Online Supplementary Materials for “Sufficient dimension reduction via distance covariance”

Wenhui Sheng and Xiangrong Yin *

December 10, 2014

Abstract

In this supplement, section A, section B and section C provide detailed proofs for Proposition 3, Proposition 4 and Corollary 1 in the paper respectively.

A Proof of Consistency

In order to prove the Proposition 3 in section 2.5 of the paper, we prove the following Lemma A first.

A.1 Proof of Lemma A

Lemma A If the support of X , say S , is compact, $E|Y| < \infty$ and furthermore, $\boldsymbol{\eta}_n \xrightarrow{P} \boldsymbol{\eta}$, then $\mathcal{V}_n^2(\boldsymbol{\eta}_n^T \mathbf{X}, \mathbf{Y}) - \mathcal{V}_n^2(\boldsymbol{\eta}^T \mathbf{X}, \mathbf{Y}) \xrightarrow{P} 0$.

Proof of Lemma A: Based on Székely, Rizzo and Bakirov (2007) definition 4, we have that for $k, l = 1, \dots, n$,

$$\begin{aligned} a_{kl}(\boldsymbol{\eta}_n) &= |\boldsymbol{\eta}_n^T \mathbf{X}_k - \boldsymbol{\eta}_n^T \mathbf{X}_l|, \bar{a}_{k.}(\boldsymbol{\eta}_n) = \frac{1}{n} \sum_{l=1}^n a_{kl}(\boldsymbol{\eta}_n), \bar{a}_{.l}(\boldsymbol{\eta}_n) = \frac{1}{n} \sum_{k=1}^n a_{kl}(\boldsymbol{\eta}_n), \\ \bar{a}_{..}(\boldsymbol{\eta}_n) &= \frac{1}{n^2} \sum_{k,l=1}^n a_{kl}(\boldsymbol{\eta}_n), A_{kl}(\boldsymbol{\eta}_n) = a_{kl}(\boldsymbol{\eta}_n) - \bar{a}_{k.}(\boldsymbol{\eta}_n) - \bar{a}_{.l}(\boldsymbol{\eta}_n) + \bar{a}_{..}(\boldsymbol{\eta}_n). \end{aligned}$$

$b_{kl} = |Y_k - Y_l|$ and $B_{kl} = b_{kl} - \bar{b}_{k.} - \bar{b}_{.l} + \bar{b}_{..}$. Then we have

$$\mathcal{V}_n^2(\boldsymbol{\eta}_n^T \mathbf{X}, \mathbf{Y}) = \frac{1}{n^2} \sum_{k,l=1}^n A_{kl}(\boldsymbol{\eta}_n) B_{kl}$$

*Wenhui Sheng is Graduate Student, Department of Statistics, University of Georgia, Athens, GA 30602 (Email: wsheng@uga.edu). Xiangrong Yin is Professor, Department of Statistics, University of Kentucky, Lexington, KY 40536 (E-mail: yinxiangrong@uky.edu).

and similarly,

$$\mathcal{V}_n^2(\boldsymbol{\eta}^T \mathbf{X}, Y) = \frac{1}{n^2} \sum_{k,l=1}^n A_{kl}(\boldsymbol{\eta}) B_{kl}.$$

Because $\boldsymbol{\eta}_n \rightarrow \boldsymbol{\eta}$ in probability, let $\boldsymbol{\eta}_n = \boldsymbol{\eta} + \boldsymbol{\varepsilon}_n$, then for any $\epsilon > 0$, $\|\boldsymbol{\varepsilon}_n\| < \epsilon$, when $n \rightarrow \infty$, here $\|\cdot\|$ is the Frobenius norm. Hence, by the condition on X and Y , we have that for a positive constant C_x , and large n , $|A_{kl}(\boldsymbol{\eta}_n) - A_{kl}(\boldsymbol{\eta})| \leq \epsilon C_x$.

Therefore,

$$\begin{aligned} \mathcal{V}_n^2(\boldsymbol{\eta}_n^T \mathbf{X}, Y) - \mathcal{V}_n^2(\boldsymbol{\eta}^T \mathbf{X}, Y) &= \frac{1}{n^2} \sum_{k,l=1}^n (A_{kl}(\boldsymbol{\eta}_n) - A_{kl}(\boldsymbol{\eta})) B_{kl}, \\ |\mathcal{V}_n^2(\boldsymbol{\eta}_n^T \mathbf{X}, Y) - \mathcal{V}_n^2(\boldsymbol{\eta}^T \mathbf{X}, Y)| &\leq \frac{1}{n^2} \sum_{k,l=1}^n |A_{kl}(\boldsymbol{\eta}_n) - A_{kl}(\boldsymbol{\eta})| |B_{kl}| \\ &\leq \epsilon C_x \frac{1}{n^2} \sum_{k,l=1}^n |B_{kl}| \\ &\leq \epsilon C_x C_y, \end{aligned}$$

where constant $C_y > 0$, which is due to condition $E|Y| < \infty$. Hence the conclusion follows. $\square$

