Supplementary Material for "Unified Tests for Nonparametric Functions in RKHS with Kernel Selection and Regularization"

Tao He[1], Ping-Shou Zhong[2], Yuehua Cui[3] and Vidyadhar Mandrekar[3]

[1]*San Francisco State University,* [2]*University of Illinois at Chicago*

and [3]*Michigan State University*

Abstract: The technical proofs of lemmas and main results are included in Section S1. The details of the regularized kernel and its oracle property are included in Section S2. Some additional simulation results are included in Section S3.

S1 Lemmas and the proof of main results

Lemma 1. *Assume that K is a positive semi-definite and symmetric kernel defined on $\mathscr{X} \times \mathscr{X}$. Let μ be a finite measure on $\mathscr{X}$. If*

$$\int_{\mathscr{X} \times \mathscr{X}} K^2(x, y) d\mu(x) d\mu(y) < \infty, \tag{S1.1}$$

then $K(x, y) = \sum_{j=1}^{\infty} \lambda_j \psi_j(x) \psi_j(y)$, $\{\psi_j(\cdot)\} \subset L^2(\mu)$ form a complete orthogonal normal system i.e., $E\{\psi_j(X)\psi_k(X)\} = \delta_{jk}$ where $\delta_{jk} = 1$ if $j = k$; $\delta_{jk} = 0$ if $j \neq k$.

Proof: Given a positive definite kernel $K(x, y)$, we can construct a reproducing kernel

Hilbert space (RKHS) $\mathscr{H}_K$, and the reproducing property implies that

$$K(x,y) =< K(x,\cdot), K(y,\cdot) >_{\mathscr{H}_K} :=< K_x, K_y >_{\mathscr{H}_K} .$$

Since for any $K_x \in \mathscr{H}_K$, $\|K_x\| = \sqrt{< K_x, K_x >_{\mathscr{H}_K}} = \sqrt{K(x,x)}$, we have

$$|K(x,y)| =< K_x, K_y >_{\mathscr{H}_K} \leq \|K_x\|\|K_y\| = \sqrt{K(x,x)K(y,y)},$$

and

$$\int_{\mathscr{X}\times\mathscr{X}} K^2(x,y)d\mu(x)d\mu(y) < \infty. \tag{S1.2}$$

Therefore, $K(x,y)$ generates a compact operator on $L^2(\mu)$ through the integral operation $(Kf)(x) = \int_{\mathscr{X}} K(x,y)f(y)d\mu(y)$. Let $\{\lambda_j\}_{j=1}^{\infty}$ and $\{\psi_j(\cdot)\}_{j=1}^{\infty}$ be the eigenvalues and corresponding complete orthogonal normal system of kernel K under measure μ, i.e.,

$$\int K(x,y)\psi_i(y)d\mu(y) = \lambda_i\psi_i(x), \quad i = 1, 2, \cdots, \infty. \tag{S1.3}$$

Since $K(x,y) \in L^2(\mu \bigotimes \mu)$, $K_x(\cdot) = K(x,\cdot) \in L^2(\mu)$, i.e., there exist $\{c_m(x)\}_{m=1}^{\infty}$ such that $K(x,y) = K_x(y) = \sum_m c_m(x)\psi_m(y)$, then we have $K_y(\cdot) = \sum_{m=1} c_m(\cdot)\psi_m(y)$. Because $K_y(\cdot) \in L^2(\mu)$ and $\{\psi_m(y)\}_{m=1}^{\infty}$ can be considered as constants once y is given, then $c_m(\cdot) \in L^2(\mu)$ and can be expanded using bases $\{\psi_m(\cdot)\}_{m=1}^{\infty}$. Therefore, we have $K(x,y) = \sum_{i,j=1}^{\infty} a_{ij}\psi_i(x)\psi_j(y)$, where $\sum_{i,j} a_{ij}^2 < \infty$ is due to (S1.2).

It will be shown in the following that $a_{ij} = \lambda_i\delta_{ij}$, which implies $K(x,y) = \sum_{j=1}^{\infty} \lambda_j\psi_j(x)\psi_j(y)$.

Actually,

$$\int\int K(x,y)\psi_i(x)\psi_j(y)d\mu(x)d\mu(y) = \int\int\sum_{k,l} a_{kl}\psi_k(x)\psi_l(y)\psi_i(x)\psi_j(y)d\mu(x)d\mu(y)$$

where the left hand side is $\int \lambda_i\psi_i(y)\psi_j(y)d\mu(y) = \lambda_i\delta_{ij}$ by using the eigen-decomposition property (S1.3), and the right hand side is $\sum_{k,l} a_{kl}\delta_{ki}\delta_{lj} = a_{ij}$. This finishes the proof of Lemma 1. □

Corollary 1. *Let $\mathbf{Z}_1,\cdots,\mathbf{Z}_n$ be q-dim IID random vectors where $q = q(n)$ is a function of n, and $q(n)\to\infty$ as $n\to\infty$. Let $H_n(x,y)$ be a symmetric kernel function of $x,y\in R^q$, which may depend on n. The kernel function satisfies $E\{H_n^2(\mathbf{Z}_1,\mathbf{Z}_2)\} = \sigma_{nH}^2$ and $E\{H_n(\mathbf{Z}_1,\mathbf{Z}_2)|\mathbf{Z}_2\} = 0$ almost surely. Then, there exists eigenvalues λ_{nk} and a complete orthonormal functions $(\psi_{nk})_{k=0,1,2,\cdots}$ such that $\sigma_{nH}^{-1}H_n(x,y) = \sum_{k=1}^{\infty}\lambda_{nk}\psi_{nk}(x)\psi_{nk}(y)$.*

Proof: Because the eigenvalue decomposition in Lemma 1 is applicable to a semi-definite and symmetric kernel defined in any space $\mathscr{X}$, we can apply it to the kernel function $\sigma_H^{-1}H_n(x,y)$ defined in R^q for any q. By definition $\sigma_H^{-2}E\{H_n^2(\mathbf{Z}_1,\mathbf{Z}_2)\} = 1 < \infty$ and Lemma 1, there exist eigenvalues and a complete orthonormal functions $(\psi_{nk})_{k=0,1,2,\cdots}$ such that $\psi_{n0} = 1$ and $\lambda_{nk}\psi_{nk}(x) = \int\sigma_{nH}^{-1}H_n(x,y)\psi_{nk}(y)d\mu_n(y)$. □

Lemma 2. *Let $\mathbf{Z}_1,\cdots,\mathbf{Z}_n$ be q-dim IID random vectors where $q = q(n)$ is a function of n, and $q(n)\to\infty$ as $n\to\infty$. Let $H_n(x,y)$ be a symmetric kernel function of $x,y\in R^q$, which may depend on n. The kernel function satisfies $E\{H_n^2(\mathbf{Z}_1,\mathbf{Z}_2)\} = \sigma_{nH}^2$ and $E\{H_n(\mathbf{Z}_1,\mathbf{Z}_2)|\mathbf{Z}_2\} = 0$ almost surely. Define $U_n = \{n(n-1)\}^{-1}\sum_{i\neq j}^{n} H_n(\mathbf{Z}_i,\mathbf{Z}_j)$,* *as a degenerate U-statistic. Let λ_{nk} and $\psi_{nk}(x)$ be eigenvalues and the corresponding eigenfunctions defined in Corollary 1. Under conditions*

(A1): $\sum_{k=N}^{\infty} \lambda_{nk}^2 \to 0$ uniformly for all $n > n_0$ as $n_0, N \to \infty$, and $\lambda_{nk} \to \lambda_k \neq 0$;

(A2): $\sup_n \sigma_{nH}^{-(2+\delta)} E|H_n(\mathbf{Z}_1, \mathbf{Z}_2)|^{2+\delta} < \infty$ for some $\delta > 0$.

Then, as $n \to \infty$, we have $n\sigma_{nH}^{-1} U_n \xrightarrow{d} \sum_{k=1}^{\infty} \lambda_k (\chi_k^2 - 1)$, where χ_k^2 are IID chi-square distributed random variables.

Proof: Let $a_n = \{\sqrt{2n(n-1)}\}^{-1}$. To prove $n\sigma_H^{-1} U_n \xrightarrow{d} \sum_{k=1}^{\infty} \lambda_k (\chi_k^2 - 1)$ is equivalent to prove

$$\sigma_{nH}^{-1} \sum_{i \neq j}^{n} a_n H_n(\mathbf{Z}_i, \mathbf{Z}_j) \xrightarrow{d} \sum_{k=1}^{\infty} \lambda_k (\chi_k^2 - 1)/\sqrt{2}. \tag{S1.4}$$

Using the eigenvalue decomposition in Corollary 1, it follows that

$$\sum_{i \neq j}^{n} a_n \sigma_{nH}^{-1} H_n(\mathbf{Z}_i, \mathbf{Z}_j) \stackrel{d}{=} \sum_{k=1}^{\infty} \lambda_k \sum_{i \neq j}^{n} a_n \psi_{nk}(\mathbf{Z}_i) \psi_{nk}(\mathbf{Z}_j) \equiv u_n,$$

where $\stackrel{d}{=}$ means equality in distribution. Let Y_j for $j = 1, \cdots, n$ be IID standard normal random variables. Then, to prove (S1.18) is equivalent to prove

$$\rho(u_n, v_n) \to 0, \tag{S1.5}$$

where $v_n = \sum_{k=1}^{\infty} \lambda_k (\sum_{j \neq l} a_n Y_j Y_l)$ and $\rho(A, B)$ is the Levy-Prokhorov distance between the probability laws of A and B (Billingsley (2013)). By carefully checking the proof of Mikosch (1993), Theorem 2.1 in Mikosch (1993) can be applied to prove (S1.5). Thus, we only need to check conditions (A1)-(A3) in Theorem 2.1 of Mikosch (1993). Condition (A1) in Mikosch (1993) can be slightly relaxed to (A1) given in this Lemma 2. For the condition (A2) in Mikosch (1993), we can see that $d^{-2}(n) = 1/\{2(n-1)\} \to 0$, therefore condition (A2) in Mikosch (1993) holds. Condition (A3) in Mikosch (1993) also holds because of the remark (5) below Theorem 2.1 in Mikosch (1993) suggests that condition (A2) is a sufficient condition. □

Lemma 3. *If kernel $K^*_\theta(x_1, x_2)$ is a positive definite kernel, then the centralized kernel $K_\theta(x_1, x_2)$ is positive semi-definite.*

Proof: Assume kernel function $K^*_\theta(x, y)$ has eigen-depcomposition $\{\lambda^*_{nm}, \psi^*_{nm}(\cdot)\}_{m=1}^{\infty}$, and centralized kernel $K_\theta(x, y)$ has eigen-decomposition $\{\lambda_{nm}, \psi_{nm}(\cdot)\}_{m=1}^{\infty}$. Since the kernel can be normalized, we assume the sum of squared eigenvalues are bounded without loss of generality. Recall the definition of the centralized kernel function $K_\theta(\mathbf{x}_1, \mathbf{x}_2) = K^*_\theta(\mathbf{x}_1, \mathbf{x}_2) - K^*_{1,\theta}(\mathbf{x}_1) - K^*_{1,\theta}(\mathbf{x}_2) + \mu_{K^*}$. Then we have $\mathrm{E}[K_\theta(\mathbf{x}_1, \mathbf{X}_2)] = \mathrm{E}[K^*_\theta(\mathbf{x}_1, \mathbf{X}_2)] - K^*_{1,\theta}(\mathbf{x}_1) = 0$, or equivalently, $\int 1 \cdot K_\theta(\mathbf{x}_1, \mathbf{x}_2) d\mu(\mathbf{x}_2) = 0$, which implies that $\psi_{nm^*}(\cdot) = 1$ is one of the eigenfunctions corresponding to zero eigenvalue. Due to the orthogonality of the system, $\mathrm{E}\{\psi_{nm}(\mathbf{X})\} = 0$ for $m \neq m^*$. By the eigen-decomposition equality in Lemma 1 we have

$$\lambda_{nm}\psi_{nm}(\mathbf{x}_1) = \mathrm{E}\{K_\theta(\mathbf{x}_1, \mathbf{X}_2)\psi_{nm}(\mathbf{X}_2)\} = \mathrm{E}\{K^*_\theta(\mathbf{x}_1, \mathbf{X}_2)\psi_{nm}(\mathbf{X}_2)\} - \mathrm{E}\{K^*_{1,\theta}(\mathbf{X}_2)\psi_{nm}(\mathbf{X}_2)\}$$
$$+ \{\mu_{K^*} - K^*_{1,\theta}(\mathbf{x}_1)\}\mathrm{E}\{\psi_{nm}(\mathbf{X}_2)\} = \mathrm{E}\{K^*_\theta(\mathbf{x}_1, \mathbf{X}_2)\psi_{nm}(\mathbf{X}_2)\} - \mathrm{E}\{K^*_{1,\theta}(\mathbf{X}_2)\psi_{nm}(\mathbf{X}_2)\},$$

for any $m \neq m^*$. By plugging in $K^*_\theta(\mathbf{x}_1, \mathbf{x}_2) = \sum_{m=1}^\infty \lambda^*_{nm}\psi^*_{nm}(\mathbf{x}_1)\psi^*_{nm}(\mathbf{x}_2)$, and multiplying $\psi_{nm}(\mathbf{x}_1)$ to both sides, we have

$$\lambda_{nm}\psi^2_{nm}(\mathbf{x}_1) = \sum_s \lambda^*_{ns}\psi^*_{ns}(\mathbf{x}_1)\psi_{nm}(\mathbf{x}_1)\mathrm{E}\{\psi^*_{ns}(\mathbf{X}_2)\psi_{nm}(\mathbf{X}_2)\} - \mathrm{E}\{K^*_{1,\theta}(\mathbf{X}_2)\psi_{nm}(\mathbf{X}_2)\}\psi_{nm}(\mathbf{x}_1),$$

for $m \neq m^*$. Taking expectation with respect to $\mathbf{X}_1$ and using the orthogonal normal property,

$$\lambda_{nm} = \sum_{m=1}^\infty \lambda^*_{nm}\mathrm{E}^2[\psi^*_{nm}(\mathbf{X})\psi_{nm}(\mathbf{X})] \geq 0, \quad m \neq m^*.$$

In addition, $\lambda_{nm^*} = 0$, then the positive semi-definiteness of centralized kernel function can be achieved. □

Proof of Theorem 1: (i) Under the null hypothesis, $Y_i = \mu + \epsilon_i$. Because the test statistic T_n is invariant to location shift, without loss of generality, we assume $\mu = 0$ in the following proof. Then $T^0_n := T^{H_0}_n$, the reduced version under null hypothesis, can be written as

$$T^0_n = \frac{1}{n(n-1)\sigma^2}\sum_{i\neq j} K(\mathbf{X}_i, \mathbf{X}_j)(\epsilon_i - \bar{\epsilon})(\epsilon_j - \bar{\epsilon})\{1 + (\frac{\sigma^2}{\hat{\sigma}^2} - 1)\} := T^0_{n1}\{1 + (\frac{\sigma^2}{\hat{\sigma}^2} - 1)\} \quad \text{(S1.6)}$$

Since $\sigma^2/\hat{\sigma}^2 - 1 = o_p(1)$, under the null, we have $T^0_n/\sqrt{V_{2n}} = T^0_{n1}/\sqrt{V_{2n}}\{1 + o_p(1)\}$.

We now study the asymptotic distribution of $T^0_{n1}/\sqrt{V_{2n}}$ using the U-statistic theory. By plugging in the full expression of $\bar{\epsilon} = n^{-1}\sum_{i=1}^n \epsilon_i$, the leading order of $T^0_{n1}/\sqrt{V_{2n}}$ can be written as the sum of three U-statistics of different orders

$$\frac{T_{n1}^0}{\sqrt{V_{2n}}} = U_n^{(2)} + U_n^{(3)} + U_n^{(4)} + \Delta_n^0, \tag{S1.7}$$

where $U_n^{(2)} = \sum_{i \neq j} \Psi^{(2)}(\mathbf{Z}_i, \mathbf{Z}_j)/P_n^2$,

$$U_n^{(3)} = \frac{1}{P_n^3} \sum_{i \neq j \neq k} \Psi^{(3)}(\mathbf{Z}_i, \mathbf{Z}_j, \mathbf{Z}_k), \quad U_n^{(4)} = \frac{1}{P_n^4} \sum_{i \neq j \neq k \neq l} \Psi^{(4)}(\mathbf{Z}_i, \mathbf{Z}_j, \mathbf{Z}_k, \mathbf{Z}_l),$$

