Supplementary materials for 'Efficient kernel-based variable selection with sparsistency '

Xin He†, Junhui Wang‡ and Shaogao Lv§

† *Shanghai University of Finance and Economics*

‡ *City University of Hong Kong*

§ *Nanjing Audit University*

S1. Technical proofs

To be self-contained, we first give a special case of Theorem 1 in Zhou [5] as a lemma on the smooth RKHS below, which plays an important role for the subsequent analysis. Its proof follows directly from that of Theorem 1 in Zhou [5] and thus is omitted here.

Lemma 1. *Let* $K : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ *be a Mercer kernel such that* $K \in C^4(\mathcal{X} \times \mathcal{X})$*, where* C^4 *is a class of functions whose fourth derivative is continuous. Then the following statements hold:*

(a) For any $\mathbf{x} \in \mathcal{X}$*,* $\partial_l K_{\mathbf{x}}, \partial_{lk} K_{\mathbf{x}} \in \mathcal{H}_K$*, for any* $l, k = 1, ..., p_n$*.*

(b) A derivative reproducing property holds true; that is, for any $f \in \mathcal{H}_K$,

$$\partial_l f(\mathbf{x}) = \langle f, \partial_l K_{\mathbf{x}} \rangle_K, \quad \textit{and } \partial_{lk} f(\mathbf{x}) = \langle f, \partial_{lk} K_{\mathbf{x}} \rangle_K.$$

Proposition 1. *Suppose Assumption 2 in the main text is met. Let* $\widetilde{f}$ *be the minimizer of* $\mathcal{E}_{\lambda_n}(f) = E(y - f(\mathbf{x}))^2 + \lambda_n \|f\|_K^2$ *in* $\mathcal{H}_K$*. Then conditioning on the event* $\{\mathcal{Z}^n : \max_{i=1,\dots,n} |y_i| \leq M_n\}$ *with* $M_n \geq (\kappa_1^2 \|f^*\|_K^2 + \sigma^2)^{1/2}$*, for any* $\delta_n \in (0,1)$*, with probability at least* $1 - \delta_n$*, there holds*

$$\|\widehat{f} - \widetilde{f}\|_K \leq \frac{6\kappa_1 M_n}{\lambda_n n^{1/2}} \log \frac{2}{\delta_n}.$$

Proof of Proposition 1: Define the sample operators $S_{\mathbf{x}} : \mathcal{H}_K \to \mathcal{R}^n$ and $S_{\mathbf{x}}^T : \mathcal{R}^n \to \mathcal{R}$ as

$$\mathcal{S}_{\mathbf{x}}(f) = (f(\mathbf{x}_1), \dots, f(\mathbf{x}_n))^T \text{ and } \mathcal{S}_{\mathbf{x}}^T \mathbf{c} = \sum_{i=1}^n c_i K_{\mathbf{x}_i}.$$

Then solving (1) in the main text is equivalent to solve

$$\widehat{f} = \underset{f \in \mathcal{H}_K}{\operatorname{argmin}} \ \frac{1}{n} \mathbf{y}^T \mathbf{y} - \frac{2}{n} \langle f, S_{\mathbf{x}}^T \mathbf{y} \rangle_K + \frac{1}{n} \langle f, \mathcal{S}_{\mathbf{x}}^T \mathcal{S}_{\mathbf{x}} f \rangle_K + \lambda_n \langle f, f \rangle_K,$$

where $\mathbf{y} = (y_1, ..., y_n)^T$, and hence that

$$\widehat{f} = \left(\frac{1}{n}\mathcal{S}_{\mathbf{x}}^T\mathcal{S}_{\mathbf{x}} + \lambda_n I\right)^{-1} \frac{1}{n}\mathcal{S}_{\mathbf{x}}^T y.$$