A.2 Proof of Proposition 3

Proof of Proposition 3: Without loss of generality, we assume $Q = I_d$. Suppose $\boldsymbol{\eta}_n$ is not a consistent estimator of $\mathcal{S}_{Y|X}$, then there exists a subsequence, still to be indexed by n , and an $\boldsymbol{\eta}^*$ satisfying $\boldsymbol{\eta}^{*T} \hat{\Sigma}_{\mathbf{X}} \boldsymbol{\eta}^* = I_d$ such that $\boldsymbol{\eta}_n \xrightarrow{P} \boldsymbol{\eta}^*$ but $\text{Span}(\boldsymbol{\eta}^*) \neq \text{Span}(\boldsymbol{\eta})$.

By Lemma A, we have $\mathcal{V}_n^2(\boldsymbol{\eta}_n^T \mathbf{X}, \mathbf{Y}) - \mathcal{V}_n^2(\boldsymbol{\eta}^{*T} \mathbf{X}, \mathbf{Y}) \xrightarrow{P} 0$ and by Theorem 2 in Székely, Rizzo and Bakirov (2007), we have $\mathcal{V}_n^2(\boldsymbol{\eta}^{*T} \mathbf{X}, \mathbf{Y}) \xrightarrow{a.s.} \mathcal{V}^2(\boldsymbol{\eta}^{*T} X, Y)$, therefore $\mathcal{V}_n^2(\boldsymbol{\eta}_n^T \mathbf{X}, \mathbf{Y}) \xrightarrow{P} \mathcal{V}^2(\boldsymbol{\eta}^{*T} X, Y)$.

On the other hand, because $\boldsymbol{\eta}_n = \arg \max_{\boldsymbol{\beta}^T \hat{\Sigma}_{\mathbf{X}} \boldsymbol{\beta} = I_d} \mathcal{V}_n^2(\boldsymbol{\beta}^T \mathbf{X}, \mathbf{Y})$, we have that $\mathcal{V}_n^2(\boldsymbol{\eta}_n^T \mathbf{X}, \mathbf{Y}) \geq \mathcal{V}_n^2(\boldsymbol{\eta}^T \mathbf{X}, \mathbf{Y})$. If we take the limit on both sides of the above inequality, we get $\mathcal{V}^2(\boldsymbol{\eta}^{*T} X, Y) \geq \mathcal{V}^2(\boldsymbol{\eta}^T X, Y)$, however, we have proved that under the assumption $P_{\boldsymbol{\eta}(\Sigma_X)}^T X \perp\!\!\!\perp Q_{\boldsymbol{\eta}(\Sigma_X)}^T X$, $\boldsymbol{\eta} = \arg \max_{\boldsymbol{\beta}^T \Sigma_{\mathbf{X}} \boldsymbol{\beta} = I_d} \mathcal{V}^2(\boldsymbol{\beta}^T \mathbf{X}, \mathbf{Y})$, and we also assume that the central subspace is unique, therefore $\mathcal{V}^2(\boldsymbol{\eta}^{*T} X, Y) \geq \mathcal{V}^2(\boldsymbol{\eta}^T X, Y)$ conflicts with the above assumption, so $\boldsymbol{\eta}_n$ is a consistent estimator of a basis of the central subspace. $\square$

B Proof of $\sqrt{n}$ -consistency

In order to prove the $\sqrt{n}$ -consistency of $\text{vec}(\boldsymbol{\eta}_n)$ in the Proposition 4 in section 2.5 of the paper, we reconstruct the optimization problem by using the Lagrange multiplier technique, and first we introduce the following notations, conditions and we also give a new definition.

B.1 Notations and Conditions

For a random sample $(\mathbf{X}, \mathbf{Y}) = \{(X_k, Y_k) : k = 1, \dots, n\}$ from the joint distribution of random vectors X in $\mathbb{R}^p$ and Y in $\mathbb{R}$.