$\Delta_n^0 = o_p(U_n^{(2)} + U_n^{(3)} + U_n^{(4)})$, and $\mathbf{Z} = (\mathbf{X}, \epsilon)$. P_n^k is the number of k-permutations of n, $\Psi^{(k)}$ is the kernel function of k-th order U-statistic $U_n^{(k)}$ for $k = 2, 3, 4$ and of the following symmetric form $\Psi^{(2)}(\mathbf{Z}_i, \mathbf{Z}_j) = \mathcal{K}(\mathbf{X}_i, \mathbf{X}_j)[\epsilon_i \epsilon_j - n^{-1}(\epsilon_i + \epsilon_j)^2]/\sigma^2$, $\Psi^{(3)}(\mathbf{Z}_i, \mathbf{Z}_j, \mathbf{Z}_k) = \varphi^{(3)}(i, j, k) + \varphi^{(3)}(j, k, i) + \varphi^{(3)}(i, k, j)$ and

$$\begin{aligned}\Psi^{(4)}(\mathbf{Z}_i, \mathbf{Z}_j, \mathbf{Z}_k, \mathbf{Z}_l) =& \varphi^{(4)}(i, j, k, l) + \varphi^{(4)}(i, k, j, l) + \varphi^{(4)}(i, l, j, k) + \varphi^{(4)}(j, k, i, l) \\ &+ \varphi^{(4)}(j, l, i, k) + \varphi^{(4)}(k, l, i, j),\end{aligned}$$

where $\varphi^{(3)}(i, j, k) = -\mathcal{K}(\mathbf{X}_i, \mathbf{X}_j)(\epsilon_i \epsilon_k + \epsilon_j \epsilon_k - \epsilon_k^2/n)/(3\sigma^2)$ and $\varphi^{(4)}(i, j, k, l) = \mathcal{K}(\mathbf{X}_i, \mathbf{X}_j)\epsilon_k \epsilon_l/(6\sigma^2)$. To study the distribution of $T_{n1}^0/\sqrt{V_{2n}}$, we will look at the asymptotic properties of each U-statistic $U_n^{(k)}$ respectively. Specifically, we are going to show the following

$$(a): \quad nU_n^{(2)} \xrightarrow{d} \sum_{m=1}^{\infty} \lambda_{\mathcal{K},m}(\chi_m^2 - 1), \tag{S1.8}$$

$$(b): \quad nU_n^{(3)} \xrightarrow{p} 0, \tag{S1.9}$$

$$(c): \quad nU_n^{(4)} \xrightarrow{p} 0. \tag{S1.10}$$

To see (a), we define the first-order and second-order projections of the kernel $\Psi^{(2)}(\cdot)$ as $\phi_1^{(2)}(\mathbf{z}_i) = E\{\Psi^{(2)}(\mathbf{z}_i, \mathbf{Z}_j)\} = 0$ and $\phi_2^{(2)}(\mathbf{z}_i, \mathbf{z}_j) = E\{\Psi^{(2)}(\mathbf{z}_i, \mathbf{z}_j)\} = \Psi_2^{(2)}(\mathbf{z}_i, \mathbf{z}_j)$, and their corresponding variances $\sigma_{2,1}^2 = \text{Var}[\phi_1^{(2)}(\mathbf{Z}_i)], \sigma_{2,2}^2 = \text{Var}[\phi_2^{(2)}(\mathbf{Z}_i, \mathbf{Z}_j)]$. It is not difficult to prove that $U_n^{(2)}$ is first-order degenerated, i.e., $\sigma_{2,1}^2 = 0$ and $\sigma_{2,2}^2 = 2V_{\mathcal{K},2}\{1 + o(1)\} \neq 0$. By Lemma 2, $nU_n^{(2)} \xrightarrow{d} \sum_{m=1}^{\infty} \lambda_{\mathcal{K}_z,m}(\chi_m^2 - 1)$, where $\{\lambda_{\mathcal{K}_z,m}\}_{m=1}^{\infty}$ are the eigenvalues of kernel function $\mathcal{K}_z(z_1, z_2) = \mathcal{K}(x_1, x_2)\epsilon_1\epsilon_2$ with respect to the distribution function F_z, i.e., solution of integral equations

$$\int \mathcal{K}_z(z_1, z_2)\psi_{\mathcal{K}_z,m}(z_1)dF_z(z_2) = \lambda_{\mathcal{K}_z,m}\psi_{\mathcal{K}_z,m}(z_2), m = 1, \ldots, \infty. \tag{S1.11}$$

It remains to prove that $\lambda_{\mathcal{K}_z,m} = \lambda_{\mathcal{K},m}$. View kernel $\mathcal{K}_z(z_1, z_2)$ as the product of kernel $\mathcal{K}_{z,1}(z_1, z_2) := \mathcal{K}(x_1, x_2)$ and kernel $\mathcal{K}_{z,2}(z_1, z_2) := \epsilon_1\epsilon_2$, where $\mathcal{K}_{z,2}$ has only one non-zero eigenvalue 1 with eigenfunction $g(\epsilon) = \epsilon/\sigma$ under the null hypothesis. Through equations (S1.11) above, it can be verified that eigenvalues and eigenfunctions of $\mathcal{K}_z(z_1, z_2)$ are $\{\lambda_{\mathcal{K},m}\}$ and $\{\psi_m(x) \cdot g(\epsilon)\}_{m=1}^{\infty}$ respectively. (b) and (c) can be achieved similarly by proving means and variances of the first- and second-order projections of $U_n^{(3)}$ and $U_n^{(4)}$ are all zero.

(ii) Based on the proof in part (i), we will only need to show that $nU_n^{(2)}/\sqrt{2} \xrightarrow{d} N(0, 1)$. To this end, we write $nU_n^{(2)}/\sqrt{2} = \sqrt{2}\sum_{i=1}^{n} \xi_i/(n-1) = a_n \sum_{i=1}^{n} \xi_i$ where $\xi_i = \sum_{j=1}^{i-1} \Psi^{(2)}(\mathbf{Z}_i, \mathbf{Z}_j)$ and $a_n = \sqrt{2}/(n-1)$. Let $\mathscr{F}_0 = \{\emptyset, \Omega\}$, $\mathscr{F}_i = \sigma\{\mathbf{Z}_1, \cdots, \mathbf{Z}_i\}$ be the σ-field generated by $\mathbf{Z}_1, \cdots, \mathbf{Z}_i$. Then, we know that $E\{\xi_i|\mathscr{F}_{i-1}\} = 0$ and $\{\xi_i, 1 \leq i \leq n\}$

is a martingale difference sequence with respect to the σ-fields $\{\mathscr{F}_i, 1 \leq i \leq n\}$. Let $\sigma_{ni}^2 = E(\xi_i^2|\mathscr{F}_{i-1})$. Then, to show the central limit theorem, it suffices to show that

$$\frac{\sum_{i=1}^n \sigma_{ni}^2}{\mathrm{Var}(\sum_{i=1}^n \xi_i)} \xrightarrow{p} 1 \quad \text{and} \quad \frac{\sum_{i=1}^n E(\xi_i^4)}{\mathrm{Var}^2(\sum_{i=1}^n \xi_i)} \to 0. \tag{S1.12}$$

We first show the first part of (S1.12). Because $E(\sum_{i=1}^n \sigma_{ni}^2) = \mathrm{Var}(\sum_{i=1}^n \xi_i)$, it suffices to show that $\mathrm{Var}(\sum_{i=1}^n \sigma_{ni}^2) = O\{\mathrm{Var}^2(\sum_{i=1}^n \xi_i)\}$. By definition, we have

$$\sigma_{ni}^2 = \sum_{j,k}^{i-1} \mathcal{K}_2(X_j, X_k)\zeta_{jk,n},$$

where $\mathcal{K}_2(x, y) = E\{\mathcal{K}(x, \mathbf{X})\mathcal{K}(\mathbf{X}, y)\}$ and $\zeta_{jk,n} = (1-\frac{2}{n})^2\sigma^2\epsilon_j\epsilon_k + n^{-2}(\kappa_4 + \sigma^2\epsilon_k^2 + \sigma^2\epsilon_j^2 + \epsilon_j^2\epsilon_k^2)$. Because $E\{\mathcal{K}_2(\mathbf{X}_j, \mathbf{X}_k)\} = 0$ for $j \neq k$, we have

$$E(\sigma_{ni}^2) = (i-1)E\{\mathcal{K}_2(\mathbf{X}_1, \mathbf{X}_1)\}E(\zeta_{n1}),$$

where $\zeta_{j,n} = (1-\frac{2}{n})^2\sigma^2\epsilon_j^2 + n^{-2}(\kappa_4 + 2\sigma^2\epsilon_j^2 + \epsilon_j^4)$. Moreover, because $E\{\mathcal{K}_2(\mathbf{X}_j, \mathbf{X}_j)\mathcal{K}_2(\mathbf{X}_j, \mathbf{X}_t)\} = 0$ for $j \neq t$, we have

$$\begin{aligned}
E(\sigma_{ni}^4) = {} & \sum_{j=1}^{i-1} E\{\mathcal{K}_2^2(\mathbf{X}_j, \mathbf{X}_j)\zeta_{j,n}\} + \sum_{j\neq s}^{i-1} E\{\mathcal{K}_2(\mathbf{X}_j, \mathbf{X}_j)\mathcal{K}_2(\mathbf{X}_s, \mathbf{X}_s)\zeta_{j,n}\zeta_{s,n}\} \\
& + 2\sum_{j\neq k}^{i-1} E\{\mathcal{K}_2^2(\mathbf{X}_j, \mathbf{X}_k)\zeta_{jk,n}^2\} \\
= {} & (i-1)E\{\mathcal{K}_2^2(\mathbf{X}_1, \mathbf{X}_1)\}E(\zeta_{1,n}^2) + (i-1)(i-2)E^2\{\mathcal{K}_2(\mathbf{X}_1, \mathbf{X}_1)\}E^2(\zeta_{1,n}) \\
& + 2(i-1)(i-2)E\{\mathcal{K}_2^2(\mathbf{X}_1, \mathbf{X}_2)\}E(\zeta_{12,n}^2).
\end{aligned}$$

Similarly, it can be shown that, for $i < j$,

$$E(\sigma_{ni}^2\sigma_{nj}^2) = (i-1)E\{\mathcal{K}_2^2(\mathbf{X}_1, \mathbf{X}_1)\}E(\zeta_{1,n}^2) + (i-1)(j-2)E^2(\zeta_{n1})E^2(\zeta_{1,n})$$
$$+ 2(i-1)(i-2)E\{\mathcal{K}_2^2(\mathbf{X}_1, \mathbf{X}_2)\}E(\zeta_{12,n}^2).$$

Thus, we have

$$\mathrm{Var}(\sum_{i=1}^n \sigma_{ni}^2) = \{\sum_{i=2}^n (i-1) + 2\sum_{i<j}(i-1)\}\mathrm{Var}\{\mathcal{K}_2(\mathbf{X}_1, \mathbf{X}_1)\}\mathrm{Var}(\zeta_{1,n})$$
$$+ 2\{\sum_{i=2}^n (i-1)(i-2) + 2\sum_{i<j}(i-1)(i-2)\}E\{\mathcal{K}_2^2(\mathbf{X}_1, \mathbf{X}_2)\}E(\zeta_{12,n}^2).$$

Because $E(\epsilon_1^8) < \infty$, the order of $\mathrm{Var}(\sum_{i=1}^n \sigma_{ni}^2) = O[\max\{n^4E\{\mathcal{K}_2^2(\mathbf{X}_1, \mathbf{X}_2), n^3E\{\mathcal{K}_2^2(\mathbf{X}_1, \mathbf{X}_1)\}]$. Moreover, we know that $\mathrm{Var}(\sum_{i=1}^n \xi_i)$ is at the order of n^2. So, $\mathrm{Var}(\sum_{i=1}^n \sigma_{ni}^2) = o\{\mathrm{Var}^2(\sum_{i=1}^n \xi_i)\}$ if $\max(E\{\mathcal{K}_2^2(\mathbf{X}_1, \mathbf{X}_2)\}, E\{\mathcal{K}_2^2(\mathbf{X}_1, \mathbf{X}_1)\}/n) = o(1)$. Define $K_2(x, y) = E\{K(x, \mathbf{X})K(y, \mathbf{X})\}$. If $E\{K_2^2(\mathbf{X}_1, \mathbf{X}_2)\} = o(V_{2n}^2)$ and $E\{K_2^2(\mathbf{X}_1, \mathbf{X}_1)\} = o(nV_{2n}^2)$, then the first part of (S1.12) holds. If $E\{K_2^4(\mathbf{X}_1, \mathbf{X}_2)\} = o(V_{2n}^4)$ holds, then $E\{K_2^2(\mathbf{X}_1, \mathbf{X}_2)\} = o(V_{2n}^2)$ holds.

We then show the second part of (S1.12). By a straightforward computation, we have

$$E(\xi_i^4) = \sum_{j=1}^{i-1} E\Big[\mathcal{K}_2^4(\mathbf{X}_i, \mathbf{X}_j)\{\epsilon_i\epsilon_j - n^{-1}(\epsilon_i + \epsilon_j)^2\}^4\Big]$$
$$+ 3\sum_{l \neq j}^{i-1} E\Big[\mathcal{K}_2^2(\mathbf{X}_i, \mathbf{X}_j)\mathcal{K}_2^2(\mathbf{X}_i, \mathbf{X}_l)\{\epsilon_i\epsilon_j - n^{-1}(\epsilon_i + \epsilon_j)^2\}^2\{\epsilon_i\epsilon_l - n^{-1}(\epsilon_i + \epsilon_l)^2\}^2\Big]$$

$$\leq C\Big[(i-1)E\{\mathcal{K}_2^4(\mathbf{X}_1,\mathbf{X}_2)\} + (i-1)(i-2)E\{\mathcal{K}_2^2(\mathbf{X}_1,\mathbf{X}_2)\mathcal{K}_2^2(\mathbf{X}_1,\mathbf{X}_3)\}\Big],$$

where C is a constant. Therefore, the second part of (S1.12) holds if $E\{K_2^2(\mathbf{X}_1,\mathbf{X}_2)K_2^2(\mathbf{X}_1,\mathbf{X}_3)\} = o(nV_{2n}^4)$ and $E\{K_2^4(\mathbf{X}_1,\mathbf{X}_2)\} = o(n^2V_{2n}^4)$. Both conditions hold if $E\{K_2^4(\mathbf{X}_1,\mathbf{X}_2)\} = o(V_{2n}^4)$. □

Remark 3: Derivation of the adjusted variance $\sigma^2_{T_n,adj}$. Consider

$$n(n-1)T_{n1} = \frac{2}{\sigma^2}\mathbf{Y}^T\mathbf{H}\mathbf{K}^0\mathbf{H}\mathbf{Y} - \frac{1}{\sigma^4(n-1)}\mathbf{Y}^T\mathbf{H}\mathbf{K}^0\mathbf{H}\mathbf{Y}\mathbf{Y}^T\mathbf{H}\mathbf{Y} \triangleq G_1 - G_2,$$

By using results from Zhong and Chen (2011), we have $\mathrm{E}(G_1) = 2\mathrm{tr}(\mathbf{H}\mathbf{K}^0)$,

$$\begin{aligned}
\mathrm{E}(G_2) &= \frac{1}{(n-1)}[\mathrm{tr}(\mathbf{H}\mathbf{K}^0)\mathrm{tr}(\mathbf{H}) + 2\mathrm{tr}(\mathbf{H}\mathbf{K}^0) + \Delta\mathrm{tr}(\mathbf{H}\mathbf{K}^0\mathbf{H}\circ\mathbf{H})],\\
\mathrm{Var}(G_1) &= 8\mathrm{tr}(\mathbf{H}\mathbf{K}^0\mathbf{H}\mathbf{K}^0) + 4\Delta\mathrm{tr}(\mathbf{A}\circ\mathbf{A}),\\
\mathrm{Cov}(G_1,G_2) &= \frac{1}{(n-1)}[16\mathrm{tr}(\mathbf{H}\mathbf{K}^0\mathbf{H}\mathbf{K}^0) + 4\mathrm{tr}^2(\mathbf{H}\mathbf{K}^0) + 4\mathrm{tr}(\mathbf{H}\mathbf{K}^0\mathbf{H}\mathbf{K}^0)\mathrm{tr}(\mathbf{H})]\\
&\quad + \frac{2\Delta}{(n-1)}[\mathrm{tr}(\mathbf{H}\mathbf{K}^0)\mathrm{tr}(\mathbf{A}\circ\mathbf{H}) + \mathrm{tr}(\mathbf{H})\mathrm{tr}(\mathbf{A}\circ\mathbf{A}) + \mathrm{tr}(\mathbf{A}^2\circ\mathbf{H}) + 2\mathrm{tr}(\mathbf{A}\circ\mathbf{A})]\\
&\quad + \frac{2(\tau_6 - 15 - 6\Delta)}{(n-1)}\mathrm{tr}(\mathbf{A}\circ\mathbf{A}\circ\mathbf{H})
\end{aligned}$$

where $\tau_k = E\left\{(Y-\mu)/\sigma\right\}^k$ for any $k \in \mathcal{N}$. Applying the results from Bao and Ullah (2010),

$$\mathrm{E}(\mathbf{Y}^T\mathbf{A}\mathbf{Y}\mathbf{Y}^T\mathbf{A}\mathbf{Y}\mathbf{Y}^T\mathbf{H}\mathbf{Y}\mathbf{Y}^T\mathbf{H}\mathbf{Y})/\sigma^8$$

$$= \mathrm{tr}^2(\mathbf{HK}^0)\mathrm{tr}^2(\mathbf{H}) + 10\mathrm{tr}^2(\mathbf{HK}^0)\mathrm{tr}(\mathbf{H}) + 2\mathrm{tr}^2(\mathbf{H})\mathrm{tr}(\mathbf{HK}^0\mathbf{HK}^0) + 20\mathrm{tr}(\mathbf{HK}^0\mathbf{HK}^0)\mathrm{tr}(\mathbf{H})$$
$$+ 24\mathrm{tr}^2(\mathbf{HK}^0) + 48\mathrm{tr}(\mathbf{HK}^0\mathbf{HK}^0) + R_n$$

where $R_n = \gamma_2 f_{\gamma_2} + \gamma_4 f_{\gamma_4} + \gamma_6 f_{\gamma_6} + \gamma_2^2 f_{\gamma_2^2}$, $\gamma_2 = \tau_4 - 3 = \Delta$, $\gamma_4 = \tau_6 - 15\Delta - 15$, $\gamma_6 = \tau_8 - 28\gamma_4 - 35\Delta^2 - 210\Delta - 105$, and

$$f_{\gamma_2} = n^{-1}\mathrm{tr}^2(\mathbf{HK}^0)\big\{5(n-1)^2 + 8n + 24(n-1)\big\} + n^{-1}\mathrm{tr}(\mathbf{HK}^0\mathbf{HK}^0)\big\{10(n-1)^2 + 64(n-1)\big\}$$
$$+ \mathrm{tr}(\mathbf{A}\circ\mathbf{A})\big\{(n-1)^2 + 2n^{-1}(n-1)^2 + 16(n-1) + 48 + 16n^{-1}(n-1)\big\},$$
$$f_{\gamma_4} = 2\mathrm{tr}^2(\mathbf{HK}^0)n^{-2}(n-1)^2 + 4\mathrm{tr}(\mathbf{HK}^0\mathbf{HK}^0)n^{-2}(n-1)^2 + \mathrm{tr}(\mathbf{A}\circ\mathbf{A})n^{-1}(n-1)(2n+18),$$
$$f_{\gamma_6} = \mathrm{tr}(\mathbf{A}\circ\mathbf{A})n^{-2}(n-1)^2,$$
$$f_{\gamma_2^2} = \mathrm{tr}^2(\mathbf{HK}^0)n^{-2}\big(2n^2 - 2n + 4\big) + \mathrm{tr}(\mathbf{A}\circ\mathbf{A})n^{-1}\big\{8n - 16 + (n-1)^2\big\}$$
$$+ \mathrm{tr}(\mathbf{HK}^0\mathbf{HK}^0)n^{-2}\big(24n^2 - 32n + 12\big).$$

Hence,

$$(n-1)^2\mathrm{Var}(G_2) = \mathrm{E}(\mathbf{Y}^T\mathbf{AYY}^T\mathbf{AYY}^T\mathbf{HYY}^T\mathbf{HY})/\sigma^8 - \{\mathrm{E}(G_2)\}^2$$
$$= \mathrm{tr}^2(\mathbf{HK}^0)\big\{6(n-1) + 20\big\} + \mathrm{tr}(\mathbf{HK}^0\mathbf{HK}^0)\big\{2(n-1)^2 + 20(n-1) + 48\big\}$$
$$-\big\{4\Delta\mathrm{tr}(\mathbf{HK}^0)\mathrm{tr}(\mathbf{A}\circ\mathbf{H}) + \Delta^2\mathrm{tr}^2(\mathbf{A}\circ\mathbf{H}) + 2\Delta\mathrm{tr}(\mathbf{HK}^0)\mathrm{tr}(\mathbf{H})\mathrm{tr}(\mathbf{A}\circ\mathbf{H})\big\} + R_n.$$

