ is in a neighborhood of } f_0 \text{ in } L_2(P) \big\}.$$

Following the same argument as in Step 1 and using the Lipschitz continuity of $l_2(y, f)$ in f, we have for $v \in (0, 2)$,

$$\log \mathcal{N}_{[]} \Big(\epsilon, \mathcal{L}_n, \|\cdot\|_{L^2(P)} \Big) \leqslant C\sigma_n^{-(2-v/2)q} \gamma_{1n}^{-v} \epsilon^{-v}.$$

In addition,

$$\| l_2(Y, f) \left\{ \sum_{j=1}^{n} \alpha_{j\boldsymbol{\lambda}_n} \kappa_{\boldsymbol{\lambda}_n, \sigma_n}(\mathbf{X}_j, \mathbf{X}) \kappa_n(X_{jm}, X_m) \right\} \|_{L_2(P)} \leqslant Cn^{-\xi_1/2}.$$

We obtain (Theorem 2.14.2 in van der Vaart and Wellner (1998))

$$E \left\{ \sup_{g \in \mathcal{L}_n} |(\mathbf{P}_n - \mathbf{P}) g| \right\} \leqslant Cn^{-1/2} \sigma_n^{-(2-\frac{3v}{4})q} \gamma_{1n}^{-\frac{v}{2}} n^{\xi_1(\frac{v}{4} - \frac{1}{2})} \equiv n^{-\xi_2},$$

where $\xi_2 = \frac{1}{2} + \alpha_1(2 - \frac{3v}{4})q + \alpha_2 \frac{v}{2} - \xi_1(\frac{v}{4} - \frac{1}{2})$.

On other hand, since $\text{abs}\left[l_2(Y, \widehat{f}_{\widetilde{\boldsymbol{\lambda}}_n}) \left\{ \sum_{j=1}^{n} \widehat{\alpha}_{j\widetilde{\boldsymbol{\lambda}}_n} K_{\widetilde{\boldsymbol{\lambda}}_n}(\mathbf{X}_j, \mathbf{X}) \kappa_n(X_{jm}, X_m) \right\} \right] \leqslant C\sigma_n^{-q/2-1} \gamma_{1n}^{-1/2}$.

By the Talagrand inequality, we have

$$\mathbf{P} \Big[\sup_{g \in \mathcal{L}_n} |(\mathbf{P}_n - \mathbf{P}) g - E \left\{ \sup_{g \in \mathcal{L}_n} |(\mathbf{P}_n - \mathbf{P}) g| \right\}| > \frac{t}{\sigma_n^{q/2+1} \gamma_{1n}^{\frac{1}{2}} \sqrt{n}} \Big] \leqslant 2e^{-\frac{1}{2} \frac{t^2}{w_n + \frac{1}{3} t/\sqrt{n}}},$$

where $w_n = \sigma_n^{q+2} \gamma_{1n} \sup_{g \in \mathcal{G}_n} var\Big(g\Big) + 2E \left\{ |\sup_{g \in \mathcal{G}_n} (\mathbf{P}_n - \mathbf{P}) g| \right\} \sigma_n^{q/2+1} \gamma_{1n}^{\frac{1}{2}}$. Combining the above results, we obtain

$$\begin{aligned} &\mathbf{P} \Big[|(\mathbf{P}_n - \mathbf{P}) l_2(Y, \widehat{f}_{\boldsymbol{\lambda}_n}) \big(\sum_{j=1}^{n} \widehat{\alpha}_j(\boldsymbol{\lambda}_n) \kappa_{\boldsymbol{\lambda}_n, \sigma_n}(\mathbf{X}_j, \mathbf{X}) \kappa_n(X_{jm}, X_m) \big)| > \frac{t}{\sqrt{n} \sigma_n^{q/2+1} \gamma_{1n}^{\frac{1}{2}}} + n^{-\xi_2} \Big] \\ &\leqslant 2 \exp(-\frac{1}{2} \frac{Ct^2}{1 + n^{-\xi_2} \sigma_n^{q/2+1} \gamma_{1n}^{\frac{1}{2}} + n^{-1/2} t}). \end{aligned} \tag{A.9}$$

We choose $t = C\sqrt{n} \sigma_n^{q/2+1} \gamma_{1n}^{\frac{1}{2}} (c_0 \gamma_{2n}/2 - n^{-\xi_2})$ in (A.9), which is positive according to