Similarly, the minimizer of $\mathcal{E}_{\lambda_n}(f)$ in $\mathcal{H}_K$ must have the form

$$\widetilde{f} = (L_K + \lambda_n I)^{-1} L_K f^*.$$

Therefore, we have

$$\begin{aligned}\widehat{f} - \widetilde{f} &= \big(\frac{1}{n}\mathcal{S}_{\mathbf{x}}^T\mathcal{S}_{\mathbf{x}} + \lambda_n I\big)^{-1}\big(\frac{1}{n}S_{\mathbf{x}}^T\mathbf{y} - \frac{1}{n}\mathcal{S}_{\mathbf{x}}^T\mathcal{S}_{\mathbf{x}}\widetilde{f} - \lambda_n\widetilde{f}\big)\\ &= \big(\frac{1}{n}\mathcal{S}_{\mathbf{x}}^T\mathcal{S}_{\mathbf{x}} + \lambda_n I\big)^{-1}\Big(\frac{1}{n}\sum_{i=1}^n \big(y_i - \widetilde{f}(\mathbf{x}_i)\big)K_{\mathbf{x}_i} - L_K\big(f^* - \widetilde{f}\big)\Big),\end{aligned}$$

and its RKHS-norm can be upper bounded as

$$\|\widehat{f} - \widetilde{f}\|_K \leq \lambda_n^{-1}\Big\|\frac{1}{n}\sum_{i=1}^n \big(y_i - \widetilde{f}(\mathbf{x}_i)K_{\mathbf{x}_i} - L_K\big(f^* - \widetilde{f}\big)\Big\|_K = \lambda_n^{-1}\Delta_1.$$

To bound Δ_1, denote $\xi_i = \big(y_i - \widetilde{f}(\mathbf{x}_i)\big)K_{\mathbf{x}_i}$, and it follows from Assumption 2 in the main text and direct calculation that

$$
\begin{aligned}
&E\xi = L_K\big(f^* - \widetilde{f}\big), \quad \|\xi\|_K \le \kappa_1\big(M_n + \|\widetilde{f}\|_\infty\big),\\
&E(\|\xi\|_K^2) \le \kappa_1^2 \int \big(y - \widetilde{f}(x)\big)^2 d\rho_{\mathbf{x},y}.
\end{aligned}
$$

By Lemma 2 of Smale and Zhou [3] and Assumption 2 in the main text, with probability at least $1-\delta_n$, there holds

$$
\begin{aligned}
\Delta_1 \le 2n^{-1}\kappa_1 \log\frac{2}{\delta_n}(M_n + \|\widetilde{f}\|_\infty) + &\\
& n^{-1/2}\kappa_1\Big(2\log\frac{2}{\delta_n}\Big)^{1/2}\Big(\int \big(y - \widetilde{f}(\mathbf{x})\big)^2 d\rho_{\mathbf{x},y}\Big)^{1/2}.
\end{aligned}
$$

For $\|\widetilde{f}\|_\infty$, by the definition of $\widetilde{f}$, we have

$$
\|\widetilde{f} - f^*\|_2^2 + \lambda_n\|\widetilde{f}\|_K^2 \le \|0 - f^*\|_2^2 + \lambda_n\|0\|_K^2 \le \|f^*\|_2^2, \tag{S.1}
$$

where $\|f^*\|_2^2$ is a bounded quantity. Hence, there holds

$$
\|\widetilde{f}\|_\infty \le \kappa_1\|\widetilde{f}\|_K \le \kappa_1\lambda_n^{-1/2}\|f^*\|_2. \tag{S.2}
$$

For $\int \big(y - \widetilde{f}(\mathbf{x})\big)^2 d\rho_{\mathbf{x},y}$, note that

$$\int \big(y - f(\mathbf{x})\big)^2 d\rho_{\mathbf{x},y} - \int \big(y - f^*(\mathbf{x})\big)^2 d\rho_{\mathbf{x},y} = \|f - f^*\|_2^2,$$

for any f. Substituting $f = 0$ and $f = \widetilde{f}$ yield that

$$\int \big(y - f^*(\mathbf{x})\big)^2 d\rho_{\mathbf{x},y} + \|f^*\|_2^2 = \int y^2 d\rho_{\mathbf{x},y} \le \kappa_1^2 \|f^*\|_K^2 + \sigma^2 \le M_n^2, \quad \text{(S.3)}$$

$$\int \big(y - \widetilde{f}(\mathbf{x})\big)^2 d\rho_{\mathbf{x},y} = \|\widetilde{f} - f^*\|_2^2 + \int \big(y - f^*(\mathbf{x})\big)^2 d\rho_{\mathbf{x},y} \le 2M_n^2, \quad \text{(S.4)}$$

where the last inequality follows from (S.1) and (S.3).