Denote $S_1 = -2\mathrm{tr}^2(\mathbf{HK}^0)/(n-1) + \mathrm{tr}(\mathbf{HK}^0\mathbf{HK}^0)\big\{2 - 12/(n-1)\big\}$ and $S_2 = -\mathrm{tr}^2(\mathbf{HK}^0)\Delta/n +$

$6\mathrm{tr}(\mathbf{H}\mathbf{K}^0\mathbf{H}\mathbf{K}^0)\Delta/n + \Delta\mathrm{tr}(\mathbf{A}\circ\mathbf{A})$, then we have

$$
\begin{aligned}
\mathrm{Var}(G_1 - G_2) &= S_1 + S_2 + o(S_1 + S_2) \\
&= \mathrm{tr}(\mathbf{H}\mathbf{K}^0\mathbf{H}\mathbf{K}^0)\{2 - 12(n-1)^{-1} + 6\Delta n^{-1}\} \\
&+ \mathrm{tr}^2(\mathbf{H}\mathbf{K}^0)\{-2(n-1)^{-1} - \Delta n^{-1}\} + \Delta\mathrm{tr}(\mathbf{A}\circ\mathbf{A}) + o(S_1 + S_2),
\end{aligned}
$$

where it can be proved $\mathrm{tr}(\mathbf{A}\circ\mathbf{A}) = n^{-1}\big\{2\mathrm{tr}(\mathbf{K}^2) - \mathrm{tr}^2(\mathbf{H}\mathbf{K}^0) - 2\mathrm{tr}(\mathbf{H}\mathbf{K}^0\mathbf{H}\mathbf{K}^0)\big\}\{1+o(1)\}$. Therefore, $\mathrm{Var}(nT_{n1}) = (n-1)^{-2}\mathrm{Var}(G_1 - G_2)$, which is an adjustment for the variance of test statistic nT_n, since $\mathrm{Var}(nT_n) = \mathrm{Var}(nT_{n1})\{1 + o(1)\}$. This finishes the proof. □

Proof of Remark 4: Let $T_n = \sum_{i\neq j} K_\theta(\mathbf{X}_i, \mathbf{X}_j)(Y_i - \bar{Y}_n)(Y_j - \bar{Y}_n)/\{n(n-1)\hat{\sigma}^2\}$ and $T_{n1} = \sum_{i\neq j} K_\theta(\mathbf{X}_i, \mathbf{X}_j)(Y_i - \bar{Y}_n)(Y_j - \bar{Y}_n)/\{n(n-1)\sigma^2\}$ be the statistics using the true centralized kernel K_θ. Let $\hat{T}_n = \sum_{i\neq j} K_{n,\theta}(\mathbf{X}_i, \mathbf{X}_j)(Y_i - \bar{Y}_n)(Y_j - \bar{Y}_n)/\{n(n-1)\hat{\sigma}^2\}$ and $\hat{T}_{n1} = \sum_{i\neq j} K_{n,\theta}(\mathbf{X}_i, \mathbf{X}_j)(Y_i - \bar{Y}_n)(Y_j - \bar{Y}_n)/\{n(n-1)\sigma^2\}$ be the ones using empirically centralized kernel $K_{n,\theta}$. Following similar arguments in the proof of Theorem 1, $nT_n/\sqrt{V_{2n}} = nT_{n1}/\sqrt{V_{2n}}\{1 + o_p(1)\}$ and $n\hat{T}_n/\sqrt{V_{2n}} = n\hat{T}_{n1}/\sqrt{V_{2n}}\{1 + o_p(1)\}$. To show $nT_n/\sqrt{V_{2n}} = n\hat{T}_n/\sqrt{V_{2n}}\{1 + o_p(1)\}$, it remains to show $(nT_{n1} - n\hat{T}_{n1})/\sqrt{V_{2n}} = o_p(1)$.

In fact, $\Delta_{n,D} = V_{2n}^{-1/2}(nT_{n1} - n\hat{T}_{n1}) = \sum_{i\neq j} \mathbf{D}_{ij}(Y_i - \bar{Y})(Y_j - \bar{Y})/\{(n-1)\sigma^2\}$, where $\mathbf{D} = V_{2n}^{-1/2}(\mathbf{K} - \mathbf{K}_n)$, $D_{ij} = V_{2n}^{-1/2}\{K_{1,\theta}^*(\mathbf{X}_j) - (n-1)^{-1}\sum_{k\neq j} K_{kj}^* + K_{1,\theta}^*(\mathbf{X}_i) - (n-1)^{-1}\sum_{k\neq i} K_{ki}^* + n^{-1}(n-1)^{-1}\sum_{k\neq l} K_{kl}^* - \mu_{K^*}\}$ and $K_{ij}^* = K^*(\mathbf{X}_i, \mathbf{X}_j)$. Viewing $\Delta_{n,D}$ as a special case that was considered in proof of Theorem 1, it is not difficult to see that $\mathrm{E}(\Delta_{n,D}) = 0$, and the asymptotic variance of $\Delta_{n,D}$ is $2\mathrm{E}(D_{ij}^2) \leq CV_{2n}^{-1}(2\sigma_{\Delta,1}^2 + \sigma_{\Delta,2}^2)$ for some constant C, where $\sigma^2_{\Delta,1} = \mathrm{E}\{K^*_{1,\theta}(\mathbf{X}_j) - (n-1)^{-1}\sum_{k\neq j} K^*_{kj}\}^2$ and $\sigma^2_{\Delta,2} = \mathrm{E}\{n^{-1}(n-1)^{-1}\sum_{k\neq l} K^*_{kl} - \mu_{K^*}\}^2$. In the following we will show that $V_{2n}^{-1}(2\sigma^2_{\Delta,1} + \sigma^2_{\Delta,2}) = o(1)$.

Let $\{\lambda^*_{nm}, \psi^*_{nm}\}_{m=1}^{\infty}$ be the eigen-decomposition of kernel K^*. Denote $V^*_{2n} = \sum_{m=1}^{\infty} \lambda^{*2}_{nm}$, $\kappa_m = \mathrm{E}\{\psi^*_{nm}(X)\}$, $\nu_1 = \sum_{m=1} \lambda^*_{nm}\kappa^2_m$ and $\nu_2 = \sum_{m=1} \lambda^{*2}_{nm}\kappa^2_m$. Since the $\Delta_{D,n}$ is invariant when the kernel is scaled, we can assume $\lambda^*_1 = 1$ without loss of generality. Then it can be shown that $\sigma^2_{\Delta,1} = n^{-1}(V^*_{2n} - \nu_2)$ and $\sigma^2_{\Delta,2} = 4n^{-1}\nu_2 + n^{-2}V^*_{2n}$. Moreover, it has been studied in Lindsay *et al.* (2014) that $V_{2n} = V^*_{2n} - 2\nu_2 + \nu_1^2$, where $1 \leq \nu_2 \leq \nu_1 \leq \sqrt{V^*_2}$, Therefore,

$$V_{2n}^{-1}(2\sigma^2_{\Delta,1} + \sigma^2_{\Delta,2}) = \frac{2}{n}\frac{V^*_{2n} + \nu_2}{V_{2n}}\{1 + o(1)\} \leq \frac{2}{n}\frac{V^*_{2n} + \sqrt{V^*_{2n}}}{V^*_{2n} - 2\nu_2 + \nu_1^2} \leq \frac{2}{n}\frac{V^*_{2n} + \sqrt{V^*_{2n}}}{V^*_{2n} - 2\sqrt{V^*_{2n}}},$$

which is $o(1)$ no matter V^*_{2n} is infinite or finite. □

Polynomial Kernel: Consider the polynomial kernel $K^*_{n,\theta}(\mathbf{Z}_1, \mathbf{Z}_2) = (\mathbf{Z}_1^T \Lambda \mathbf{Z}_2)^d$ where $\mathbf{Z}_1$ and $\mathbf{Z}_2$ are independent multivariate distributed normal random vectors with mean $\boldsymbol{\mu}^*$ and variance $\mathbf{I}_p$. Then the corresponding centralized kernel is

$$K_{n,\theta}(\mathbf{Z}_1, \mathbf{Z}_2) = K^*_{n,\theta}(\mathbf{Z}_1, \mathbf{Z}_2) - E\{K^*_{n,\theta}(\mathbf{Z}_1, \mathbf{Z}_2)|\mathbf{Z}_1\} - E\{K^*_{n,\theta}(\mathbf{Z}_1, \mathbf{Z}_2)|\mathbf{Z}_2\} + E\{K^*_{n,\theta}(\mathbf{Z}_1, \mathbf{Z}_2)\}.$$

Let $J = \{j_1, \cdots, j_p\}$ be a set of non-negative integers such that $j_1 + \cdots + j_p = d$ and $\{k_1, \cdots, k_{S_J}\}$ be a subset of $\{1, \cdots, p\}$ for which $j_{k_1} \neq 0, \cdots, j_{k_{S_J}} \neq 0$. It follows that

$$E\{K^*_{n,\theta}(\mathbf{Z}_1, \mathbf{Z}_2)|\mathbf{Z}_1\} = \sum_{j_1 + j_2 + \cdots + j_p = d} \frac{d!}{j_1! \cdots j_p!} E\Big\{\prod_{k=1}^{p} \eta_k^{j_k} Z_{1k}^{j_k} Z_{2k}^{j_k} |\mathbf{Z}_1\Big\}$$

$$= \sum_{j_1+j_2+\cdots+j_p=d} \frac{d!}{j_1!\cdots j_p!} E \prod_{l=1}^{S_J} \eta_{k_l}^{j_{k_l}} Z_{1k_l}^{j_{k_l}} E(Z_{2k_l}^{j_{k_l}}).$$

We then write the centralized kernel as

$$K_{n,\theta}(\mathbf{Z}_1, \mathbf{Z}_2) = \sum_{j_1+j_2+\cdots+j_p=d} \frac{d!}{j_1!\cdots j_p!} \prod_{l=1}^{S_J} \eta_{k_l}^{j_{k_l}} \{Z_{1k_l}^{j_{k_l}} - E(Z_{1k_l}^{j_{k_l}})\}\{Z_{2k_l}^{j_{k_l}} - E(Z_{2k_l}^{j_{k_l}})\}.$$

Following the derivation in Liang and Lee (2013), the eigenvalues λ_n and eigenfunctions $\phi(\cdot)$ of K satisfy the following equation

$$\int \sum_{j_1+j_2+\cdots+j_p=d} \frac{d!}{j_1!\cdots j_p!} \prod_{l=1}^{S_J} \eta_{k_l}^{j_{k_l}} \{Z_{k_l}^{j_{k_l}} - E(Z_{k_l}^{j_{k_l}})\}\{W_{k_l}^{j_{k_l}} - E(W_{k_l}^{j_{k_l}})\}\phi(\mathbf{Z})f(\mathbf{Z})d\mathbf{Z} = \lambda_n \phi(\mathbf{W}),$$

where $f(\mathbf{Z})$ is the density function of $\mathbf{Z}$. The above equation then can be written as

$$\sum_{j_1+j_2+\cdots+j_p=d} \left(\frac{d!}{j_1!\cdots j_p!}\right)^{1/2} \prod_{l=1}^{S_J} \eta_{k_l}^{j_{k_l}/2} \{W_{k_l}^{j_{k_l}} - E(W_{k_l}^{j_{k_l}})\} C_{j_1,\cdots,j_p} = \lambda_n \phi(\mathbf{W}), \qquad \text{(S1.13)}$$

where

$$C_{j_1,\cdots,j_p} = \left(\frac{d!}{j_1!\cdots j_p!}\right)^{1/2} \prod_{l=1}^{S_J} \int \eta_{k_l}^{j_{k_l}/2} \{Z_{k_l}^{j_{k_l}} - E(Z_{k_l}^{j_{k_l}})\}\phi(\mathbf{Z})f(\mathbf{Z})d\mathbf{Z}.$$

From equation (S1.13), we obtain

$$\phi(\mathbf{W}) = \lambda_n^{-1} \sum_{j_1+j_2+\cdots+j_p=d} \left(\frac{d!}{j_1!\cdots j_p!}\right)^{1/2} \prod_{l=1}^{S_J} \eta_{k_l}^{j_{k_l}/2} \{W_{k_l}^{j_{k_l}} - E(W_{k_l}^{j_{k_l}})\} C_{j_1,\cdots,j_p}. \qquad \text{(S1.14)}$$

Plugging the equation $\phi(\cdot)$ into the expression of $C_{j_1,\cdots,j_p}$, we have

$$C_{j_1,\cdots,j_p} = \Big(\frac{d!}{j_1!\cdots j_p!}\Big)^{1/2}\prod_{l=1}^{S_J}\int \eta_{k_l}^{j_{k_l}/2}\{Z_{k_l}^{j_{k_l}} - E(Z_{k_l}^{j_{k_l}})\}$$
$$\times\Big[\lambda_n^{-1}\sum_{i_1+i_2+\cdots+i_p=d}\Big(\frac{d!}{i_1!\cdots i_p!}\Big)^{1/2}\prod_{l=1}^{S_I}\eta_{k_l}^{i_{k_l}/2}\{Z_{k_l}^{i_{k_l}} - E(Z_{k_l}^{i_{k_l}})\}C_{i_1,\cdots,i_p}\Big]f(\mathbf{Z})d\mathbf{Z},$$

where $I = \{k_1, \cdots, k_{S_I}\} \subset \{1, \cdots, p\}$ for which $j_{k_1} \neq 0, \cdots, j_{k_{S_I}} \neq 0$. Because of the independence among $\{Z_1, \cdots, Z_p\}$, the integration in the above equation is zero whenever I is not the same as J. Therefore, we can write the above equation as

$$\lambda_n C_{j_1,\cdots,j_p} = \sum_{i_{k_1}+i_{k_2}+\cdots+i_{k_{S_J}}=d}\Big(\frac{d!}{j_1!\cdots j_p!}\Big)^{1/2}\Big(\frac{d!}{i_{k_1}!\cdots i_{k_{S_J}}!}\Big)^{1/2}\mu_J C_{i_1,\cdots,i_p},$$

where $\mu_J = \prod_{l=1}^{S_J}\eta_{k_l}^{(i_{k_l}+j_{k_l})/2}E\Big[\{Z_{k_l}^{j_{k_l}} - E(Z_{k_l}^{j_{k_l}})\}\{Z_{k_l}^{i_{k_l}} - E(Z_{k_l}^{i_{k_l}})\}\Big]$. Here $\{k_1, \cdots, k_{S_J}\}$ is a subset of $\{1, \cdots, p\}$ for which $j_{k_1} \neq 0, \cdots, j_{k_{S_J}} \neq 0$ and $i_1, \cdots, i_p$ has the same non-zero support as $\{j_1, \cdots, j_p\}$ on J.

Let $M_{p,d} = (M_{p,d}^{(J)})$ be a $d_p \times d_p$ matrix with the entries $M_{p,d}^{(J)}$ given by

$$M_{p,d}^{(J)} = \Big(\frac{d!}{j_1!\cdots j_p!}\Big)^{1/2}\Big(\frac{d!}{i_{k_1}!\cdots i_{k_{S_J}}!}\Big)^{1/2}\mu_J,$$

where $d_p = \binom{d+p-1}{d}$. Then, the eigenvalues of the kernel K^* are given by the following equation:

$$M_{p,d}C = \lambda_n C,$$

where $C = (C_J)$ is a d_p-dimensional eigenvector that containing all $C_{j_1,\cdots,j_p}$ for which $j_1 + \cdots + j_p = d$.