Let $\mathcal{L}(\zeta) = \mathcal{V}^2(\beta^T X, Y) + \lambda^T(\text{vec}(\beta^T \Sigma_X \beta) - \text{vec}(I_d))$ and $\mathcal{L}_n(\zeta) = \mathcal{V}_n^2(\beta^T \mathbf{X}, \mathbf{Y}) + \lambda^T(\text{vec}(\beta^T \hat{\Sigma}_{\mathbf{X}} \beta) - \text{vec}(I_d))$. Here $\zeta = \begin{pmatrix} \text{vec}(\beta) \\ \lambda \end{pmatrix} \in \mathbb{R}^{pd+d^2}$, $\beta \in \mathbb{R}^{p \times d}$, $\lambda \in \mathbb{R}^{d^2}$, Σ_X is the covariance matrix of X , and $\hat{\Sigma}_{\mathbf{X}}$ is the sample estimate for Σ_X . Let $\eta_n = \arg \max_{\beta^T \hat{\Sigma}_{\mathbf{X}} \beta = I_d} \mathcal{V}_n^2(\beta^T \mathbf{X}, \mathbf{Y})$, then there exists a λ_n such that $\begin{pmatrix} \text{vec}(\eta_n) \\ \lambda_n \end{pmatrix}$ is a stationary point for $\mathcal{L}_n(\zeta)$. Let $\theta_n = \begin{pmatrix} \text{vec}(\eta_n) \\ \lambda_n \end{pmatrix}$, then $\mathcal{L}'_n(\theta_n) = 0$. Let η to be a basis of CS, then under the assumption $P_{\eta(\Sigma_X)}^T \mathbf{X} \perp Q_{\eta(\Sigma_X)}^T \mathbf{X}$, there exists a rotation matrix $Q : Q^T Q = I_d$, such that $\eta Q = \arg \max_{\beta^T \Sigma_X \beta = I_d} \mathcal{V}^2(\beta^T X, Y)$. Without loss of generality, we assume $Q = I_d$ here, therefore there exists a λ_0 such that $\begin{pmatrix} \text{vec}(\eta) \\ \lambda_0 \end{pmatrix}$ is a stationary point for $\mathcal{L}(\zeta)$. Let $\theta = \begin{pmatrix} \text{vec}(\eta) \\ \lambda_0 \end{pmatrix}$.

In the proof, we need to take derivatives of $\mathcal{V}^2(\eta^T X, Y)$ and $\mathcal{V}_n^2(\eta^T \mathbf{X}, \mathbf{Y})$ with respect to $\text{vec}(\eta)$, so for the simplicity of notation, when we consider the derivatives of $\mathcal{V}^2(\eta^T X, Y)$ and $\mathcal{V}_n^2(\eta^T \mathbf{X}, \mathbf{Y})$, we use $\mathcal{V}(\eta)$ and $\mathcal{V}_n(\eta)$ to denote $\mathcal{V}^2(\eta^T X, Y)$ and $\mathcal{V}_n^2(\eta^T \mathbf{X}, \mathbf{Y})$ respectively.

Here are additional notations, which will be used later in the following proof. $I_{(d,d)}$ is the vec-permutation matrix. I_m is a identity matrix with rank m , and $I_m(:, i)$ denotes the i th column of I_m . $\mathbf{A} \otimes \mathbf{B}$ denotes Kronecker product between matrix $\mathbf{A}$ and $\mathbf{B}$. $\text{vec}(\cdot)$ is a vec operator.

Furthermore, we give the following definition and assumptions.

Definition 1 Let $\Delta(\eta) = \{\alpha : \|\alpha - \eta\| \leq c\}$, where α is a $p \times d$ matrix and $\alpha^T \Sigma_X \alpha = I_d$, c is a fixed small constant, $\|\cdot\|$ is the Frobenius norm. We define an indicator function

$$\rho(X, X') = \begin{cases} 0 & \text{if } |\alpha^T (X - X')| \leq \epsilon_0, \text{ for } \alpha \in \Delta(\eta) \\ 1 & \text{if } |\alpha^T (X - X')| > \epsilon_0, \text{ for } \alpha \in \Delta(\eta) \end{cases}$$

where X' is an i.i.d. copy of X and ϵ_0 is a small number. We define the second and third derivative of $\mathcal{V}(\eta)$ with respect to $\text{vec}(\eta)$ as $\mathcal{V}''(\eta)\rho(X, X')$ and $\mathcal{V}'''(\eta)\rho(X, X')$. For the simplicity of notation, we will still use $\mathcal{V}''(\eta)$ and $\mathcal{V}'''(\eta)$ to denote $\mathcal{V}''(\eta)\rho(X, X')$ and $\mathcal{V}'''(\eta)\rho(X, X')$, respectively.

The reason we use this definition is that under definition 1, the second and third derivative of $\mathcal{V}(\eta)$ and $\mathcal{V}_n(\eta)$ are bounded, near the neighborhood of the central subspace.