Combing (S.2) and (S.4), we have with probability at least $1 - \delta_n$ that

$$\begin{aligned}
\Delta_1 &\le 2n^{-1}\kappa_1 \log\frac{2}{\delta_n} M_n (1 + \kappa_1 \lambda_n^{-1/2}) + 2n^{-1/2}\kappa_1 \big(\log\frac{2}{\delta_n}\big)^{1/2} M_n^2 \\
&\le \frac{2\kappa_1 M_n}{n}\log\frac{2}{\delta_n} + \frac{2\kappa_1 M_n}{n^{1/2}}\log\frac{2}{\delta_n}\frac{\kappa_1}{\lambda_n^{1/2} n^{1/2}} + \frac{2\kappa_1 M_n}{n^{1/2}}\big(\log\frac{2}{\delta_n}\big)^{1/2}.
\end{aligned}$$

Note that when $\frac{\kappa_1}{\lambda_n^{1/2} n^{1/2}} \le \big(3\log\frac{2}{\delta_n}\big)^{-1}$, the above upper bound simplifies to

$$\|\widehat{f} - \widetilde{f}\|_K \le \lambda_n^{-1}\Delta_1 \le \frac{6\kappa_1 M_n}{\lambda_n n^{1/2}}\log\frac{2}{\delta_n}.$$

When $\frac{\kappa_1}{\lambda_n^{1/2} n^{1/2}} > \left(3 \log \frac{2}{\delta_n}\right)^{-1}$, we have

$$\|\widehat{f} - \widetilde{f}\|_K \le \|\widehat{f}\|_K + \|\widetilde{f}\|_K \le \frac{2M_n}{\lambda_n^{1/2}} \le \frac{6\kappa_1 M_n}{\lambda_n n^{1/2}} \log \frac{2}{\delta_n},$$

where the second inequality follows from (S.2), (S.3) and the definition of $\widehat{f}$ that $\frac{1}{n}\sum_{i=1}^n (y_i - \widehat{f}(\mathbf{x}_i))^2 + \lambda_n \|\widehat{f}\|_K^2 \le \frac{1}{n}\sum_{i=1}^n y_i^2 \le M_n^2$. The desired inequality then follows immediately. ■

Proof of Theorem 3: For simplicity, denote

$$\mathcal{C}_3 = \left\{ \mathcal{Z}^n : \max_{l,k \in \widehat{\mathcal{A}}} \left| \|\widehat{g}_{lk}\|_n^2 - \|g_{lk}^*\|_2^2 \right| > b_{n,2} \log\left(\frac{8p_0^2}{\delta_n}\right) n^{-\frac{(2r-1)}{2(2r+1)}} \right\}.$$

Note that $P(\mathcal{C}_3)$ can be decomposed as

$$\begin{aligned} P(\mathcal{C}_3) &= P\big(\mathcal{C}_3 \cap \{\widehat{\mathcal{A}} = \mathcal{A}^*\}\big) + P\big(\mathcal{C}_2 \cap \{\widehat{\mathcal{A}} \ne \mathcal{A}^*\}\big) \\ &\le P\big(\widehat{\mathcal{A}} \ne \mathcal{A}^*\big) + P\big(\mathcal{C}_3 | \widehat{\mathcal{A}} = \mathcal{A}^*\big) P(\widehat{\mathcal{A}} = \mathcal{A}^*) = \Delta_n + P_3(1 - \Delta_n), \end{aligned}$$

where $\Delta_n \to 0$ according to Theorem 2 in the main text, and P_3 can be bounded as follows.