Based on the above results, the eigenvalues of the kernel K^* are given by the eigenvalues of the matrix $M_{p,d}$. Therefore, V_{2n} is a summation of the squares of $M_{p,d}^{(J)}$. If $\mathrm{Cov}(Z_k^u, Z_k^v)$ for any integers $u, v \in \{1, \cdots, d\}$ are uniformly bounded (for $k = 1, \cdots, p$) above and below, then applying the Cauchy-schwarz inequality, we have

$$V_{2n} \asymp \sum_{i_{k_1}+\cdots+i_{k_{S_J}}=d} \sum_{j_{k_1}+\cdots+j_{k_{S_J}}=d} \frac{d!}{i_{k_1}!\cdots i_{k_{S_J}}!} \frac{d!}{j_{k_1}!\cdots j_{k_{S_J}}!} \prod_{l=1}^{k_{S_J}} \eta_k^{(i_{k_l}+j_{k_l})}$$
$$\asymp \sum_{i_{k_1}+\cdots+i_{k_{S_J}}=d} \frac{d!}{i_{k_1}!\cdots i_{k_{S_J}}!} \prod_{l=1}^{k_{S_J}} \eta_k^{2i_{k_l}} = \sum_{i_1+\cdots+i_p=d} \frac{d!}{i_1!\cdots i_p!} \prod_{k=1}^{p} \eta_k^{2i_k} = \mathrm{tr}^d(\mathbf{\Sigma}^2).$$

This completes the proof of the polynomial example. □

Proof of Theorem 2: Under the alternative hypothesis H_{1n}, $Y_i = \mu + d_n(\mathbf{X}_i) + \epsilon_i$, where $\mathrm{E}d_n(\mathbf{X}_i) = 0$, $\mathrm{E}(\epsilon_i) = 0$, and $\mathrm{Var}(\epsilon_i) = \sigma^2$. Without loss of generality, we also assume $\mu = 0$ in the following proof. Similar to (S1.6), considering the following expansion

$$T_n = \frac{1}{n(n-1)\sigma^2} \sum_{i \neq j} K(\mathbf{X}_i, \mathbf{X}_j)(Y_i - \bar{Y}_n)(Y_j - \bar{Y}_n)\{1 + (\frac{\sigma^2}{\hat{\sigma}^2} - 1)\} := T_{n1}\{1 + (\frac{\sigma^2}{\hat{\sigma}^2} - 1)\}.$$

Under condition (3.9), $\mathrm{E}(Y_i) = 0$ and $\mathrm{E}(Y_i^2) = \sigma^2 + \mathrm{E}\{d_n^2(\mathbf{X}_i)\} = \sigma^2\{1 + o(1)\}$, it is not difficult to see $\hat{\sigma}^2 \xrightarrow{p} \sigma^2$. Hence it is enough to study the behavior of T_{n1}. By plugging in expression of Y_i and Y_j under H_{1n}, T_{n1} could be decomposed into two parts: T_{n1}^0 and $\tilde{T}_{n1}^0$, where asymptotic distribution of T_{n1}^0 is the null distribution and has been studied in

Theorem 1. The remainder term $\tilde{T}_{n1}^0$ can be expressed as the sum of the following three terms

$$\tilde{T}_{n1}^0 = \{\Theta_n^{(2)} + \Theta_n^{(3)} + \Theta_n^{(4)}\}\{1 + o_p(1)\}, \tag{S1.15}$$

where

$$\Theta_n^{(2)} = \frac{\sqrt{V_{2n}}}{n^2\sigma^2}\sum_{i\neq j}\mathcal{K}_\theta(\mathbf{X}_i,\mathbf{X}_j)\big[U_iU_j + U_i\epsilon_j + U_j\epsilon_i - \frac{1}{n}(U_i+U_j)^2 - \frac{2}{n}(U_i+U_j)(\epsilon_i+\epsilon_j)\big],$$

$$\Theta_n^{(3)} = \frac{\sqrt{V_{2n}}}{n^3\sigma^2}\sum_{i\neq j\neq k}\mathcal{K}_\theta(\mathbf{X}_i,\mathbf{X}_j)\big[-(U_i+U_j)(U_k+\epsilon_k) - (\epsilon_i+\epsilon_j)U_k + \frac{1}{n}(U_k^2 + 2U_k\epsilon_k)\big],$$

$$\Theta_n^{(4)} = \frac{\sqrt{V_{2n}}}{n^4\sigma^2}\sum_{i\neq j\neq k\neq l}\mathcal{K}_\theta(\mathbf{X}_i,\mathbf{X}_j)(U_kU_l + U_k\epsilon_l + U_l\epsilon_k),$$

and $U_i = d_n(\mathbf{X}_i)$. Denote the eigenvalues of normalized kernel $\mathcal{K}_\theta(x,y)$ as $\{\lambda_{\mathcal{K},m}\}_{m=1}^\infty$, where $\sum_m \lambda_{\mathcal{K},m}^2 = 1$ and $\lambda_{\mathcal{K},m} \geq 0$ for each m. Notice that kernel K_θ and $\mathcal{K}_\theta$ have the same eigenfucntions $\{\psi_{nm}(\mathbf{X})\}_{m=1}^\infty$. Besides, for centralized kernel, $\mu_m := \mathrm{E}\{\psi_{nm}(\mathbf{X})\} = 0$ for $m \in \mathcal{N}\backslash\{m^*\}$, where $\mu_{m^*} = 1$ corresponds to zero eigenvalue ($\lambda_{m^*} = 0$) (See Lemma 3 in the first section). Let $\mathbb{G} = \{m : \lambda_m > 0\}$, then $K_\theta(x_1,x_2) = \sum_{m\in\mathbb{G}}\lambda_m\psi_m(x_1)\psi_m(x_2)$ and $\mu_m = 0$ for all $m \in \mathbb{G}$. Define $b_{nm} := \mathrm{E}\big\{\psi_{nm}(\mathbf{X})d_n(\mathbf{X})\big\}$ representing the projection of function $d_n(\mathbf{X})$ onto the eigen-function $\psi_{nm}(\mathbf{X})$.

In the following we will show that (a) $\mathrm{E}(\sigma_{T_n}^{-1}n\tilde{T}_{n1}^0) - \Psi(d_n) = o(1)$, and (b) $\mathrm{Var}(n\tilde{T}_{n1}^0) = o\{\mathrm{Var}(nT_{n1}^0)\}$. To prove (a) and (b), let us study the asymptotic behavior of each term in $n\sigma_{T_n}^{-1}\tilde{T}_{n1}^0 = n\sigma_{T_n}^{-1}(\Theta_n^{(2)} + \Theta_n^{(3)} + \Theta_n^{(4)})$. Firstly split

$$\sigma_{T_n}^{-1} n\Theta_n^{(2)} = n\big(\tilde{S}_{11} + 2\tilde{S}_{12} + 2\tilde{S}_{13} + 2\tilde{S}_{14}\big)\{1 + o_p(1)\}, \tag{S1.16}$$

where

$$\tilde{S}_{11} := \frac{\sqrt{V_{2n}}}{n^2\sigma^2\sigma_{T_n}} \sum_{i\neq j} \mathcal{K}_\theta(\mathbf{X}_i, \mathbf{X}_j) U_i U_j, \quad \tilde{S}_{12} = \frac{\sqrt{V_{2n}}}{n^2\sigma^2\sigma_{T_n}} \sum_{i\neq j} \mathcal{K}_\theta(\mathbf{X}_i, \mathbf{X}_j) U_i \epsilon_j,$$

$$\tilde{S}_{13} = -\frac{\sqrt{V_{2n}}}{n^3\sigma^2\sigma_{T_n}} \sum_{i\neq j} \mathcal{K}_\theta(\mathbf{X}_i, \mathbf{X}_j) U_i^2, \quad \tilde{S}_{14} = -\frac{\sqrt{V_{2n}}}{2n^3\sigma^2\sigma_{T_n}} \sum_{i\neq j} \mathcal{K}_\theta(\mathbf{X}_i, \mathbf{X}_j) U_i \epsilon_i.$$

We want to show $n\tilde{S}_{11} \xrightarrow{p} \Psi(d_n)$ and $n\tilde{S}_{1j} \xrightarrow{p} 0$ for $j = 2, 3, 4$. Actually, $\mathrm{E}(n\tilde{S}_{11}) = n\sigma_{T_n}^{-1}\sqrt{V_{2n}}\sum_{m=1}^{\infty} b_{nm}^2 \lambda_{\mathcal{K},m} = \Psi(d_n)\{1 + o(1)\}$, and

$$n^2\tilde{S}_{11}^2 = \frac{V_{2n}}{n^2\sigma^4\sigma_{T_n}^2} \sum_{\substack{i\neq j, k\neq l \\ m_1, m_2 \in \mathbb{G}}} \lambda_{\mathcal{K},m_1}\lambda_{\mathcal{K},m_2}\psi_{nm_1}(\mathbf{X}_i)\psi_{nm_1}(\mathbf{X}_j)\psi_{nm_2}(\mathbf{X}_k)\psi_{nm_2}(\mathbf{X}_l) U_i U_j U_k U_l.$$

Define index subsets $I_c = \{(i,j,k,l) || \{i,j\} \cap \{k,l\}| = c, i,j,k,l \in \{1, \cdots, n\}, i \neq j, k \neq l\}$ for $c = 0, 1, 2$, where $|\cdot|$ denotes the set cardinality. For example, I_0 represents set $\{(i,j,k,l) \in |i,j,k,l \in \{1, \cdots, n\}, i \neq j \neq k \neq l\}$. Then $n^2\tilde{S}_{11}^2 = J_0 + J_1 + J_2$, where

$$J_c = \frac{V_{2n}}{n^2\sigma^4\sigma_{T_n}^2} \sum_{\substack{i,j,k,l \in I_c \\ m_1, m_2 \in \mathbb{G}}} \lambda_{\mathcal{K},m_1}\lambda_{\mathcal{K},m_2}\psi_{nm_1}(\mathbf{X}_i)\psi_{nm_1}(\mathbf{X}_j)\psi_{nm_2}(\mathbf{X}_k)\psi_{nm_2}(\mathbf{X}_l) U_i U_j U_k U_l.$$

By using the orthogonal and centralized properties of eigen-functions, it can be proved that

$$
\begin{aligned}
\mathrm{E}(J_0) &= \mathrm{E}^2(n\tilde{S}_{11})\{1+o(1)\} = \Psi^2(d_n)\{1+o(1)\},\\
\mathrm{E}(J_1) &= \frac{4V_{2n}}{n^2\sigma^4\sigma_{T_n}^2}\sum_{\substack{i\neq j\neq k\\ m_1,m_2\in\mathbb{G}}}\lambda_{\mathcal{K},m_1}\lambda_{\mathcal{K},m_2}\mathrm{E}\{\psi_{nm_1}(\mathbf{X}_i)\psi_{nm_2}(\mathbf{X}_i)U_i^2\}\mathrm{E}\{\psi_{nm_1}(\mathbf{X}_j)U_j\}\mathrm{E}\{\psi_{nm_2}(\mathbf{X}_k)U_k\}\\
&= 4nV_{2n}\sigma^{-4}\sigma_{T_n}^{-2}\Bigg(\sum_{m_1,m_2\in\mathbb{G}}\lambda_{\mathcal{K},m_1}\lambda_{\mathcal{K},m_2}b_{nm_1}b_{nm_2}e_{m_1,m_2}\Bigg)\{1+o(1)\},\\
\mathrm{E}(J_2) &= 2nV_{2n}\sigma^{-4}\sigma_{T_n}^{-2}\Bigg(\sum_{m_1,m_2\in\mathbb{G}}\lambda_{\mathcal{K},m_1}\lambda_{\mathcal{K},m_2}e_{m_1,m_2}^2\Bigg)\{1+o(1)\},
\end{aligned}
$$

where $e_{m_1,m_2} = \mathrm{E}[\psi_{nm_1}(\mathbf{X})\psi_{nm_2}(\mathbf{X})d_n^2(\mathbf{X})]$. Under condition (3.9), we can prove that $|b_{nm}| \leq D_1\big[\mathrm{E}d_n^8(\mathbf{X})\big]^{1/8}$, and $|e_{m_1,m_2}| \leq D_2\big[\mathrm{E}d_n^8(\mathbf{X})\big]^{1/4}$ for some finite constants D_1 and D_2, by using Cauchy-Schwartz inequality. Therefore, $\mathrm{E}(J_1) \leq 2\sigma^{-4}D_1^2D_2\cdot n\big[\mathrm{E}d_n^8(\mathbf{X})\big]^{1/2} = o(1)$, $\mathrm{E}(J_2) \leq \sigma^{-4}D_2^2\cdot n\big[\mathrm{E}d_n^8(\mathbf{X})\big]^{1/2} = o(1)$, and $\mathrm{Var}(n\tilde{S}_{11}) = o(1)$ under condition (3.9). Hence $n\tilde{S}_{11} \xrightarrow{p} \Psi(d_n) = O(1)$. It remains to prove $n\tilde{S}_{1j} \xrightarrow{p} 0$ for $j = 2,3,4$ in (S1.16).

It is easy to see that $\mathrm{E}(n\tilde{S}_{12}) = \mathrm{E}(n\tilde{S}_{13}) = \mathrm{E}(n\tilde{S}_{14}) = 0$ by using the centralized kernel property. Moreover, it can be proved that $\mathrm{Var}(n\tilde{S}_{12}) = \mathrm{Var}(n\tilde{S}_{13}) = \mathrm{Var}(n\tilde{S}_{14}) = o(1)$. Actually,

$$
\begin{aligned}
\mathrm{Var}(n\tilde{S}_{12}) &= \frac{V_{2n}}{\sigma^2\sigma_{T_n}^2}\sum_{m\in\mathbb{G}}\lambda_{\mathcal{K},m}^2 e_{m,m}\{1+o(1)\} \leq (\sqrt{2}\sigma)^{-2}D_2[\mathrm{E}d_n^8(\mathbf{X})]^{1/4},\\
\mathrm{Var}(n\tilde{S}_{13}) &= \frac{V_{2n}}{n^2\sigma^4\sigma_{T_n}^2}\left\{\sum_{m\in\mathbb{G}}\lambda_{\mathcal{K},m}^2 f_m + \sum_{m_1,m_2}\lambda_{\mathcal{K},m_1}\lambda_{\mathcal{K},m_2}e_{m_1,m_2}^2\right\}\{1+o(1)\}\\
&\leq (\sqrt{2}n\sigma^2)^{-2}[\mathrm{E}d_n^8(\mathbf{X})]^{1/2}(D_3+D_2^2)
\end{aligned}
$$

$$\mathrm{Var}(n\tilde{S}_{14}) = \frac{V_{2n}}{n^2\sigma^2\sigma_{T_n}^2}\sum_{m\in\mathbb{G}}\lambda_{\mathcal{K},m}^2 e_{m,m}\{1+o(1)\} \le (\sqrt{2}n\sigma)^{-2}D_2[\mathrm{E}d_n^8(\mathbf{X})]^{1/4},$$

where $f_m = \mathrm{E}[\psi_m^2(\mathbf{X})d_n^4(\mathbf{X})] \le D_3[\mathrm{E}d_n^8(\mathbf{X})]^{1/2}$ for some constant $D_3 > 0$. The variance above are all of order $o(1)$ under condition (3.10). For the triple sum terms $\Theta_n^{(3)}$ in (S1.15),

$$\tilde{S}_{21} = \frac{\sqrt{V_{2n}}}{n^3\sigma^2\sigma_{T_n}}\sum_{i\neq j\neq k}\mathcal{K}_\theta(\mathbf{X}_i,\mathbf{X}_j)U_i\epsilon_k, \quad \tilde{S}_{22} = \frac{\sqrt{V_{2n}}}{n^3\sigma^2\sigma_{T_n}}\sum_{i\neq j\neq k}\mathcal{K}_\theta(\mathbf{X}_i,\mathbf{X}_j)U_iU_k,$$

$$\tilde{S}_{23} = \frac{\sqrt{V_{2n}}}{n^3\sigma^2\sigma_{T_n}}\sum_{i\neq j\neq k}\mathcal{K}_\theta(\mathbf{X}_i,\mathbf{X}_j)U_k\epsilon_i, \quad \tilde{S}_{24} = \frac{\sqrt{V_{2n}}}{n^4\sigma^2\sigma_{T_n}}\sum_{i\neq j\neq k}\mathcal{K}_\theta(\mathbf{X}_i,\mathbf{X}_j)U_k\epsilon_k,$$

$$\tilde{S}_{25} = \frac{\sqrt{V_{2n}}}{n^4\sigma^2\sigma_{T_n}}\sum_{i\neq j\neq k}\mathcal{K}_\theta(\mathbf{X}_i,\mathbf{X}_j)U_k^2.$$

Similarly, it is not difficult to see that $\mathrm{E}(n\tilde{S}_{2j}) = 0$ for $j = 1, ..., 5$. Furthermore, up to a factor of $\{1 + o(1)\}$, we have the following

$$\begin{aligned}
\mathrm{Var}(n\tilde{S}_{21}) =& \frac{1}{2\sigma^2}\sum_{m\in\mathbb{G}}\lambda_{\mathcal{K},m}^2\big(b_{nm}^2 + 2n^{-1}e_{m,m}\big),\\
\mathrm{Var}(n\tilde{S}_{22}) =& \frac{1}{2\sigma^4}\sum_{m_1,m_2\in\mathbb{G}}\lambda_{\mathcal{K},m_1}\lambda_{\mathcal{K},m_2}\big(b_{nm_1}^2b_{nm_2}^2 + n^{-1}d_{m_1,m_2}^2\mathrm{E}[d_n^2(\mathbf{X})] + 2n^{-1}b_{nm_1}c_{m_2}d_{m_1,m_2}\big)\\
&+ \frac{1}{2\sigma^4}\sum_{m\in\mathbb{G}}\lambda_{\mathcal{K},m}^2\big(b_{nm}^2\mathrm{E}[d_n^2(\mathbf{X})] + n^{-1}e_{m,m}\mathrm{E}[d_n^2(\mathbf{X})] + n^{-1}c_m^2\big),\\
\mathrm{Var}(n\tilde{S}_{23}) =& (2n\sigma^4)^{-1}\big(\sum_{m\in\mathbb{G}}\lambda_{\mathcal{K},m}^2b_{nm}^2 + \mathrm{E}[d_n^2(\mathbf{X})]\big),\\
\mathrm{Var}(n\tilde{S}_{24}) =& (n^3\sigma^4)^{-1}\mathrm{E}[d_n^2(\mathbf{X})],
\end{aligned}$$

$$\mathrm{Var}(n\tilde{S}_{25}) = (2n^2\sigma^4)^{-1}\bigg(\mathrm{E}^2[d_n^2(\mathbf{X})] + 2n^{-1}\mathrm{E}[d_n^4(\mathbf{X})]\bigg) + \frac{2}{n^3\sigma^4}\sum_{m\in\mathbb{G}}\lambda_{\mathcal{K},m}^2 b_{nm}^2,$$

where $c_m = \mathrm{E}[\psi_{nm}(\mathbf{X})d_n^2(\mathbf{X})]$, and $d_{m_1,m_2} = \mathrm{E}[\psi_{nm_1}(\mathbf{X})\psi_{nm_2}(\mathbf{X})d_n(\mathbf{X})]$. Under condition (3.9), all the triple sum terms are of small order. Finally consider the following quadruple sum terms $\Theta_n^{(4)}$ in (S1.15),

$$\tilde{S}_{31} = \frac{V_1}{n^4\sigma^2\sigma_{T_n}}\sum_{i\neq j\neq k\neq l}\mathcal{K}_\theta(\mathbf{X}_i,\mathbf{X}_j)U_kU_l, \quad \tilde{S}_{32} = \frac{V_1}{n^4\sigma^2\sigma_{T_n}}\sum_{i\neq j\neq k\neq l}\mathcal{K}_\theta(\mathbf{X}_i,\mathbf{X}_j)U_k\epsilon_l.$$

It can be shown that $\mathrm{E}(n\tilde{S}_{31}) = \mathrm{E}(n\tilde{S}_{31}) = 0$, and

$$\mathrm{Var}(n\tilde{S}_{31}) = \frac{1}{2n^2\sigma^2}\big(\mathrm{E}[d_n^2(\mathbf{X})]\sum_{m\in\mathbb{G}}\lambda_{\mathcal{K},m}^2 b_{nm}^2 + \big(\sum_{m\in\mathbb{G}}\lambda_{\mathcal{K},m}b_{nm}\big)^2 + \mathrm{E}^2[d_n^2(\mathbf{X})]V_{\mathcal{K},2}\big)\{1+o(1)\},$$

$$\mathrm{Var}(n\tilde{S}_{32}) = \frac{1}{2n^2\sigma^2}\big(\sum_{m\in\mathbb{G}}\lambda_{\mathcal{K},m}^2 b_{nm}^2 + \mathrm{E}[d_n^2(\mathbf{X})]V_{\mathcal{K},2}\big)\{1+o(1)\},$$

are of order $o(1)$ under condition (3.10). Therefore, $n\sigma_{T_n}^{-1}\tilde{T}_{n1}^0 \xrightarrow{d} \Psi(d_n)$ under the local hypothesis H_{1n} (3.9). This finishes the proof of Theorem 2. □

Proof of Proposition 1: (i) We will firstly show that the leading order of $\hat{\sigma}_{T_n}^2$ is $\hat{\sigma}_{T_{n1}}^2 = 2n^{-1}(n-1)^{-1}\mathrm{tr}[(\mathbf{K}^0)^2]$, which can be written as a U-statistic. Denote $\mathbf{\Lambda}_K = \mathrm{diag}(\mathbf{K})$, a diagonal matrix with the same diagonal elements as matrix $\mathbf{K}$. Then on one hand, we can see $\mathbf{K}^0 = \mathbf{K} - \mathbf{\Lambda}_K$ and $\mathrm{tr}[(\mathbf{K}^0)^2] = \mathrm{tr}(\mathbf{K}^2) - \mathrm{tr}(\mathbf{\Lambda}_K^2)$. On the other hand,

$$\mathrm{tr}(\mathbf{HK}^0\mathbf{HK}^0) = \mathrm{tr}(\mathbf{K}^2) - 2\mathrm{tr}(\mathbf{HKH\Lambda}_K) + \mathrm{tr}(\mathbf{H\Lambda}_K\mathbf{H\Lambda}_K) = \{\mathrm{tr}(\mathbf{K}^2) - \mathrm{tr}(\mathbf{\Lambda}_K^2)\}\{1+o_p(1)\}.$$

It remains to show that $\hat{\sigma}^2_{T_{n1}}/\sigma^2_{T_n} \xrightarrow{p} 1$ as $n \to \infty$. As a U-statistic, the variance of $\hat{\sigma}^2_{T_{n1}}$ can be derived following the classical method. In fact, $\mathrm{Var}(\hat{\sigma}^2_{T_{n1}}/2V_{2n}) = O(n^{-1})$, and $\mathrm{E}(\hat{\sigma}^2_{T_{n1}}/2V_{2n}) = \mathrm{E}(\hat{\sigma}^2_{T_{n1}}/\sigma^2_{T_n}) = 1$. This finishes the proof of part (i).