Condition C5. Constraint $\alpha_3 < \xi_2$ Then (A.8) gives

$$\mathbf{P}\left[\frac{\partial}{\partial\lambda_{nm}}\bigg|_{\widetilde{\lambda}_{nm}=0+}\left[\mathbf{P}_n\left\{l(Y,f_{\boldsymbol{\lambda}_n})\right\}+\gamma_{1n}\|f\|^2_{\mathcal{H}_{\boldsymbol{\lambda}_n,\sigma_n}}+\gamma_{2n}P(\boldsymbol{\lambda}_n)\right]\leqslant 0\right]$$
$$\leqslant 2\exp(-Cn\sigma_n^{2+q}\gamma_{1n}\gamma_{2n})\equiv 2\exp(-Cn^{\xi_3}),$$

where $\xi_3=(1-(2+q)\alpha_1-\alpha_2-\alpha_3)>0$. Finally, we conclude that the last (p_n-q) inequalities in the KKT conditions hold with probability at least

$$1-2(p_n-q)\exp(-Cn^{\xi_3}).$$

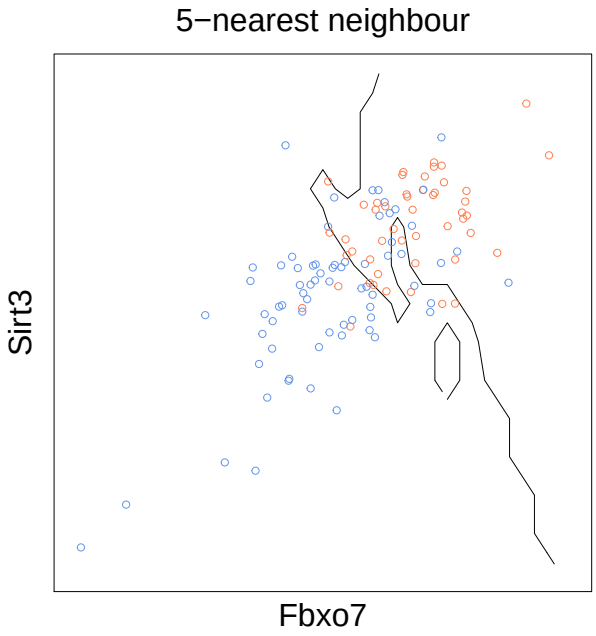
From Condition C5, this probability goes to 1. Hence, $\widetilde{\boldsymbol{\lambda}}_n$ is a local minimizer with probability tending to 1 which is exactly the local minimizer, $\widehat{\boldsymbol{\lambda}}_n$, needed for the main theorem. We have completed the proof.

Remark 1 Recall that Step 1 in the proof to Theorem 3 has established the convergence rate for the predicted function assuming that we know which feature variables are important, and in addition, this convergence rate is $O(n^{-\xi_1})$. Thus, after establishing the oracle property of our proposed approach, we conclude that the same convergence rate applies to the estimated prediction function $\widehat{f}_{\widehat{\boldsymbol{\lambda}}}$. That is, the prediction error of the estimated prediction function converges to the minimal error in a polynomial rate of the sample size, regardless of the number of the unimportant feature variables.

Web Appendix C: Additional Results for Analysis of Gene Expression Data

For the analysis of the gene expression data, the following plot reveals some nonlinear relationship between Sirt 3 and Fbxo7 using 5-Nearest-Neighbors model.

Web Figure 2. 5-Nearest-Neighbor Plot of Sirt3 versus Fbxo7 in Real Data Study



We also redo the analysis using all 31,000 probes without feature screening. The results of prediction performance and feature selection are in the following table. We notice that our proposed method gave the comparable small classification error and the most sparse selection results regarding to the number of selected features (9.1 variables selected on the average from 500 random splittings). This shows that even without prescreening the probes, the proposed method still performs better than the other methods.

Web Table 1. Summary of Analysis Results Using 31,000 Probes

Feature selection result				classification error
	min	max	avg	
Proposed	0	24	9. 1	0.324 (0.057)
HSICLasso	1	31098	2129	0.318 (0.049)
SpAM	3	68	46.2	0.325 (0.058)
l_1-SVM	10	27679	8063.7	0.371 (0.057)

Note: The numbers are the mean of misclassification rates from 500 replicates. The numbers within parentheses are the median absolute deviations from 500 replicates. "min#" is the minimum number of the selected features, "max#" is the max number of the selected features, and "avg.#" is the average number of the selected features in 500 random splittings.

Web Appendix D: Additional Simulation Study

We conducted one additional simulation study with the same setting as the first simulation study in the main paper, but we allowed the dependence among the important variables $X_1, X_2, \cdots, X_5$ and the unimportant variables $X_6, \cdots, X_8$ and also between them. Specifically, $corr(X_1, X_2) = 0.4, corr(X_1, X_3) = -0.3, corr(X_2, X_3) = 0.5, corr(X_3, X_4) = 0.2, corr(X_1, X_7) = -0.2$, $corr(X_6, X_7) = 0.3, corr(X_7, X_8) = 0.2$, while the others were all independent. We continued to consider sample size $n = 100, 200$ and 400 and varied the feature dimension from $p = 200, 400$ to 1000. Each simulation setting was repeated 500 times. For comparison, we continue to compared our proposed method with HSICLasso and SpAM and LASSO. The feature selection and prediction results based on 500 replicates are summarized in the following table. Most of the findings are similar to the first simulation study in the main paper. Thee true positive rates for all the methods become smaller because the correlations among important and unimportant variables give more chances for important variables to be selected as unimportant ones.