To bound P_3, we first introduce some additional notations. Denote the operators for the second-order gradients as

$$D_{lk}^* D_{lk} f = \int \partial_{lk}^2 K_{\mathbf{x}} g_{lk}(\mathbf{x}) d\rho_{\mathbf{x}} \quad \text{and} \quad \widehat{D}_{lk}^* \widehat{D}_{lk} f = \frac{1}{n} \sum_{i=1}^n \partial_{lk}^2 K_{\mathbf{x}_i} \widehat{g}_{lk}(\mathbf{x}_i),$$

where $\partial_{lk}^2 K_{\mathbf{x}} = \frac{\partial^2 K(\mathbf{x},\cdot)}{\partial x^l \partial x^k}$. Hence, for any $l, k \in \mathcal{A}^*$, we have

$$\begin{aligned}
&\big|\|\widehat{g}_{lk}\|_n^2 - \|g_{lk}^*\|_2^2\big| \\
&= \Big|\frac{1}{n}\sum_{i=1}^n \big(\widehat{g}_{lk}(\mathbf{x}_i)\big)^2 - \int \big(g_{lk}^*(\mathbf{x})\big)^2 d\rho_{\mathbf{x}}\Big| \\
&= \Big|\frac{1}{n}\sum_{i=1}^n \widehat{g}_{lk}(\mathbf{x}_i)\big\langle \widehat{f}, \partial_{lk}^2 K_{\mathbf{x}_i}\big\rangle_K - \int g_{lk}^*(\mathbf{x})\big\langle f^*, \partial_{lk}^2 K_{\mathbf{x}}\big\rangle_K d\rho_{\mathbf{x}}\Big| \\
&= \Big|\big\langle \widehat{f}, \widehat{D}_{lk}^* \widehat{D}_{lk} \widehat{f}\big\rangle_K - \big\langle f^*, D_{lk}^* D_{lk} f^*\big\rangle_K\Big| \\
&= \Big|\big\langle \widehat{f} - f^*, \widehat{D}_{lk}^* \widehat{D}_{lk}(\widehat{f} - f^*)\big\rangle_K + 2\big\langle f^*, \widehat{D}_{lk}^* \widehat{D}_{lk}(\widehat{f} - f^*)\big\rangle_K + \\
&\qquad\qquad \big\langle f^*, (\widehat{D}_{lk}^* \widehat{D}_{lk} - D_{lk}^* I_{lk}) f^*\big\rangle_K\Big| \\
&\le \kappa_3^2 \|\widehat{f} - f^*\|_K^2 + 2\kappa_3^2 \|f^*\|_K \|\widehat{f} - f^*\|_K + \|f^*\|_K^2 \|\widehat{D}_{lk}^* \widehat{D}_{lk} - D_{lk}^* D_{lk}\|_{HS},
\end{aligned}$$

where the last inequality follows from the Cauthy-Schwartz inequality.

Note that $\|f^*\|_K$ is bounded, and D_{lk} and $\widehat{D}_{lk}$ are Hilbert-Schmidt operators on $\mathcal{H}_K$ by Assumption 5 in the main text and a slightly modified proof of Proposition 6 in Vito et al. [**?**]. It then follows from Rosasco et al. [**?**] that $\max_{l,k \in \mathcal{A}^*} \|\widehat{D}_{lk}^* \widehat{D}_{lk}\|_{HS} \le \kappa_3^2$. Hence, conditional on $\widehat{\mathcal{A}} = \mathcal{A}^*$, we have

$$
\begin{aligned}
&\max_{l,k\in\mathcal{A}^*} \big|\|\widehat{g}_{lk}\|_n^2 - \|g_{lk}^*\|_2^2\big| \\
\leq\ & s_3\Big(\|\widehat{f}-f^*\|_K^2 + 2\|\widehat{f}-f^*\|_K + \max_{l,k\in\mathcal{A}^*} \|\widehat{D}_{lk}^*\widehat{D}_{lk} - D_{lk}^*D_{lk}\|_{HS}\Big) \\
\leq\ & s_3\Big(3\|\widehat{f}-f^*\|_K + \max_{l,k\in\mathcal{A}^*} \|\widehat{D}_{lk}^*\widehat{D}_{lk} - D_{lk}^*D_{lk}\|_{HS}\Big),
\end{aligned}
$$