(ii) It is sufficient to prove $\mathrm{E}(\hat{\delta}_K/\delta_K) \to 1$ and $\mathrm{Var}(\hat{\delta}_K/\delta_K) \to 0$ as $n \to \infty$. Let $\sigma_n^2(\boldsymbol{a}) = n\mathrm{Var}(\hat{\delta}_K)$. Then we have

$$\mathrm{E}\left(\frac{\hat{\delta}_K}{\delta_K}\right) = \mathrm{E}\left(\frac{\sqrt{n}\sigma_n^{-1}(\boldsymbol{a})(\hat{\delta}_K - \delta_K)}{\sqrt{n}\sigma_n^{-1}(\boldsymbol{a})\delta_K}\right) + 1,$$

and

$$\mathrm{Var}\left(\frac{\hat{\delta}_K}{\delta_K}\right) = \mathrm{Var}\left(\frac{\sqrt{n}\hat{\delta}_K}{\sqrt{n}\delta_K}\right) = \frac{\sigma_n^2(\boldsymbol{a})}{n\delta_K^2}.$$

Therefore, it remains to show that $\sigma_n^2(\boldsymbol{a})/(n\delta_K^2) = o(1)$. Since the leading order of $\hat{\delta}_K$ is a U-statistic, we can obtain its variance using classical results. That is,

$$\sigma_n^2(\boldsymbol{a}) = n\mathrm{Var}(\hat{\delta}_K) = \left[4\mathrm{Var}\{h(\mathbf{X})g(\mathbf{X})\} + 4\sigma^2\mathrm{Var}\{g(\mathbf{X})\} + 4\sigma^2\delta_K^2 + R_n\right]\{1 + o(1)\},$$

where $R_n = 2n^{-1}\mathrm{Var}\{K(\mathbf{X}_1,\mathbf{X}_2)h(\mathbf{X}_1)h(\mathbf{X}_2)\}+4n^{-1}\sigma^2\mathrm{Var}\{K(\mathbf{X}_1,\mathbf{X}_2)h(\mathbf{X}_1)\}+2n^{-1}\sigma^4V_{2n}$, and $g(x) = \mathbb{E}[K(x,\mathbf{X})h(\mathbf{X})]$. Applying the conditional variance formula, we obtain that $\mathrm{Var}\{h(\mathbf{X}_2)g(\mathbf{X}_2)\} \leq \mathrm{Var}\{K(\mathbf{X}_1,\mathbf{X}_2)h(\mathbf{X}_1)h(\mathbf{X}_2)\}$ and $\mathrm{Var}\{g(\mathbf{X}_2)\} \leq \mathrm{Var}\{K(\mathbf{X}_1,\mathbf{X}_2)h(\mathbf{X}_1)\}$. Then the ratio consistency is proved.

In the following we will show that the kernel $\hat{\mathbb{K}}_\theta \in \mathscr{F}_{\mathbb{K},1}$ with probability 1. This is equivalent to show that $P\big(\cup_{\mathbb{K}\in\mathscr{F}_{\mathbb{K},1}}\{\hat{\mathbb{K}}_\theta=\mathbb{K}\}\big)=1$. Because

$$P\big(\cup_{\mathbb{K}\in\mathscr{F}_{\mathbb{K},1}}\{\hat{\mathbb{K}}_\theta=\mathbb{K}\}\big)\geq\max_{\mathbb{K}\in\mathscr{F}_{\mathbb{K},1}}P(\hat{\mathbb{K}}_\theta=\mathbb{K})\geq P(\hat{\mathbb{K}}_\theta=\widetilde{\mathbb{K}}_\theta),$$

one only need to show that $P(\hat{\mathbb{K}}_\theta=\widetilde{\mathbb{K}}_\theta)=1$. Define $\mathscr{F}^{(1)}=\{\mathbb{K}_\theta\in\mathscr{F}_{\mathbb{K},0}:V_{\max,\mathbb{K}_\theta}=o(n\delta^2_{\mathbb{K}_\theta})\}$, and $\mathscr{F}^{(2)}=\mathscr{F}_{\mathbb{K},0}\setminus\mathscr{F}^{(1)}$. It is not difficult to see that for any $\mathbb{K}_\theta\in\mathscr{F}^{(2)}$, $n\delta^2_{\mathbb{K}_\theta}=O(V_{\max,\mathbb{K}_\theta})$. Let us consider two different cases: $\mathbb{K}_\theta\in\mathscr{F}^{(1)}$, or $\mathbb{K}_\theta\in\mathscr{F}^{(2)}$. Specifically, we will show for each of the two cases, for $j=1,2$,

$$\lim_{n\to\infty}P\Big(\bigcap_{\mathbb{K}_\theta\in\mathscr{F}^{(j)},\mathbb{K}_\theta\in\mathscr{F}_{\mathbb{K},0}}\{\hat{\sigma}^{-1}_{T_n,\tilde{\mathbb{K}}_\theta}\hat{\delta}_{\tilde{\mathbb{K}}_\theta}>\hat{\sigma}^{-1}_{T_n,\mathbb{K}_\theta}\hat{\delta}_{\mathbb{K}_\theta}\}\Big)=1. \tag{S1.17}$$

To show (S1.17), we note that

$$\begin{aligned}P\Big(\bigcap_{\mathbb{K}_\theta\in\mathscr{F}_j,\mathbb{K}_\theta\in\mathscr{F}_{\mathbb{K},0}}\{\hat{\sigma}^{-1}_{T_n,\tilde{\mathbb{K}}_\theta}\hat{\delta}_{\tilde{\mathbb{K}}_\theta}>\hat{\sigma}^{-1}_{T_n,\mathbb{K}_\theta}\hat{\delta}_{\mathbb{K}_\theta}\}\Big)&=1-P\Big(\bigcup_{\mathbb{K}_\theta\in\mathscr{F}_j,\mathbb{K}_\theta\in\mathscr{F}_{\mathbb{K},0}}\{\hat{\sigma}^{-1}_{T_n,\tilde{\mathbb{K}}_\theta}\hat{\delta}_{\tilde{\mathbb{K}}_\theta}\leq\hat{\sigma}^{-1}_{T_n,\mathbb{K}_\theta}\hat{\delta}_{\mathbb{K}_\theta}\}\Big)\\&\geq 1-\sum_{\mathbb{K}_\theta\in\mathscr{F}_j,\mathbb{K}_\theta\in\mathscr{F}_{\mathbb{K},0}}P\Big(\hat{\sigma}^{-1}_{T_n,\tilde{\mathbb{K}}_\theta}\hat{\delta}_{\tilde{\mathbb{K}}_\theta}\leq\hat{\sigma}^{-1}_{T_n,\mathbb{K}_\theta}\hat{\delta}_{\mathbb{K}_\theta}\Big).\end{aligned}$$

Thus, to show (S1.17) is equivalent to show that, for $\mathbb{K}_\theta\in\mathscr{F}_{\mathbb{K},0}$,

$$P\Big(\hat{\sigma}^{-1}_{T_n,\tilde{\mathbb{K}}_\theta}\hat{\delta}_{\tilde{\mathbb{K}}_\theta}\leq\hat{\sigma}^{-1}_{T_n,\mathbb{K}_\theta}\hat{\delta}_{\mathbb{K}_\theta}\Big)=o(\frac{1}{|\mathscr{F}_{\mathbb{K},0}|}), \tag{S1.18}$$

where $|\mathscr{F}_{\mathbb{K},0}|$ is the cardinality of the candidate kernel set $\mathscr{F}_{\mathbb{K},0}$.

Consider the first case where $\mathbb{K}_\theta\in\mathscr{F}^{(1)}$ and $\mathbb{K}_\theta\in\mathscr{F}_{\mathbb{K},0}$. Let $\Delta_{\mathbb{K}_\theta}=\sigma^{-1}_{T_n,\tilde{\mathbb{K}}_\theta}\delta_{\tilde{\mathbb{K}}_\theta}-\sigma^{-1}_{T_n,\mathbb{K}_\theta}\delta_{\mathbb{K}_\theta}$. Because $\tilde{\mathbb{K}}_\theta$ maximizes $\sigma^{-1}_{T_n,\mathbb{K}_\theta}\delta_{\mathbb{K}_\theta}$, we have $\Delta_{\mathbb{K}_\theta}>0$ for any $\mathbb{K}_\theta\neq\tilde{\mathbb{K}}_\theta$. Then, we have the following

$$
\begin{aligned}
&P\Big(\hat{\sigma}^{-1}_{T_n,\tilde{\mathbb{K}}_\theta}\hat{\delta}_{\tilde{\mathbb{K}}_\theta} \le \hat{\sigma}^{-1}_{T_n,\mathbb{K}_\theta}\hat{\delta}_{\mathbb{K}_\theta}\Big)\\
=&P\Big\{\big(\hat{\sigma}^{-1}_{T_n,\tilde{\mathbb{K}}_\theta}\hat{\delta}_{\tilde{\mathbb{K}}_\theta} - \sigma^{-1}_{T_n,\tilde{\mathbb{K}}_\theta}\delta_{\tilde{\mathbb{K}}_\theta}\big) - \big(\hat{\sigma}^{-1}_{T_n,\mathbb{K}_\theta}\hat{\delta}_{\mathbb{K}_\theta} - \sigma^{-1}_{T_n,\mathbb{K}_\theta}\delta_{\mathbb{K}_\theta}\big) \le -\Delta_{\mathbb{K}_\theta}\Big\}\\
\le&P\Big\{|\hat{\sigma}^{-1}_{T_n,\tilde{\mathbb{K}}_\theta}\hat{\delta}_{\tilde{\mathbb{K}}_\theta} - \sigma^{-1}_{T_n,\tilde{\mathbb{K}}_\theta}\delta_{\tilde{\mathbb{K}}_\theta}| + |\hat{\sigma}^{-1}_{T_n,\mathbb{K}_\theta}\hat{\delta}_{\mathbb{K}_\theta} - \sigma^{-1}_{T_n,\mathbb{K}_\theta}\delta_{\mathbb{K}_\theta}| \ge \Delta_{\mathbb{K}_\theta}\Big\}\\
\le&P\Big\{\frac{|\hat{\sigma}^{-1}_{T_n,\tilde{\mathbb{K}}_\theta}\hat{\delta}_{\tilde{\mathbb{K}}_\theta} - \sigma^{-1}_{T_n,\tilde{\mathbb{K}}_\theta}\delta_{\tilde{\mathbb{K}}_\theta}|}{\sigma^{-1}_{T_n,\tilde{\mathbb{K}}_\theta}\delta_{\tilde{\mathbb{K}}_\theta}} \ge \frac{\Delta_{\mathbb{K}_\theta}}{2\sigma^{-1}_{T_n,\tilde{\mathbb{K}}_\theta}\delta_{\tilde{\mathbb{K}}_\theta}}\Big\} + P\Big\{\frac{|\hat{\sigma}^{-1}_{T_n,\mathbb{K}_\theta}\hat{\delta}_{\mathbb{K}_\theta} - \sigma^{-1}_{T_n,\mathbb{K}_\theta}\delta_{\mathbb{K}_\theta}|}{\sigma^{-1}_{T_n,\mathbb{K}_\theta}\delta_{\mathbb{K}_\theta}} \ge \frac{\Delta_{\mathbb{K}_\theta}}{2\sigma^{-1}_{T_n,\mathbb{K}_\theta}\delta_{\mathbb{K}_\theta}}\Big\}\\
\le&4\mathrm{Var}(\hat{\delta}_{\tilde{\mathbb{K}}_\theta}/\delta_{\tilde{\mathbb{K}}_\theta})\frac{\sigma^{-2}_{T_n,\tilde{\mathbb{K}}_\theta}\delta^2_{\tilde{\mathbb{K}}_\theta}}{\Delta^2_{\mathbb{K}_\theta}} + 4\mathrm{Var}(\hat{\delta}_{\mathbb{K}_\theta}/\delta_{\mathbb{K}_\theta})\frac{\sigma^{-2}_{T_n,\mathbb{K}_\theta}\delta^2_{\mathbb{K}_\theta}}{\Delta^2_{\mathbb{K}_\theta}} \le \frac{4V_{\max,\tilde{\mathbb{K}}_\theta}}{n\delta^2_{\tilde{\mathbb{K}}_\theta}}\frac{\sigma^{-2}_{T_n,\tilde{\mathbb{K}}_\theta}\delta^2_{\tilde{\mathbb{K}}_\theta}}{\Delta^2_{\mathbb{K}_\theta}} + \frac{4V_{\max,\mathbb{K}_\theta}}{n\delta^2_{\mathbb{K}_\theta}}\frac{\sigma^{-2}_{T_n,\mathbb{K}_\theta}\delta^2_{\mathbb{K}_\theta}}{\Delta^2_{\mathbb{K}_\theta}}.
\end{aligned}
$$

Because $\sigma^{-2}_{T_n,\tilde{\mathbb{K}}_\theta}\delta^2_{\tilde{\mathbb{K}}_\theta}/\Delta^2_{\mathbb{K}_\theta} = O(1)$ and $V_{\max,\tilde{\mathbb{K}}_\theta}/(n\delta^2_{\tilde{\mathbb{K}}_\theta}) = o(1/|\mathscr{F}_{\mathbb{K},0}|)$, the first term in the above inequality is a smaller order of $1/|\mathscr{F}_{\mathbb{K},0}|$. Since $\mathbb{K}_\theta \in \mathscr{F}^{(1)}$, we have $V_{\max,\mathbb{K}_\theta}/(n\delta^2_{\mathbb{K}_\theta}) = o(1)$. If $\sigma^{-2}_{T_n,\mathbb{K}_\theta}\delta^2_{\mathbb{K}_\theta}/\Delta^2_{\mathbb{K}_\theta} = O(1/|\mathscr{F}_{\mathbb{K},0}|)$, then the second term in the above inequality is a smaller order of $1/|\mathscr{F}_{\mathbb{K},0}|$. Thus, (S1.17) holds.

In the second case when $\mathbb{K}_\theta \in \mathscr{F}^{(2)}$, we have

$$
\begin{aligned}
&P\left(\hat{\sigma}^{-1}_{T_n,\tilde{\mathbb{K}}_\theta}\hat{\delta}_{\tilde{\mathbb{K}}_\theta} \le \hat{\sigma}^{-1}_{T_n,\mathbb{K}_\theta}\hat{\delta}_{\mathbb{K}_\theta}\right)\\
=&P\left(\hat{\sigma}^{-1}_{T_n,\tilde{\mathbb{K}}_\theta}\hat{\delta}_{\tilde{\mathbb{K}}_\theta} - \sigma^{-1}_{T_n,\tilde{\mathbb{K}}_\theta}\delta_{\tilde{\mathbb{K}}_\theta} \le \hat{\sigma}^{-1}_{T_n,\mathbb{K}_\theta}\hat{\delta}_{\mathbb{K}_\theta} - \sigma^{-1}_{T_n,\tilde{\mathbb{K}}_\theta}\delta_{\tilde{\mathbb{K}}_\theta}\right)\\
=&P\Big\{\hat{\sigma}^{-1}_{T_n,\tilde{\mathbb{K}}_\theta}\hat{\delta}_{\tilde{\mathbb{K}}_\theta} - \sigma^{-1}_{T_n,\tilde{\mathbb{K}}_\theta}\delta_{\tilde{\mathbb{K}}_\theta} \le \frac{1}{n}\frac{\sqrt{n}(\hat{\delta}_{\mathbb{K}_\theta} - \delta_{\mathbb{K}_\theta})}{\sigma_{n,\mathbb{K}_\theta}}\frac{\sqrt{n}\sigma_{n,\mathbb{K}_\theta}}{\sigma_{T_n,\mathbb{K}_\theta}}\frac{\sigma_{T_n,\mathbb{K}_\theta}}{\hat{\sigma}_{T_n,\mathbb{K}_\theta}} + \frac{\delta_{\mathbb{K}_\theta}}{\sigma_{T_n,\mathbb{K}_\theta}}\frac{\sigma_{T_n,\mathbb{K}_\theta}}{\hat{\sigma}_{T_n,\mathbb{K}_\theta}} - \sigma^{-1}_{T_n,\tilde{\mathbb{K}}_\theta}\delta_{\tilde{\mathbb{K}}_\theta}\Big\},
\end{aligned}
$$

where $\sigma^2_{n,\mathbb{K}_\theta}$ is the variance of $\sqrt{n}\hat{\delta}_{\mathbb{K}_\theta}$. We can show that

$$
\frac{\sqrt{n}\sigma_{n,\mathbb{K}_\theta}}{\sigma_{T_n,\mathbb{K}_\theta}} = \sqrt{\frac{n\sigma^2_{n,\mathbb{K}_\theta}}{2V_{2n}(\mathbb{K}_\theta)}} \asymp \sqrt{\frac{n\delta^2_{\mathbb{K}_\theta} + 2V_{2n}(\mathbb{K}_\theta)}{2V_{2n}(\mathbb{K}_\theta)}} = O(1),
$$

indicating that the second term on right hand side of the inequality in the probability is of order $O_p(1/n)$. Similarly, the third term on the right hand side is also of $O_p(1/n)$ for $\mathbb{K}_\theta \in \mathscr{F}^{(2)}$. Hence, we have

$$
\begin{aligned}
P\left(\hat{\sigma}^{-1}_{T_n,\tilde{\mathbb{K}}_\theta}\hat{\delta}_{\tilde{\mathbb{K}}_\theta} \leq \hat{\sigma}^{-1}_{T_n,\mathbb{K}_\theta}\hat{\delta}_{\mathbb{K}_\theta}\right) &= P\left(\sigma^{-1}_{T_n,\tilde{\mathbb{K}}_\theta}\hat{\delta}_{\tilde{\mathbb{K}}_\theta} - \sigma^{-1}_{T_n,\tilde{\mathbb{K}}_\theta}\delta_{\tilde{\mathbb{K}}_\theta} \leq -\sigma^{-1}_{T_n,\tilde{\mathbb{K}}_\theta}\delta_{\tilde{\mathbb{K}}_\theta}\right)\{1+o(1)\} \\
&\leq P\Bigg\{\frac{|\sigma^{-1}_{T_n,\tilde{\mathbb{K}}_\theta}\hat{\delta}_{\tilde{\mathbb{K}}_\theta} - \sigma^{-1}_{T_n,\tilde{\mathbb{K}}_\theta}\delta_{\tilde{\mathbb{K}}_\theta}|}{\sigma^{-1}_{T_n,\tilde{\mathbb{K}}_\theta}\delta_{\tilde{\mathbb{K}}_\theta}} \geq 1\Bigg\}\{1+o(1)\} \leq \mathrm{Var}(\hat{\delta}_{\tilde{\mathbb{K}}_\theta}/\delta_{\tilde{\mathbb{K}}_\theta}) = o(1/|\mathscr{F}_{\mathbb{K},0}|).
\end{aligned}
$$

In summary, we have (S1.18) holds for any $\mathbb{K}_\theta \in \mathscr{F}_{\mathbb{K},0}$. This completes the proof of second part.