[Table 1 about here.]

References

Steinwart, I. and Christmann, A. (2008). *Support Vector Machines*. Springer.

Steinwart, I. and Scovel, C. (2007). Fast rates for support vector machines using gaussian kernels. *Annals of Statistics* **35,** 575–607.

van der Vaart, A. and Wellne, A. (1996). *Weak Convergence and Empirical Processes*. Springer.

Vert, R. and Vert, J.-P. (2006). Consistency and convergence rates of one-class svms and related algorithms. *Journal of Machine Learning Research* **7,** 817–854.

Wainwright, M. (2006). *High-Dimensional Statistics: A Non-Asymptotic Viewpoint*. Cambridge University Press.

Table 1
Results from The Additional Simulation Study with Continuous Outcome

(a) Summary of Feature Selection Performance

		Proposed Method			HSICLasso			SPAM			LASSO		
p	n	TPR	TNR	Avg#	TPR	TNR	Avg#	TPR	TNR	Avg#	TPR	TNR	Avg#
100	100	59.6%	97.0%	5.8	71.7%	77.7%	24.8	93.9%	33.5%	67.8	99.0 %	1.3%	98.7
	200	63.4%	99.0%	4.1	88.1%	59.9 %	42.5	99.6%	3.8%	95.8	100.0%	0.1%	99.9
	400	70.0%	99.2%	4.3	92.0 %	74.4 %	28.9	100.0%	0.2%	99.8	100.0%	0.1%	99.9
200	100	54.6%	98.5%	5.7	64.5%	86.9%	49.1	88.2%	62.3%	77.9	85.8%	52.2%	97.5
	200	61.1%	99.5%	4.0	80.3%	76.2%	44.0	96.6%	33.0%	135.5	99.5%	0.8%	198.5
	400	66.1%	99.8%	3.8	91.5 %	82.8%	39.0	99.5%	4.6%	190.9	100.0%	0.1%	199.7
400	100	51.2%	99.1%	6.2	58.7%	89.7%	43.5	81.3%	80.2%	82.3	77.7%	76.4%	97.1
	200	60.6%	99.7%	4.1	67.8%	90.7%	40.0	92.3%	61.6%	156.2	88.2 %	51.3%	196.6
	400	63.5%	99.9%	3.7	89.8%	86.2%	58.9	97.0%	33.3%	268.5	99.5%	0.7%	397.3
1000	100	45.8%	99.6%	5.9	49.3%	91.1%	91.3	73.4%	91.8%	84.3	72.1%	90.6%	97.1
	200	57.2%	99.9%	4.1	58.0%	98.3%	19.5	86.8%	83.9%	164.8	80.8%	80.7%	195.8
	400	63.7%	99.9%	3.6	87.6%	90.2%	101.8	92.8%	68.3%	319.2	88.8%	60.8%	394.5

(b) Summary of Prediction Errors

p	n	Proposed Method	HSICLasso	SPAM	LASSO
100	100	6.733 (0.441)	7.085 (0.258)	7.047 (0.310)	36.815 (10.144)
	200	5.906 (0.401)	6.754 (0.094)	7.660 (0.430)	8.546 (0.429)
	400	5.216 (0.339)	6.532 (0.058)	7.335 (0.332)	7.107 (0.147)
200	100	6.693 (0.471)	6.986 (0.328)	6.536 (0.311)	9.066 (0.687)
	200	5.927 (0.251)	6.631 (0.143)	6.461 (0.295)	38.112 (8.374)
	400	5.485(0.150)	6.244 (0.058)	7.064 (0.316)	8.307 (0.351)
400	100	6.876 (0.561)	7.310 (0.432)	6.521 (0.368)	8.100 (0.471)
	200	6.042 (0.276)	6.731 (0.428)	6.161 (0.240)	8.998 (0.555)
	400	5.674 (0.098)	6.310 (0.056)	6.206 (0.182)	34.457 (5.114)
1000	100	7.203 (0.658)	8.021 (0.315)	6.775 (0.346)	7.954 (0.343)
	200	6.325 (0.333)	6.880 (0.218)	6.177 (0.223)	7.906 (0.306)
	400	5.864 (0.128)	6.742 (0.075)	5.972 (0.146)	8.581 (0.342)

Note. In (a), "TPR" is the true positive rate, "TNR" is the true negative rate, and "Avg#" is the average number of the selected variables from 500 replicates. In (b), the numbers are the mean squared errors from prediction, and the numbers within parentheses are the median absolute deviations from 500 replicates.