where $s_3 = \max\{\kappa_3^2, \|f^*\|_K^2, \kappa_3^2\|f^*\|_K\}$, and the second inequality holds when $\|\widehat{f}-f^*\|_K^2$ is sufficiently small. Here, by Theorem 1 in the main text, with probability at least $1-\delta_n/2$, we have $\|\widehat{f}-f^*\|_K$ is bounded. Moreover, for any $\epsilon_n \in (0,1)$ and $l,k\in\mathcal{A}^*$, by the concentration inequalities in $HS(K)$ on $\mathcal{H}_K$ [?], we have

$$
P\Big(\|\widehat{D}_{lk}^*\widehat{D}_{lk} - D_{lk}^*D_{lk}\|_{HS} \geq \epsilon_n\Big) \leq 2\exp\Big(-\frac{n\epsilon_n^2}{8\kappa_3^4}\Big).
$$

Let $\epsilon_n = \big(\frac{8\kappa_3^4}{n}\log\frac{4}{\delta_n}\big)^{1/2}$, then with probability at least $1-\delta_n/2$, there holds

$$
\max_{l,k\in\mathcal{A}^*} \big\|\widehat{D}_{lk}^*\widehat{D}_{lk} - D_{lk}^*D_{lk}\big\|_{HS} \leq \Big(\frac{8\kappa_3^4}{n}\log\frac{4p_0^2}{\delta_n}\Big)^{1/2}.
$$

Therefore, conditional on $\widehat{\mathcal{A}} = \mathcal{A}^*$, we have with probability at least

$1-\delta_n$, there holds

$$\max_{l,k\in\widehat{\mathcal{A}}} \left| \|\widehat{g}_{lk}\|_n^2 - \|g_{lk}^*\|_2^2 \right|$$
$$\leq s_3\Big(3\log\frac{8}{\delta_n}\big(\frac{3\kappa_1}{n^{1/2}\lambda_n}(\kappa_1\|f^*\|_K + q^{-1}(\log\frac{4c_1n}{\delta_n})) + \lambda_n^{r-1/2}\|L_K^{-r}f^*\|_2\big) + \big(\frac{8\kappa_3^4}{n}\log\frac{4p_0^2}{\delta_n}\big)^{1/2}\Big).$$

Furthermore, with $\lambda_n = n^{-\frac{1}{2r+1}}$, the upper bound reduces to

$$\max_{l,k\in\widehat{\mathcal{A}}} \left| \|\widehat{g}_{lk}\|_n^2 - \|g_{lk}^*\|_2^2 \right| \leq b_{n,2}\Big(\log\frac{8p_0^2}{\delta_n}\Big)n^{-\frac{(2r-1)}{2(2r+1)}},$$

where $b_{n,2}$ is given in Theorem 3 of the main text, and hence that $P_3 \leq \delta_n$. Therefore, $P(\mathcal{C}_3) \leq \Delta_n + \delta_n(1-\Delta_n) \leq \Delta_n + \delta_n$, and the desired result follows immediately. ■