(iii) Let $\phi_n(\hat{K}_\theta) \in \{0,1\}$ be the decision rule associated with the kernel $\hat{K}_\theta$ where $\phi_n(\hat{K}_\theta) = 1$ indicates the rejection of null hypothesis. Similarly, we can define $\phi_n(\tilde{K}_\theta)$ and $\phi_n(K)$ for the decision rules associated with the $\tilde{K}_\theta$ and K respectively. Note that for any kernel $\mathbb{K}$ that satisfies equation (4.10) in the main text and alternative $h(x)$ with $\sigma_h^2 = \mathrm{E}\{h^2(\mathbf{X})\} < \infty$,

$$
\begin{aligned}
\phi_n(\mathbb{K}) &= \mathbb{1}\left(\frac{nT_n}{\hat{\sigma}_{T_n,\mathbb{K}}} > z_{1-\alpha}\right) = \mathbb{1}\left(\frac{n\hat{\delta}_{\mathbb{K}}}{\hat{\sigma}^2\hat{\sigma}_{T_n,\mathbb{K}}} > z_{1-\alpha}\right) \\
&= \mathbb{1}\left(\frac{n\delta_{\mathbb{K}}}{\sigma_{T_n,\mathbb{K}}(\sigma^2+\sigma_h^2)} \cdot \frac{\hat{\delta}_{\mathbb{K}}}{\delta_{\mathbb{K}}} \cdot \frac{\sigma_{T_n,\mathbb{K}}}{\hat{\sigma}_{T_n,\mathbb{K}}} > z_{1-\alpha}\right) \xrightarrow{p} 1,
\end{aligned}
$$

as $n \to \infty$, where $\mathbb{1}(\cdot)$ is an indicator function. Since both $\tilde{K}_\theta$ and K satisfy condition (4.10), we have $\phi_n(\tilde{K}_\theta) - \phi_n(K) \xrightarrow{p} 0$. Therefore,

$$
\phi_n(\hat{K}_\theta) - \phi_n(K) = \phi_n(\hat{K}_\theta) - \phi_n(\tilde{K}_\theta) + \phi_n(\tilde{K}_\theta) - \phi_n(K) \xrightarrow{p} 0.
$$

This completes the proof. □

Let T_K be the integral operator defined using kernel K, i.e., $T_K f = \int K(x,\cdot)f(x)d\mu(x)$ for $f \in L^2(\mu)$. Then the eigenvalues corresponding to kernel function K are actually the ones correspond to integral operator, and we denote them by $\lambda(T_K)$ in the following.

Lemma 4. *For the given regularized kernel $K_{R,\gamma}$ in the paper, we have*

$$\lambda_m(T_{K_{R,\gamma}}) = \frac{\gamma \lambda_m(T_K)}{\gamma + \lambda_m(T_K)}.$$

Proof: Applying the result from Dauxois et al. (1982), we have

$$n^{-1}\lambda_m(\mathbf{K}) - \lambda_m(T_K) \overset{a.s.}{\to} 0, \quad n \to \infty \tag{S1.19}$$

and

$$n^{-1}\lambda_m(\mathbf{K}_{R,\gamma}) - \lambda_m(T_{K_{R,\gamma}}) \overset{a.s.}{\to} 0, \quad n \to \infty \tag{S1.20}$$

for any integer m. Next we will show that $\lambda(\mathbf{K}_{R,\gamma}) = \gamma\lambda(\mathbf{K})/\{n^{-1}\lambda(\mathbf{K}) + \gamma\}$. If we assume kernel matrix $\mathbf{K}$ has eigen-decomposition $\mathbf{Q}\boldsymbol{\Lambda}\mathbf{Q}^T$, where $\boldsymbol{\Lambda} = \text{diag}\{\Lambda_1, ..., \Lambda_n\}$, then

$$\begin{aligned}
\mathbf{K}_{R,\gamma} &= \mathbf{Q}\boldsymbol{\Lambda}\mathbf{Q}^T - \mathbf{Q}\boldsymbol{\Lambda}\mathbf{Q}^T(n\gamma\mathbf{Q}\mathbf{Q}^T + \mathbf{Q}\boldsymbol{\Lambda}\mathbf{Q}^T)^{-1}\mathbf{Q}\boldsymbol{\Lambda}\mathbf{Q}^T \\
&= \mathbf{Q}\big\{\boldsymbol{\Lambda} - \boldsymbol{\Lambda}(n\gamma\mathbf{I} + \boldsymbol{\Lambda})^{-1}\boldsymbol{\Lambda}\big\}\mathbf{Q}^T,
\end{aligned}$$

which implies $\lambda_m(\mathbf{K}_{R,\gamma}) = \Lambda_m - \dfrac{\Lambda_m^2}{n\gamma + \Lambda_m} = \dfrac{\gamma\Lambda_m}{\gamma + \Lambda_m/n} = \dfrac{\gamma\lambda_m(\mathbf{K})}{\gamma + \lambda_m(\mathbf{K})/n}$. Hence we have

$$n^{-1}\lambda_m(\mathbf{K}_{R,\gamma}) - \frac{\gamma\lambda_m(T_K)}{\gamma + \lambda_m(T_K)} = \frac{\gamma\lambda_m(\mathbf{K})/n}{\gamma + \lambda_m(\mathbf{K})/n} - \frac{\gamma\lambda_m(T_K)}{\gamma + \lambda_m(T_K)} \xrightarrow{a.s.} 0. \tag{S1.21}$$

Combing (S1.20) and (S1.21), we can see $\lambda_m(T_{K_{R,\gamma}}) = \gamma\lambda_m(T_K)/\{\gamma + \lambda_m(T_K)\}$. $\square$

S2 Regularized kernel and its oracle property

The regularization is most effective in the "sparse" case where the non-zero projections reside only in the first N coordinates corresponding to the N largest eigenvalues. To appreciate that, we hereafter consider the setting where $\lambda_{nm} = c_m\lambda_{1n}$ and $\{c_m\}_{m=1}^{\infty}$ is a decreasing sequence satisfying $c_1 = 1$. Let $\{b_{nm}^2 \asymp B_p, m \in S_1\}$ be the set of non-zero projections whose squares are of the same order as B_p, and S_1 is a subset of $\{1, \cdots, N\}$. Here $a \asymp b$ means that a and b are of the same order.

To show the effectiveness of regularization, we compare the SNR $\Psi_R(d_n, \gamma)$ to an "oracle" SNR $\Psi_R^O(d_n, \gamma)$ using regularized kernel. The oracle SNR is an ideal SNR which eliminates all the coordinates with zero projections. The oracle SNR is used for comparison purpose but it cannot be realized by any test procedure in practice. The oracle SNR $\Psi_R^O(d_n, \gamma)$ is defined as

$$\Psi_R^O(d_n, \gamma) \quad = \quad C_n \cdot \frac{\sum_{m\in S_1} \lambda_{nm} b_{nm}^2/(\lambda_{nm} + \gamma)}{\sqrt{\sum_{m\in S_1} \lambda_{nm}^2/(\lambda_{nm} + \gamma)^2}}.$$

The following theorem provides the maximum orders of $\Psi_R^O(d_n,\gamma)$ and $\Psi_R(d_n,\gamma)$.

Theorem 3. *Let $|S_1|$ be the cardinality of signal set S_1. Assume that the regularization parameter γ^* satisfies $\gamma^* = o(\lambda_{nN}), \gamma^* = O(\lambda_{nN_1}), \lambda_{nN_2} = o(\gamma^*)$, and $R_2/N\gamma^{*2} = o(1)$ where $N_1 = [N\log\log N]$, $N_2 = [N\log N]$, and $R_2 = \sum_{m=N_2}^{\infty}\lambda_{nm}^2$. Then (i) $\max_\gamma \Psi_R^O(d_n,\gamma) \asymp \Psi_R^O(d_n,\gamma^*)$ and both at the order $\sqrt{|S_1|}C_nB_p$ for large p; (ii) there exist constants J_0, J_1 and J_2 such that, for large p,*

$$\frac{J_1|S_1|C_nB_p}{\sqrt{N\log N}} \le \Psi_R(d_n,\gamma^*) \le \frac{J_2|S_1|C_nB_p}{\sqrt{N\{1+J_0\log N(c_{N_2}/c_{N_1})^2\}}}.$$

Proof: (i) Consider the regularized oracle location shift $\Psi_R^O(d_n,\gamma)$, whose order is proportional to

$$f(\gamma) := \frac{\sum_{m\in S_1} g_m(\gamma)}{\sqrt{\sum_{m\in S_1} g_m^2(\gamma)}},$$

where $g_m(\gamma) = \lambda_{nm}/(\lambda_{nm}+\gamma)$. It can be shown that function $f(\gamma)$ is maximized when $g_m(\gamma)$ is a non-zero constant for $m\in S_1$. Denote $f_1(\gamma) = \sum_{m\in S_1} g_m(\theta)$, and $f_2(\gamma) = \sqrt{\sum_{m\in S_1} g_m^2(\gamma)}$. Since $f_1' = \sum g_m'(\gamma)$ and $f_2' = \sum g_m(\gamma)g_m'(\gamma)/f_2$, then $f'(\gamma)=0$ (*i.e.*, $f_1'f_2 - f_1f_2' = 0$) is equivalent to

$$\sum_{m_1\neq m_2\in S_1} g_{m_1}(\gamma)g_{m_2}'(\gamma)\Big(g_{m_1}(\gamma) - g_{m_2}(\gamma)\Big) = 0,$$

where $\hat\gamma = 0$ (i.e., $g_m(\hat\gamma)=1$) is one of the solutions. Then we can show that $\mathrm{sgn}(f'')|_{\gamma=\hat\gamma} = \mathrm{sgn}(f_1''f_2 - f_1f_2'')_{\gamma=\hat\gamma}$, where $(f_1''f_2 - f_1f_2'')_{\gamma=\hat\gamma} = -\sum_{m\in S_1}\lambda_{nm}^{-2}|S_1|^2 + (\sum_{m\in S_1}\lambda_{nm}^{-1})^2|S_1|$ is strictly less than zero when there exists at least one $m \in S_1$ such that $\lambda_{nm} \neq 1$, by using Cauchy-Schwarz inequality. For the case where $\lambda_{nm} = 1$ for all $m \in S_1$, $f(\gamma) = \sqrt{|S_1|}$ does not depend on $\hat{\gamma}$. On the other hand, using the Cauchy-Schwarz inequality, we have $|f(\gamma)| \leq \sqrt{|S_1|}$. Therefore, $\max_\gamma \Psi_R^O(d_n, \gamma) \asymp \max_\gamma f(\gamma) C_n B_p = f(\hat{\gamma}) C_n B_p = \sqrt{|S_1|} C_n B_p$. Furthermore, if $\gamma^* = o(\lambda_{nN})$, then $g_m(\gamma^*)$ at the order of 1 for $m \leq N$ and $\mu_R^O(d_n, \gamma^*) \asymp \sqrt{|S_1|} C_n B_p$. Hence $\max_\gamma \Psi_R^O(d_n, \gamma) \asymp \Psi_R^O(d_n, \gamma^*) \asymp \sqrt{|S_1|} C_n B_p$.

(ii) It is not difficult to see for a regularization parameter γ^* satisfying conditions in Theorem 3, $g_m(\gamma^*) \to 1$ for $m = 1, \cdots, N$, and $g_m(\gamma^*) \to 0$ for $m \geq N_2$. Since $\gamma^* = o(\lambda_{nN})$, there exists $\epsilon_1 > 0$ small enough s.t. $\gamma^* < \epsilon_1 \lambda_{nN}$, hence $|S_1|/(1+\epsilon_1) \leq \sum_{m \in S_1} \lambda_{nm}/\{\lambda_{nm} + \gamma^*\} \leq |S_1|$. Similarly, there exists $\epsilon_2 > 0$ small enough s.t. $\lambda_{nN_2} < \epsilon_2 \gamma^*$, and $R_2/\{(1+\epsilon_2)^2 \gamma^{*2}\} \leq \sum_{m=N_2}^{\infty} \{\lambda_{nm}(\lambda_{nm} + \gamma^*)^{-1}\}^2 \leq R_2/\gamma^{*2}$. Then, we have

$$\Psi_R(d_n, \gamma^*) \geq \frac{J_1 |S_1| C_n B_p}{\sqrt{N \log N + R_2/\gamma^{*2}}} \tag{S2.1}$$

for some positive constant J_1. Assuming $\gamma^* = J_0 \lambda_{nN_1}$, then

$$\Psi_R(d_n, \gamma^*) \leq \frac{J_2 |S_1| C_n B_p}{\sqrt{N(1+\epsilon_2)^2/(1+\epsilon_1)^2 + (N \log N - N) J_0^{-2} (c_{N_2}/c_{N_1})^2 + R_2/\gamma^{*2}}}. \tag{S2.2}$$

Since ϵ_1 and ϵ_2 go to 0, and $R_2/\gamma^{*2} = o(N)$, we obtain the conclusion in part (ii) by combining (S2.1) and (S2.2). □

From Theorem 3, if $|S_1| \asymp N$, we have

$$\frac{J_1\sqrt{|S_1|}C_nB_p}{\sqrt{\log|S_1|}} \leq \Psi_R(d_n,\gamma^*) \leq \frac{J_2\sqrt{|S_1|}C_nB_p}{\sqrt{1+J_0\log|S_1|(c_{N_2}/c_{N_1})^2}}.$$

Therefore, the SNR $\Psi_R(d_n,\gamma^*)$ of the proposed test with regularized kernel can attain the SNR $\Psi_R^O(d_n,\gamma^*)$ of the oracle test within a factor of a slowly varying function $\log(N)$.

The above regularization could enhance the dimensionality that the proposed test could handle. Recall the local alternatives considered in Theorem 2 in the paper. Let $d_n(\mathbf{x}) = b_{R,n}(\gamma^*)\Delta_n(\mathbf{x})$ where $\Delta_n(\mathbf{x})$ is a function such that $\mathrm{E}\left\{K_\theta(\mathbf{X}_1,\mathbf{X}_2)\Delta_n(\mathbf{X}_1)\Delta_n(\mathbf{X}_2)\right\}$ is a constant. Using the regularized kernel with regularization parameter γ^*, the proposed test has a non-trivial power if $b_{R,n}(\gamma^*)$ is at the order $b_{R,n}(\gamma^*) = V_{2n}^{1/4}\rho^{1/4}(\gamma^*)/\sqrt{n}$ where

$$\rho(\gamma^*) = \Big(\frac{\sum_{m=1}^{\infty}\lambda_{nm}^2/(\lambda_{nm}/\gamma^*+1)^2}{\{\sum_{m\in S_1}\lambda_{nm}b_{nm}^2/(\lambda_{nm}/\gamma^*+1)\}^2}\Big)\Big(\frac{V_2}{\{\sum_{m\in S_1}\lambda_{nm}b_{nm}^2\}^2}\Big)^{-1}.$$

Assume γ^* satisfies the conditions in Theorem 3. Then we have

$$\rho(\gamma^*) \asymp \frac{N}{|S_1|^2}\cdot\big(\sum_{m\in S_1} c_m\big)^2.$$

If $|S_1| \asymp N$ and $c_m = m^{-\alpha}$ for $\alpha > 1/2$, then $\rho(\gamma^*) = O(N^{-\min\{2\alpha-1,1\}}) = o(1)$. This means that the smallest detectable order using a regularized kernel is smaller than that of an unregularized kernel. The improvement is significant when N is large and $\alpha > 1$. Moreover, the test is consistent if $V_{2n} = o\{n^2\rho^{-1}(\gamma^*)\}$. Comparing to the unregularized case which requires $V_{2n} = o\{n^2\}$, the regularized kernel is powerful for higher dimensional functions since $\rho(\gamma^*) \to 0$.

S3 Some additional simulation results

S3.1 Simulation studies with $p < n$

In this subsection, we present some additional simulation studies in the cases of $p < n$. The simulation settings are introduced in Section 5 of the main paper with the $h(\mathbf{x}) = h_L(\mathbf{x}) - E(h_L)$ in setting (i) under the alternatives where

$$h_L(\mathbf{x}) = c_1(x_1 + x_2 - x_3) + c_2\{\exp(-x_2^2)H_2(x_2) + \exp(-x_3^2)H_5(x_3)\} + c_3\{x_1x_3 + \cos(x_3^2)\},$$

where $H_k(\cdot)$ is the kth order Hermite polynomial. We considered two scenarios with $\mathcal{S}_1 = \{c_1 = 0.002, c_2 = 0.2, c_3 = 0.002\}$ and $\mathcal{S}_2 = \{c_1 = 1.2, c_2 = 0.012, c_3 = 0.012\}$. In scenarios $\mathcal{S}_1$, c_2 are chosen to be much larger than c_1 such that the non-linear parts dominate the functions while in $\mathcal{S}_2$, c_1 are much larger than c_2 so that the linear parts dominate.