Proof of Theorem 4: Note that

$$\begin{aligned}
&P\left(\widehat{\mathcal{A}}_2 = \mathcal{A}_2^*, \widehat{\mathcal{A}}_1 = \mathcal{A}_1^*\right) \\
&= P\left(\widehat{\mathcal{A}}_2 = \mathcal{A}_2^*, \widehat{\mathcal{A}}_1 = \mathcal{A}_1^*, \widehat{\mathcal{A}} = \mathcal{A}^*\right) \\
&= P\left(\widehat{\mathcal{A}}_2 = \mathcal{A}_2^*, \widehat{\mathcal{A}}_1 = \mathcal{A}_1^* | \widehat{\mathcal{A}} = \mathcal{A}^*\right) P\left(\widehat{\mathcal{A}} = \mathcal{A}^*\right) \\
&\geq \left(1 - P\left(\widehat{\mathcal{A}}_2 \neq \mathcal{A}_2^* | \widehat{\mathcal{A}} = \mathcal{A}^*\right) - P\left(\widehat{\mathcal{A}}_1 \neq \mathcal{A}_1^* | \widehat{\mathcal{A}} = \mathcal{A}^*\right)\right) P\left(\widehat{\mathcal{A}} = \mathcal{A}^*\right) \\
&= \left(1 - 2P\left(\widehat{\mathcal{A}}_2 \neq \mathcal{A}_2^* | \widehat{\mathcal{A}} = \mathcal{A}^*\right)\right)(1-\Delta_n),
\end{aligned}$$

where the last equality follows from the fact that $\widehat{A}_1 \cap \widehat{\mathcal{A}}_2 = A_1^* \cap \mathcal{A}_2^* = \emptyset$, and then $\{\widehat{\mathcal{A}}_1 \neq \mathcal{A}_1^*\} = \{\widehat{\mathcal{A}}_2 \neq \mathcal{A}_2^*\}$ given $\widehat{\mathcal{A}} = \mathcal{A}^*$. By Theorem 2 in the main text, $\Delta_n \to 0$ as n diverges. Therefore, it suffices to show $P(\widehat{\mathcal{A}}_2 \neq \mathcal{A}_2^* | \widehat{\mathcal{A}} = \mathcal{A}^*) \to 0$ as n diverges.

We first show that $\mathcal{A}_2^* \subset \widehat{\mathcal{A}}_2$ in probability conditional on $\widehat{\mathcal{A}} = \mathcal{A}^*$. If not, suppose that there exists some $l' \in \mathcal{A}_2^*$, which directly implies that $\|g_{l'k}^*\|_2^2 > b_{n,2} \max\{\kappa_1 \|f^*\|_K, q^{-1}(\log \frac{4c_1 n}{\delta_n})\} n^{-\xi_4} \log p_0$, for some $k \in \mathcal{A}^*$ but $l' \notin \widehat{\mathcal{A}}_2$, and thus $\|\widehat{g}_{l'k}\|_n^2 \leq v_n^{int}$. By Assumption 6 in the main text, we have with probability at least $1 - \Delta_n - \delta_n$ that

$$\big|\|\widehat{g}_{l'k}\|_n^2 - \|g_{l'k}^*\|_2^2\big| \geq \|g_{l'k}^*\|_2^2 - \|\widehat{g}_{l'k}\|_n^2 > \frac{b_{n,2}}{2} \max\{\kappa_1 \|f^*\|_K, q^{-1}(\log \frac{4c_1 n}{\delta_n})\} n^{-\xi_4} \log p_0,$$

which contradicts with Theorem 3 in the main text. This implies that conditional on $\widehat{\mathcal{A}} = \mathcal{A}^*$, $\mathcal{A}_2^* \subset \widehat{\mathcal{A}}_2$ with probability at least $1 - \Delta_n - \delta_n$.

Next, we show that $\widehat{\mathcal{A}}_2 \subset \mathcal{A}_2^*$ in probability conditional on $\widehat{\mathcal{A}} = \mathcal{A}^*$. If not, suppose there exists some $l' \in \widehat{\mathcal{A}}_2$ but $l' \notin \mathcal{A}_2^*$, which implies $\|\widehat{g}_{l'k}\|_n^2 > v_n^{int}$ for some $k \in \mathcal{A}^*$ but $\|g_{l'k}^*\|_2^2 = 0$. Then with probability at least $1 - \Delta_n - \delta_n$, there holds

$$\big|\|\widehat{g}_{l'k}\|_n^2 - \|g_{l'k}^*\|_2^2\big| = \|\widehat{g}_{l'k}\|_n^2 > \frac{b_{n,2}}{2} \max\{\kappa_1 \|f^*\|_K, q^{-1}(\log \frac{4c_1 n}{\delta_n})\} n^{-\xi_4} \log p_0,$$

which contradicts with Theorem 3 in the main text again. Therefore, conditional on $\widehat{\mathcal{A}} = \mathcal{A}^*$, $\widehat{\mathcal{A}}_2 \subset \mathcal{A}_2^*$ with probability at least $1 - \Delta_n - \delta_n$.