Table S1 summarizes the empirical size of the proposed test under low-dimensional cases with normally and Laplace distributed errors at the nominal level 5%. We can see that the empirical size of the proposed test was reasonably controlled at the nominal level for all three types of kernels and different error distributions.

Table S2 and S3, respectively, contain the empirical power of the proposed test for scenarios $\mathcal{S}_1$ and $\mathcal{S}_2$ under the setting (i). Several observations are given below. (1) There is a clear difference in power among the three types of kernels K_E, K_G and K_L, especially when p and n are relatively small. The power difference was especially striking in Table S3 for scenarios $\mathcal{S}_1$. The power based on the exponential and Gaussian kernels were both higher than that using the linear kernel. This is understandable since the non-linear parts dominate the function $h_L(\mathbf{x})$ in scenarios $\mathcal{S}_1$ and exponential kernel and Gaussian kernel contain richer non-linear eigenfunctions than that of the linear kernel, which can capture more information of non-linear functions; (2) The power increased as the sample size increased in all the cases; and (3) The proposed test was very robust to the change of error distributions.

Table S1: Empirical size (in percentages) of the proposed test for Gaussian and Laplace errors with low-dimensional dependent covariates using different kernels.

		Gaussian Error			Laplace Error		
n	p	K_E	K_L	K_G	K_E	K_L	K_G
40	3	5.8	5.5	5.6	5.1	4.6	4.7
	5	6.0	5.8	5.9	4.9	4.8	4.8
	10	6.0	6.2	6.2	4.7	4.4	4.6
60	3	5.5	5.4	5.6	4.6	4.6	4.6
	5	5.8	5.5	5.6	5.3	5.4	5.5
	10	5.6	5.4	5.6	4.6	4.8	4.8
100	3	5.3	5.4	5.6	4.8	4.6	4.7
	5	5.3	5.3	5.3	4.3	4.0	4.1
	10	5.5	5.3	5.4	4.9	4.5	4.6

Table S2: Empirical power (in percentages) of the proposed test for Gaussian and Laplace errors with dependent covariates using different kernels in the setting of $p << n$ under scenario $\mathcal{S}_1$. The estimated theoretical power is given in the parenthesis, and the percentage of a kernel being selected among the three candidate kernels using the proposed kernel selection method is displayed underneath it.

		Gaussian Error			Laplace Error		
n	p	K_E	K_L	K_G	K_E	K_L	K_G
40	3	72.6 (65.5)	10.5 (17.3)	29.6 (37.4)	76.0 (67.5)	11.7 (19.0)	32.0 (39.2)
		(96.8)	(0.0)	(3.2)	(95.9)	(0.0)	(4.1)
	5	27.0 (34.8)	9.9 (16.7)	15.2 (24.1)	26.2 (34.0)	10.8 (17.5)	15.5 (24.1)
		(79.5)	(0.0)	(20.5)	(84.2)	(0.2)	(15.6)
	10	14.9 (21.8)	9.6 (16.0)	11.5 (18.6)	15.8 (23.1)	10.4 (16.8)	13.6 (20.0)
		(63.1)	(1.7)	(35.2)	(64.1)	(1.9)	(34.0)
60	3	99.2 (93.4)	12.6 (21.3)	70.4 (59.7)	99.0 (92.9)	13.9 (22.2)	70.8 (60.9)
		(99.5)	(0.0)	(0.5)	(98.6)	(0.0)	(1.4)
	5	51.5 (51.4)	11.7 (19.3)	24.1 (32.2)	52.3 (52.4)	12.4 (19.6)	24.9 (33.0)
		(90.9)	(0.0)	(9.1)	(93.2)	(0.0)	(6.8)
	10	18.5 (26.4)	9.8 (17.0)	12.4 (20.9)	19.6 (27.0)	10.9 (17.7)	14.2 (20.9)
		(72.0)	(0.4)	(27.6)	(70.5)	(0.2)	(29.3)
100	3	100 (100)	28.6 (35.9)	100 (98.8)	100 (100)	30.3 (36.7)	100 (100)
		(100)	(0.0)	(0.0)	(100)	(0.0)	(0.0)
	5	98.2 (88.8)	17.8 (26.6)	64.6 (57.7)	95.8 (87.1)	12.9 (27.2)	63.0 (57.8)
		(97.9)	(0.0)	(2.1)	(97.8)	(0.0)	(2.2)
	10	36.5 (42.0)	15.0 (21.8)	22.2 (30.0)	34.7 (40.4)	13.6 (20.7)	19.4 (28.5)
		(81.2)	(0.1)	(18.7)	(80.1)	(0.0)	(19.9)

S3.2 Simulation studies with $p > n$

The simulation setups for $p > n$ are similar to those given in Section 5 of the paper except that we consider $n = 40, 60$ and 100, and $p = (150, 200, 250)$ in this subsection. The true function $h(\cdot)$ was chosen as $h_H(\cdot)$ as defined in Section 5 except that we consider two more scenarios $\mathcal{S}_5$ and $\mathcal{S}_6$ regarding the choice of c_1, c_2 and c_3. In particular, $\mathcal{S}_5 = \{c_1 = 0.01, c_2 = 10, c_3 = 0.01\}$ and $\mathcal{S}_6 = \{c_1 = 100u, c_2 = 0.1u, c_3 = 0.1u, u = 0.0015\}$. In scenarios $\mathcal{S}_5$, c_2 are chosen to be much larger than c_1 such that the non-linear parts

Table S3: Empirical power (in percentages) of the proposed test for Gaussian and Laplace errors with dependent covariates using different kernels in the setting of $p << n$ under scenario $\mathcal{S}_2$. The estimated theoretical power is given in the parenthesis, and the percentage of a kernel being selected among the three candidate kernels using the proposed kernel selection method is displayed underneath it.

		Gaussian Error			Laplace Error		
n	p	K_E	K_L	K_G	K_E	K_L	K_G
40	3	60.5 (60.5)	62.1 (61.0)	63.4 (62.3)	60.2 (61.0)	61.0 (60.5)	62.1 (61.8)
		(37.5)	(32.3)	(30.2)	(37.9)	(35.5)	(26.6)
	5	49.6 (49.5)	49.4 (49.8)	50.7 (51.1)	43.4 (46.2)	45.0 (46.5)	46.3 (47.8)
		(38.6)	(30.2)	(31.2)	(37.7)	(29.6)	(32.7)
	10	33.8 (37.5)	32.9 (36.8)	34.6 (37.9)	29.8 (35.3)	31.5 (35.5)	32.9 (36.4)
		(39.7)	(33.3)	(27.0)	(36.6)	(32.7)	(30.7)
60	3	79.3 (78.2)	79.0 (77.6)	80.3 (78.8)	78.6 (77.4)	78.1 (77.1)	79.1 (78.3)
		(38.5)	(29.4)	(32.1)	(35.6)	(29.6)	(34.8)
	5	66.8 (65.6)	69.4 (66.3)	70.5 (67.8)	69.8 (68.3)	70.3 (68.5)	72.5 (70.0)
		(35.1)	(24.3)	(40.6)	(33.9)	(25.0)	(41.1)
	10	49.2 (50.1)	49.9 (50.7)	51.7 (52.0)	49.8 (50.5)	50.6 (50.6)	51.8 (51.9)
		(35.0)	(29.9)	(35.1)	(35.0)	(30.7)	(34.3)
100	3	97.1 (95.6)	96.9 (95.6)	97.4 (96.1)	96.8 (96.0)	96.5 (96.0)	96.8 (96.4)
		(39.4)	(23.2)	(37.4)	(35.0)	(26.2)	(38.8)
	5	92.1 (89.7)	92.8 (90.5)	93.2 (91.2)	89.6 (87.7)	89.7 (87.6)	90.7 (88.6)
		(33.2)	(18.2)	(48.6)	(34.4)	(18.2)	(47.4)
	10	78.6 (75.6)	80.0 (76.6)	81.0 (77.7)	78.1 (75.0)	78.0 (75.7)	79.4 (76.9)
		(31.4)	(23.3)	(45.3)	(32.7)	(20.1)	(47.2)

dominate the functions while in $\mathcal{S}_6$, c_1 are much larger than c_2 so that the linear parts dominate. All the results for evaluating empirical power are based on 1000 simulation replicates and that for empirical size are based on 5000 simulation replicates. The kernels considered here are the same as those considered in Section 5 of the paper.

Table S4 summarizes the empirical size of the proposed test with normally and Laplace distributed errors at the nominal level 5%. We can see that the empirical size of the proposed test was reasonably controlled at the nominal level for all three types of kernels and different error distributions. The corresponding empirical powers are reported in Tables S5 and S6 under setting (ii). The phenomena we observed in Tables S5 and S6 are very similar to those in Table 2 in the paper.

To check the performance of the proposed method when the underlying function contains "equivalent" linear and non-linear functions, we enlarge the coefficient c_3 but make the coefficients c_1 and c_2 to be much smaller. We also compare our proposed method with the method proposed by Liu et al. (2007). The simulation results are summarized in Table S7.

Table S4: Empirical size (in percentages) of the proposed test with $p > n$ for Gaussian and Laplace errors with dependent covariates using different kernels.

		Gaussian Error			Laplace Error		
n	p	K_E	K_L	K_G	K_E	K_L	K_G
40	150	5.9	5.8	5.9	4.2	4.2	4.0
	200	6.4	6.4	6.4	4.7	4.7	4.7
	250	5.7	5.5	5.6	4.9	4.7	4.7
60	150	5.4	5.3	5.3	4.9	4.7	4.8
	200	5.7	5.9	5.8	4.1	4.0	4.1
	250	6.0	5.9	5.9	4.9	4.7	4.8
100	150	5.3	5.4	5.4	5.0	5.0	5.0
	200	5.5	5.4	5.4	5.4	5.2	5.3
	250	5.4	5.4	5.5	4.9	4.8	4.9

S3.3 Simulation studies with $p >> n$ and Laplace errors

In this subsection, we include simulation results for $p >> n$ with the same settings as those reported in Section 5 of the main paper. But we replace the Gaussian errors by Laplace errors to check the robustness of the proposed methods to the changes of the

Table S5: Empirical power (in percentages) of the proposed test for Gaussian and Laplace errors with dependent covariates using different kernels in the setting of $p > n$ under scenario $\mathcal{S}_5$. The estimated theoretical power is given in the parenthesis, and the percentage of a kernel being selected among the three candidate kernels using the proposed kernel selection method is displayed underneath it.

		Gaussian Error			Laplace Error		
n	p	K_E	K_L	K_G	K_E	K_L	K_G
40	150	70.6 (71.6)	64.0 (64.8)	63.9 (65.3)	71.0 (72.6)	65.4 (72.6)	65.3 (66.5)
		(94.8)	(2.9)	(2.3)	(95.3)	(1.9)	(2.8)
	200	42.7 (50.7)	39.6 (45.7)	39.6 (46.3)	39.0 (48.1)	34.4 (43.2)	34.8 (43.7)
		(89.4)	(3.4)	(7.2)	(86.6)	(4.8)	(8.6)
	250	26.1 (37.1)	23.8 (33.4)	26.5 (34.0)	27.0 (38.3)	24.6 (34.8)	24.6 (35.3)
		(81.8)	(4.4)	(13.8)	(82.9)	(5.0)	(12.1)
60	150	89.0 (86.9)	84.2 (81.5)	84.1 (81.8)	89.8 (87.6)	86.1 (82.6)	86.4 (82.8)
		(96.6)	(2.4)	(1.0)	(96.9)	(2.4)	(0.7)
	200	59.6 (61.5)	53.5 (56.0)	53.1 (56.4)	59.1 (61.3)	53.8 (56.2)	53.4 (56.6)
		(89.7)	(6.2)	(4.1)	(89.6)	(5.2)	(5.2)
	250	40.2 (45.7)	35.3 (41.9)	35.4 (42.3)	35.0 (43.0)	32.1 (39.5)	32.1 (40.0)
		(81.9)	(8.9)	(9.2)	(82.7)	(8.1)	(9.2)
100	150	99.2 (98.4)	98.5 (96.9)	98.6 (97.0)	99.0 (98.2)	98.4 (96.9)	98.4 (96.9)
		(99.2)	(0.8)	(0.0)	(99.1)	(0.8)	(0.1)
	200	88.5 (85.7)	84.2 (81.8)	84.4 (82.0)	87.0 (84.0)	82.8 (79.8)	82.9 (79.9)
		(94.1)	(3.9)	(2.0)	(94.9)	(3.6)	(1.5)
	250	61.8 (63.1)	58.4 (59.0)	58.5 (59.3)	63.8 (64.0)	60.0 (60.2)	59.9 (60.4)
		(84.3)	(10.7)	(5.0)	(87.4)	(8.1)	(4.5)

Table S6: Empirical power (in percentages) of the proposed test for Gaussian and Laplace errors with dependent covariates using different kernels in the setting of $p > n$ under scenario $\mathcal{S}_6$. The estimated theoretical power is given in the parenthesis, and the percentage of a kernel being selected among the three candidate kernels using the proposed kernel selection method is displayed underneath it.

		Gaussian Error			Laplace Error		
n	p	K_E	K_L	K_G	K_E	K_L	K_G
40	150	67.1 (69.1)	67.3 (68.2)	67.4 (68.7)	66.7 (68.7)	67.5 (68.0)	67.2 (68.4)
		(47.8)	(20.8)	(31.4)	(44.8)	(22.1)	(33.1)
	200	59.5 (61.6)	60.5 (62.6)	60.4 (63.1)	59.4 (63.4)	60.0 (63.2)	60.0 (63.7)
		(53.9)	(14.2)	(31.9)	(53.8)	(13.6)	(32.6)
	250	54.4 (59.7)	54.1 (58.0)	54.3 (58.7)	52.7 (58.9)	53.4 (60.0)	52.5 (58.4)
		(62.4)	(8.8)	(28.8)	(63.1)	(9.6)	(27.3)
60	150	90.3 (87.1)	90.3 (87.1)	90.4 (87.3)	86.0 (84.1)	86.5 (84.1)	86.6 (84.2)
		(36.1)	(34.6)	(29.3)	(36.7)	(33.1)	(30.2)
	200	81.7 (81.0)	82.1 (80.8)	82.3 (81.1)	82.4 (80.9)	83.2 (80.7)	83.3 (81.0)
		(46.6)	(27.6)	(27.8)	(44.2)	(26.6)	(31.2)
	250	73.3 (76.1)	73.3 (75.7)	74.4 (76.4)	76.8 (76.2)	77.3 (76.8)	77.2 (76.5)
		(51.2)	(19.7)	(29.1)	(50.7)	(20.1)	(29.2)
100	150	99.3 (98.6)	99.1 (98.5)	99.2 (98.6)	98.9 (97.9)	99.0 (98.0)	99.0 (98.0)
		(24.9)	(45.8)	(29.3)	(26.2)	(46.4)	(27.4)
	200	98.3 (97.2)	98.6 (97.3)	98.6 (97.3)	98.5 (96.6)	98.5 (96.7)	98.5 (96.7)
		(31.6)	(39.1)	(29.3)	(31.1)	(41.9)	(27.0)
	250	98.0 (96.0)	98.1 (96.0)	98.1 (96.0)	97.2 (95.1)	97.2 (95.1)	97.2 (95.1)
		(33.3)	(38.3)	(28.4)	(35.2)	(34.8)	(30.0)

error distributions.

Table S8 summarizes the empirical sizes of the proposed test and the test procedure (LLD) proposed by Liu et al. (2007) for high-dimensional and functional covariates with Laplace errors. We see that both methods were robust to the change of error distributions and can control the type I errors reasonably well under Laplace errors. The empirical power of the proposed test and LLD test are summarized in Table S9. We can see that the power patterns for Laplace errors were very similar those for Gaussian errors. The

Table S7: Empirical power (in percentages) of the proposed test (Proposed) and the method (LLD) proposed by Liu et al. (2007) for Gaussian and Laplace errors with dependent covariates using different kernels when $(c_1, c_2, c_3) = (0.1, 0.1, 10)$. The estimated theoretical power is given in the parenthesis, and the percentage of a kernel being selected among the three candidate kernels using the proposed kernel selection method is displayed underneath it.

			Gaussian Error			Laplace Error		
n	p	method	K_E	K_L	K_G	K_E	K_L	K_G
60	200	Proposed	83.0(76.4)	83.0(76.5)	83.2(76.7)	83.8(77.8)	83.7(77.7)	83.8(77.9)
			(35.9)	(32.3)	(31.8)	(36.1)	(33.6)	(30.3)
		LLD	79.1	80.5	80.7	79.7	81.9	81.8
	250	Proposed	76.3(70.4)	74.9(70.2)	75.3(70.4)	74.9(69.6)	75.1(69.4)	75.1(69.7)
			(39.7)	(32.3)	(28.0)	(37.7)	(31.3)	(31.0)
		LLD	71.9	73.0	73.3	69.2	71.9	72.1
100	200	Proposed	98.8(96.6)	98.9(96.7)	98.9(96.7)	99.9(96.7)	99.0(96.7)	99.0(96.8)
			(34.7)	(34.1)	(31.2)	(31.8)	(37.2)	(31.0)
		LLD	98.6	98.7	98.7	98.9	99.0	99.0
	250	Proposed	97.6(94.9)	97.8(94.9)	97.8(95.0)	98.9 (94.9)	99.1(94.9)	99.0(95.0)
			(35.1)	(36.6)	(28.3)	(35.9)	(34.9)	(29.2)
		LLD	97.6	97.8	97.7	98.3	98.7	98.7

kernel selection results in Table S9 were also very similar to the kernel selection results with Gaussian errors in the main paper. This indicates that the proposed kernel selection method was also robust to the change of error distributions.

S3.4 Simulation studies with regularization

In addition to the regularized tests with exponential kernels reported in the main text, we include more simulation results with regularized Gaussian and linear kernels in this subsection. We generated data and chose the function $h(\mathbf{x})$ similar to those described in Section 5.2 in the main paper. For each kernel K_L and K_G, we constructed the regularized kernels with regularization parameter γ. We selected a sequence of regularization

Table S8: Empirical size (in percentages) of the proposed test (Proposed) and Liu et al. (2007)'s method (LLD) for Laplace errors with high-dimensional ($p >> n$) and functional covariates using different kernels.

			High-dimensional covariates			Functional covariates		
n	p	method	K_E	K_L	K_G	K_E	K_L	K_G
40	1500	Proposed	4.7	4.8	4.7	4.3	4.9	4.8
		LLD	4.1	4.2	4.3	4.3	4.1	4.2
	3000	Proposed	4.4	4.3	4.3	5.2	6.1	5.8
		LLD	3.8	4.0	4.0	5.0	5.8	5.7
	4500	Proposed	4.1	4.4	4.1	5.5	5.2	5.4
		LLD	4.4	4.0	4.0	5.1	5.2	5.3
60	1500	Proposed	5.2	5.1	5.0	4.8	5.0	5.0
		LLD	5.3	5.0	5.2	4.8	5.0	5.0
	3000	Proposed	5.6	5.4	5.5	4.5	4.5	4.2
		LLD	5.7	5.5	5.6	4.5	4.3	4.2
	4500	Proposed	4.9	4.8	4.8	4.9	4.2	4.4
		LLD	5.2	5.4	5.4	4.9	4.5	5.0
100	1500	Proposed	4.7	4.3	4.3	5.3	5.0	5.1
		LLD	4.9	4.4	4.4	5.5	4.9	4.9
	3000	Proposed	5.0	4.9	5.0	4.6	5.1	4.9
		LLD	5.0	4.8	4.8	4.8	5.3	5.1
	4500	Proposed	4.6	4.5	4.4	4.6	4.5	4.7
		LLD	4.3	4.7	4.6	4.4	4.8	5.0

Table S9: Empirical power (in percentages) of the proposed test (Proposed) and Liu et al. (2007)'s method (LLD) for Laplace errors with dependent covariates using different kernels under scenarios $\mathcal{S}_3$ and $\mathcal{S}_4$. The estimated theoretical power is given in the parenthesis, and the percentage of a kernel being selected among the three candidate kernels is displayed underneath it.

			$\mathcal{S}_3$			$\mathcal{S}_4$		
n	p	method	K_E	K_L	K_G	K_E	K_L	K_G
	1500	Proposed	49.8(49.8)	47.3(47.6)	47.4(47.7)	58.9(57.2)	59.3(57.0)	59.2(57.1)
			(84.1)	(11.0)	(4.9)	(38.7)	(34.0)	(27.3)
		LLD	42.8	42.3	42.3	52.6	53.8	53.9
	3000	Proposed	27.7(33.1)	27.0(32.7)	27.1(32.8)	38.0(41.2)	37.8(41.1)	38.2(41.2)
40			(50.4)	(29.9)	(19.7)	(36.2)	(38.1)	(25.7)
		LLD	22.4	21.5	21.8	31.7	32.3	32.5
	4500	Proposed	20.6(27.2)	19.8(27.0)	20.1(27.1)	31.8(35.5)	31.8(35.4)	32.0(35.4)
			(42.5)	(37.0)	(20.5)	(36.2)	(40.4)	(23.4)
		LLD	15.8	16.7	16.8	26.9	27.8	27.8
	1500	Proposed	74.2(69.1)	71.3(66.4)	71.3(66.5)	85.3(79.2)	85.0(79.1)	85.2(79.1)
			(93.9)	(3.6)	(2.5)	(37.1)	(33.0)	(29.9)
		LLD	70.8	69.1	69.1	83.9	83.5	83.5
	3000	Proposed	45.2(45.5)	44.5(45.0)	44.4(45.0)	62.6(60.0)	62.4(59.7)	62.5(59.8)
60			(59.7)	(24.6)	(15.7)	(39.4)	(31.3)	(29.3)
		LLD	40.9	40.8	40.7	58.8	59.8	59.8
	4500	Proposed	31.1(34.9)	30.4(34.7)	30.1(34.8)	51.7(51.3)	51.2(51.1)	51.7(51.2)
			(43.4)	(35.0)	(21.6)	(39.0)	(33.3)	(27.7)
		LLD	27.5	27.4	27.4	48.1	48.2	48.3
	1500	Proposed	98.3(94.1)	97.7(92.6)	97.7(92.7)	99.8(98.5)	99.8(98.5)	99.8(98.5)
			(98.6)	(0.6)	(0.8)	(34.8)	(37.3)	(27.9)
		LLD	98.0	97.7	97.7	99.8	99.8	99.8
	3000	Proposed	76.8(69.2)	75.2(68.4)	75.3(68.5)	95.1(88.5)	95.0(88.5)	95.0(88.5)
100			(68.5)	(17.7)	(13.8)	(37.8)	(33.9)	(28.3)
		LLD	74.4	73.4	73.4	94.5	94.7	94.7
	4500	Proposed	56.0(55.8)	55.7(54.5)	55.7(54.6)	85.1(78.1)	85.2(78.0)	85.3(78.0)
			(52.9)	(26.9)	(20.2)	(40.9)	(33.0)	(26.1)
		LLD	54.0	54.0	54.0	83.6	83.2	83.2

parameters of different orders ($\gamma = 10^{-a}/n$, $a \in (-5, 2)$) to check their effects on empirical power. For each regularization parameter value, we constructed the corresponding regularized test statistic and applied the test, respectively, to data generated under H_0 and H_1.

Figure S1 shows the empirical power and size of the proposed test using regularized kernel $K_{R,\gamma}$. The x-axis represents the $-\log_{10}(\gamma)$ and y-axis is the empirical power or size. The power with large regularization parameters γ was not displayed in the graph for a better view for small γ range. When γ is large $-\log_{10}(\gamma) \in (-3.222, 1.778)$, not shown in Figure S1, the power of the test was the same as the one using non-regularized kernels (0.674, and 0.672 for K_G and K_L), and then started to grow slowly. As for $-\log_{10}\gamma \in (1.778, 3.778)$, the power peak (0.720 and 0.710, for K_G and K_L respectively) of the proposed test can be observed for all the three kernels. It can be seen from Figure S1 that the empirical size of the regularized test was all reasonably controlled.

To evaluate the method for selecting regularization parameters proposed in Section 4.2, we also marked the regularization parameter selection results in Figure S1. The three vertical lines correspond to the first quantile (Q_1), median and third quantile (Q_3) of the stabilized $\tilde{\gamma}$ obtained from the 1000 simulation replicates, where $L = 5$ were chosen in stability selection. It can be seen from Figure S1 that the vertical lines were all very close to the place where the maximum power was achieved. This suggests that the proposed regularization selection method can locate the optimal regularization parameter to maximize the power of the proposed test.

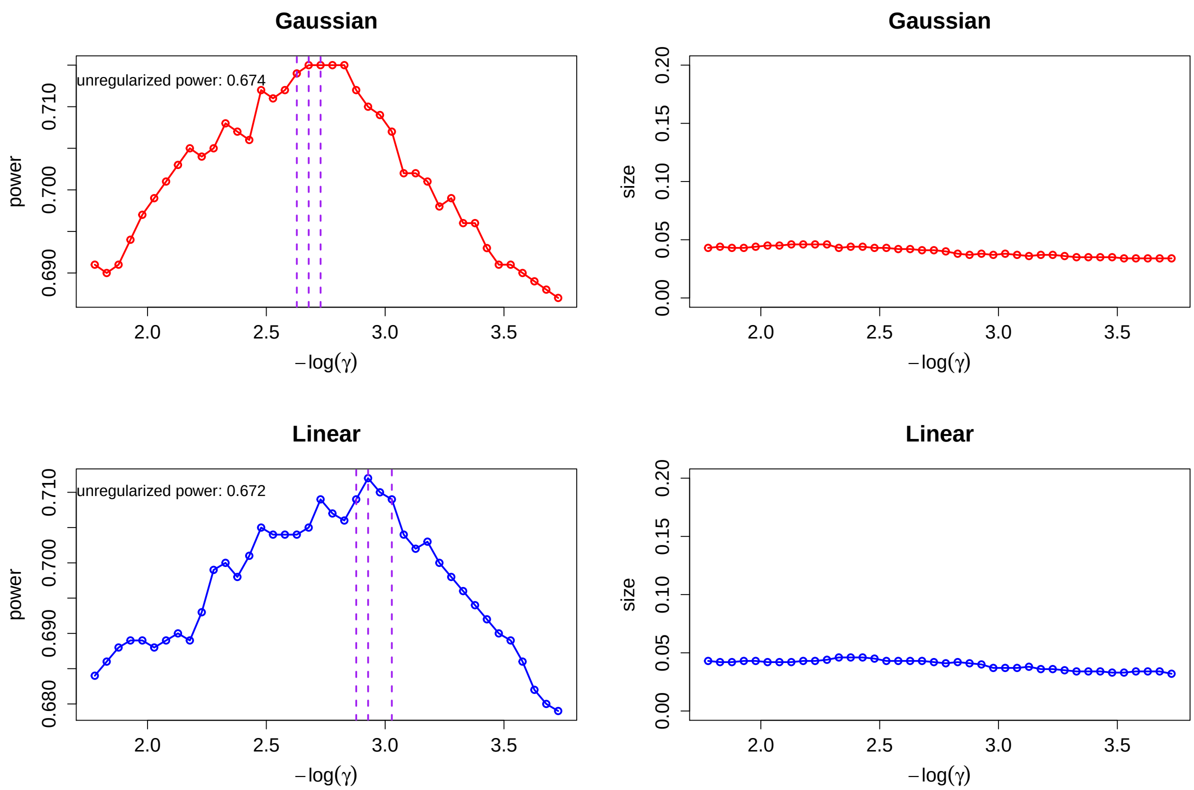

Figure S1: The empirical power (left panel) and size (right panel) for regularized kernels, where the vertical purple lines in the left panel denote the first, second and third quantile of the selected regularization parameters among 1000 simulation replicates. For each replicate, the regularization parameter was selected by the method introduced in Section 4.2 in the main paper.

S3.5 Impact of kernel parameters

To understand the impact of the kernel parameters on the proposed method, we evaluate our proposed method and the LLD method using linear kernel $K_L(\mathbf{x}, \mathbf{y}) = \mathbf{x}^T\mathbf{y}/\theta$, Gaussian kernel $K_G(\mathbf{x}, \mathbf{y}) = \exp\{-\|\mathbf{x}-\mathbf{y}\|^2/\theta\}$ and the exponential kernel $K_E(\mathbf{x}, \mathbf{y}) = \exp\{-(\|\mathbf{x}\|^2 + 3\|\mathbf{x}-\mathbf{y}\|^2 + \|\mathbf{y}\|^2)/\theta\}$ with two different kernel parameters $\theta = 2\sqrt{p}$ and $\theta = p$. The results are summarized in the Table S10. The results show that our proposed method is robust in the kernel parameters in the sense that they have similar empirical size and power with different kernel parameters. However, we observe that the LLD test may be sensitive to the kernel parameter. In particular, when $\theta = 2\sqrt{p}$, one can see that the LLD test lost power for exponential kernel in both scenarios $\mathcal{S}_3$ and $\mathcal{S}_4$, and had a size distortion for exponential kernel with Laplace error. These results are understandable because the LLD test was not designed for high dimensional settings.

S3.6 Computational time

To understand the computational cost for the proposed tests with and without regularized kernels, in Figures S2 and S3, we summarize the mean and standard deviation of computational time using the simulation experiments in the scenario $\mathcal{S}_3$ described at the beginning of Section 5.2 in the main paper. We considered various data dimension $p = 200, 500, 1500$, and sample size $n = 60, 100$. We chose exponential kernel for our illustration. The mean and standard deviation of computational time were calculated

Table S10: Empirical size and power (in percentages) of the proposed test (Proposed) and the method (LLD) proposed by Liu et al. (2007) for Gaussian and Laplace errors using different kernels and different kernel parameter θ when $n = 60$ and $p = 1500$ in scenarios $\mathcal{S}_3$ and $\mathcal{S}_4$.

			Gaussian Error			Laplace Error		
θ	Scenarios	method	K_E	K_L	K_G	K_E	K_L	K_G
$2\sqrt{p}$	Size	Proposed	4.9	5.1	5.5	4.0	4.9	5.3
		LLD	4.7	4.5	4.8	12.7	4.3	4.9
	$\mathcal{S}_3$	Proposed	70.5	76.2	76.2	68.0	71.5	72.3
		LLD	5.8	73.1	73.6	4.9	69.2	69.6
	$\mathcal{S}_4$	Proposed	69.2	84.0	85.0	68.5	84.6	84.9
		LLD	5.0	81.9	82.7	4.8	83.6	83.3
p	Size	Proposed	5.3	5.1	5.1	5.6	5.4	5.5
		LLD	3.9	4.0	4.0	5.3	5.0	5.2
	$\mathcal{S}_3$	Proposed	76.3	74.1	74.2	74.2	71.3	71.3
		LLD	73.4	71.9	71.8	70.8	69.1	69.1
	$\mathcal{S}_4$	Proposed	84.4	84.3	84.2	85.3	85.0	85.2
		LLD	83.0	83.9	83.7	83.9	83.5	83.5

based on 1000 simulation replications. We observed that the average computational cost both tests with and without regularization are low. As expected, the computation time for tests with regularization is longer than tests without regularization. The computational time is about linear in data dimension, which indicates that the proposed tests are computationally efficient for high-dimensional data sets.

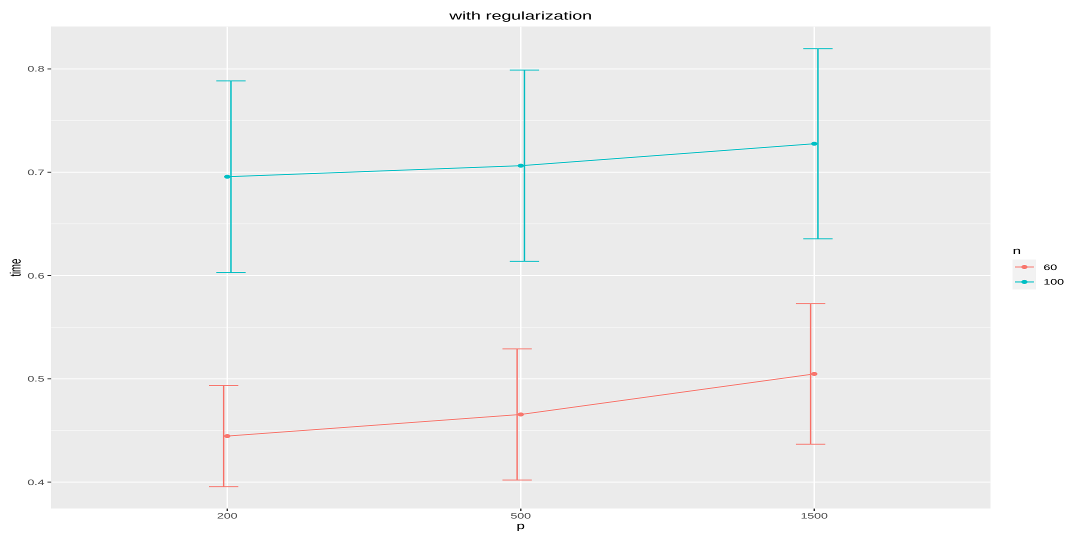

Figure S2: The mean and standard deviation (vertical bars) of the computational time for the proposed tests with kernel regularization when $p = 200, 500$ and 1500, and $n = 60$ and 100. The tests with kernel regularization are described in Section 5.2 in the main paper. The simulation settings are the same as those described in Section 5.2 in the main paper.

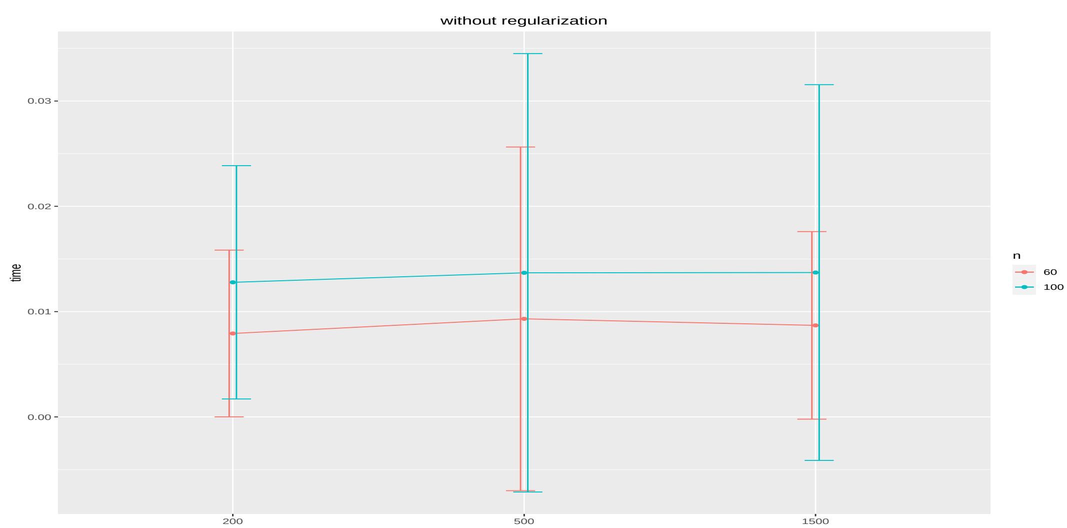

Figure S3: The mean and standard deviation (vertical bars) of the computational time for the proposed tests without kernel regularization when $p = 200, 500$ and 1500, and $n = 60$ and 100. The simulation settings are the same as those described in Section 5.2 in the main paper.

References

Bao, Y. and A. Ullah (2010). Expectation of quadratic forms in normal and nonnormal variables with applications. *Journal of Statistical Planning and Inference 140*(5), 1193–1205.

Billingsley, P. (2013). *Convergence of probability measures.* John Wiley & Sons.

Dauxois, J., A. Pousse, and Y. Romain (1982). Asymptotic theory for the principal component analysis of a vector random function: some applications to statistical inference. *Journal of multivariate analysis 12*(1), 136–154.

Liang, Z. and Y. Lee (2013). Eigen-analysis of nonlinear pca with polynomial kernels. *Statistical Analysis and Data Mining: The ASA Data Science Journal 6*(6), 529–544.

Liu, D., X. Lin, and D. Ghosh (2007). Semiparametric regression of multidimensional genetic pathway data: least-squares kernel machines and linear mixed models. *Biometrics 63*(4), 1079–1088.

Mikosch, T. (1993). A weak invariance principle for weighted u-statistics with varying kernels. *Journal of multivariate analysis 47*(1), 82–102.

Zhong, P.-S. and S. X. Chen (2011). Tests for high-dimensional regression coefficients with factorial designs. *Journal of the American Statistical Association 106*(493), 260–274.