Combining these two results yields that $P(\widehat{\mathcal{A}}_2 = \mathcal{A}_2^* | \widehat{\mathcal{A}} = \mathcal{A}^*) \geq 1 - 2\Delta_n - 2\delta_n$, or equivalently, $P\big(\widehat{\mathcal{A}}_2 \neq \mathcal{A}_2^* | \widehat{\mathcal{A}} = \mathcal{A}^*\big) \leq 2\Delta_n + 2\delta_n \to 0$. The desired sparsistency then follows immediately. ∎

S1.1 Verification of theoretical examples

The following two additional assumptions are made to establish the sparsistency.

Assumption S1: There exist some positive constant τ_1 such that the smallest eigenvalue of $\mathrm{E}(\mathbf{x}\,\mathbf{x}^T)$, $\lambda_{\min}(\mathrm{E}(\mathbf{x}\,\mathbf{x}^T)) = O(n^{-\tau_1})$.

Assumption S2: There exist some positive constants s_1 and $\xi_2 > 1/3$ such that $\min\limits_{l \in \mathcal{A}^*} |\boldsymbol{\beta}_l^*| > s_1 p_n^{1/6} n^{-\frac{1-2\tau_1}{6}} (\log n)^{\xi_2}$.

Assumption S1 implies that $\mathrm{E}(\mathbf{x}\,\mathbf{x}^T)$ is invertible, and that Assumption 1 in the main text is satisfied for the scaled linear kernel with $r = 1$. Assumption 2 in the main text is also satisfied due to the fact that $\|\widetilde{\mathbf{x}}\|^2 = p_n^{-1}\mathbf{x}^T\mathbf{x}$ belongs to a compact set $\mathcal{X}$. A similar assumption is made in Shao and Deng [1], assuming the decay order of the smallest eigenvalue of $n^{-1}\,\mathbf{X}^T\,\mathbf{X}$. Assumption S2 is similar to Assumption 3 in the main text, and requires the true regression coefficients contains sufficient information about the truly informative variables in the linear model. Similar assumptions are also assumed in Shao and Deng [1] and Wang and Leng [4].

Proof of Corollary 1**:** The estimation consistency for the linear case is a direct application of Theorem 1 in the main text for the scaled linear kernel $K(\mathbf{x}, \mathbf{u}) = \mathbf{x}^T\mathbf{u}/p_n$, and we just need to verify the assumptions of Theorem 1 in the main text . In fact, Assumption S1 implies that $\mathrm{E}(\mathbf{x}\,\mathbf{x}^T)$ is invertible, and thus Assumption 1 in the main text is satisfied for the scaled linear kernel with $r = 1$. Assumption 2 in the main text is also satisfied due to the fact that $\sup_{\mathbf{x}\in\mathcal{X}} \|K_{\mathbf{x}}\|_K = p_n^{-1/2}\|\mathbf{x}\|$ belongs to a compact set $\mathcal{X} \subset \mathcal{R}^{p_n}$. Furthermore,

$$\begin{aligned}\|L_K^{-1}f^*\|_2 &= \|(\mathrm{E}\,\widetilde{\mathbf{x}}\widetilde{\mathbf{x}}^T)^{-1}\boldsymbol{\beta}^*\|_2 = (\boldsymbol{\beta}^{*T}(\mathrm{E}\,\mathbf{x}\,\mathbf{x}^T/p_n)^{-1}\boldsymbol{\beta}^*)^{1/2}\\ &\le p_n^{1/2}\lambda_{min}(\mathrm{E}(\mathbf{x}\,\mathbf{x}^T))^{-1/2}\|\boldsymbol{\beta}^*\| = O(p_n^{1/2}n^{\tau_1/2}),\end{aligned}$$

where $\|\boldsymbol{\beta}^*\|$ is a bounded quantity. Then, following from Theorem 1 in the main text, let $\lambda_n = O(p_n^{1/3}n^{-(1+\tau_1)/3}(\log n)^{2/3})$, for any $\delta_n \ge 4(\sigma^2 + \|\boldsymbol{\beta}^*\|_2^2)(\log n)^{-2}$, there exists some positive constant c_3 such that, with probability at least $1 - \delta_n$, there holds

$$\|\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*\| \le c_3 \log\Big(\frac{4}{\delta_n}\Big)p_n^{1/6}n^{-\frac{1-2\tau_1}{6}}(\log n)^{1/3}. \tag{S.5}$$

To establish the selection consistency, note that $\mathcal{A}^* = \{l : \beta_l^* \neq 0\}$ and $\widehat{\mathcal{A}}_{v_n} = \{l : |\widehat{\beta}_l| > v_n\}$. Clearly, (S.5) directly implies that for any $l = 1, ..., p_n$, with probability at least $1 - \delta_n$, there holds

$$|\widehat{\beta}_l - \beta_l^*| \leq 2c_3 \log \frac{4}{\delta_n} p_n^{1/6} n^{-\frac{1-2\tau_1}{6}} (\log n)^{1/3}.$$

Therefore, following the proof of Theorem 2 in the main text and let $v_n = \frac{s_1}{2} p_n^{1/6} n^{-\frac{1-2\tau_1}{6}} (\log n)^{\xi_2}$, we have $P(\widehat{\mathcal{A}}_{v_n} = A^*) \to 1$. ■

Additional assumptions are made to establish the sparsistency for the proposed method with quadratic kernel.

Assumption S3: There exists a positive constant τ_2 such that the smallest eigenvalue of $\mathrm{E}(\bar{\mathbf{x}}\bar{\mathbf{x}}^T)$, $\lambda_{\min}(\mathrm{E}(\bar{\mathbf{x}}\bar{\mathbf{x}}^T)) = O(n^{-\tau_2})$.

Assumption S4: There exist some positive constants s_2 and $\xi_3 > 1/3$ such that $\min\limits_{l \in \mathcal{A}^*} |\beta_l^*| + \sum_{k=1}^{p_n} |\gamma_{lk}^*| > s_2 p_n^{1/3} n^{-\frac{1-2\tau_2}{6}} (\log n)^{\xi_3}$.

Assumptions S3 and S4 can be regarded as an extension of Assumptions S1 and S2 by requiring $\mathrm{E}(\bar{\mathbf{x}}\bar{\mathbf{x}}^T)$ is invertible, and that the true regression coefficients have sufficient information in the quadratic model.

References

[1] Shao, J. and Deng, X. (2013). Estimation in high-dimensional linear models with deterministic design matrices. *Annals of Statistics*, **40**, 1821–1831.

[2] She, Y. and Wang, Z. and Jiang, H. (2018). Group regularized estimation under structural hierarchy. *Journal of the American Statistical Association*, **113**, 445–454 .

[3] Smale, S. and Zhou, D. (2005). Shannon sampling II: connections to learning theory. *Applied and Computational Harmonic Analysis*, **19**, 285–302.

[4] Wang, X. and Leng, C. (2016). High dimensional ordinary least squares projection for screening variables. *Journal of the Royal Statistical Society, Series B*, **78**, 589–611.

[5] Zhou, D. (2007). Derivative reproducing properties for kernel methods in learning theory. *Journal of Computational and Applied Mathematics*, **220**, 456–463.

School of Statistics and Management, Shanghai University of Finance and Economics, Shanghai 200433, China

E-mail: he.xin17@mail.shufe.edu.cn

School of Data Science, City University of Hong Kong, Hong Kong, China

E-mail: j.h.wang@cityu.edu.hk

School of Statistics and Mathematics, Nanjing Audit University, China

E-mail: lvsg716@swufe.edu.cn