Assumption A: $\text{Var} [\phi^{(1)}((X_1, Y_1), (X_2, Y_2))]$, $\text{Var} [\phi^{(2)}((X_1, Y_1), (X_2, Y_2), (X_3, Y_3))]$, $\text{Var} [\phi^{(3)}((X_1, Y_1), (X_2, Y_2), (X_3, Y_3), (X_4, Y_4))]$, $\text{Var} [\phi^{(4)}((X_1, Y_1))]$,

$Var [\phi^{(5)}((X_1, Y_1), (X_2, Y_2))], Var [\phi^{(6)}((X_1, Y_1))], Var [\phi^{(7)}((X_1, Y_1), (X_2, Y_2))],$
 $Var [\phi^{(8)}((X_1, Y_1))]$ are all $< \infty$. Here

$$\begin{aligned}
\phi^{(1)}((X_1, Y_1), (X_2, Y_2)) &= \frac{(I_d \otimes (X_1 - X_2))(I_d \otimes (X_1 - X_2)^T)vec(\boldsymbol{\eta})}{|(I_d \otimes (X_1 - X_2)^T)vec(\boldsymbol{\eta})|} |Y_1 - Y_2|, \\
\phi^{(2)}((X_1, Y_1), (X_2, Y_2), (X_3, Y_3)) \\
&= \frac{1}{6} \sum_{P_1} \frac{(I_d \otimes (X_{i_1} - X_{i_2}))(I_d \otimes (X_{i_1} - X_{i_2})^T)vec(\boldsymbol{\eta})}{|(I_d \otimes (X_{i_1} - X_{i_2})^T)vec(\boldsymbol{\eta})|} |Y_{i_1} - Y_{i_3}|, \\
\phi^{(3)}((X_1, Y_1), (X_2, Y_2), (X_3, Y_3), (X_4, Y_4)) \\
&= \frac{1}{24} \sum_{P_2} \frac{(I_d \otimes (X_{i_1} - X_{i_2}))(I_d \otimes (X_{i_1} - X_{i_2})^T)vec(\boldsymbol{\eta})}{|(I_d \otimes (X_{i_1} - X_{i_2})^T)vec(\boldsymbol{\eta})|} |Y_{i_3} - Y_{i_4}|, \\
\phi^{(4)}((X_1, Y_1)) &= (I_d \otimes X_1 X_1^T \boldsymbol{\eta})(I_{d^2} + I_{d,d}^T) \lambda_0, \\
\phi^{(5)}((X_1, Y_1), (X_2, Y_2)) &= \frac{1}{2} (I_d \otimes (X_1 X_2^T + X_2 X_1^T) \boldsymbol{\eta})(I_{d^2} + I_{d,d}^T) \lambda_0, \\
\phi^{(6)}((X_1, Y_1)) &= vec(\boldsymbol{\eta}^T X_1 X_1^T \boldsymbol{\eta}), \\
\phi^{(7)}((X_1, Y_1), (X_2, Y_2)) &= \frac{1}{2} vec(\boldsymbol{\eta}^T (X_1 X_2^T + X_2 X_1^T) \boldsymbol{\eta}), \\
\phi^{(8)}((X_1, Y_1)) &= vec(X_1 - EX_1)(X_1 - EX_1)^T
\end{aligned}$$

and $\sum_{P_1}$ denotes summation over the $3!$ permutations (i_1, i_2, i_3) of $(1, 2, 3)$, $\sum_{P_2}$ denotes summation over the $4!$ permutations (i_1, i_2, i_3, i_4) of $(1, 2, 3, 4)$.

Assumption B: $\begin{pmatrix} \mathcal{V}''(\boldsymbol{\eta}) + L & (I_d \otimes \Sigma_{\mathbf{X}} \boldsymbol{\eta})(I_{d^2} + I_{(d,d)}) \\ (I_{d^2} + I_{(d,d)}^T)(I_d \otimes \boldsymbol{\eta}^T \Sigma_{\mathbf{X}}) & 0 \end{pmatrix}$ is nonsingular.

Assumption A is needed for Proposition 4 in the main article and Lemma B in the next Section, which is similar to the assumed conditions of Theorem 6.1.6 (Lehmann, 1999, Ch. 6). This assumption is required by the asymptotic properties of U-statistics.

Assumption B is in the spirit of von Mises proposition (Serfling, 1980, Section 6.1). In this proposition, it claims that if the first nonvanishing term of Taylor expansion is the linear term, then the root-n consistency of the differentiable statistical function can be achieved. In our case, we assume the corresponding matrix is nonsingular, which guarantees the root-n consistency. If the matrix is singular, then n or higher order consistency of some parts of our estimates can be proved.

B.2 Proof of Lemma B

In order to prove Proposition 4 in section 2.5 of the paper, we prove the following Lemma B first.

Lemma B Under assumption A,B and the assumptions in Proposition 4, then $\sqrt{n}(\boldsymbol{\theta}_n - \boldsymbol{\theta}) \xrightarrow{\mathcal{D}} N(0, V)$. The explicit expression for V is in the proof.

Proof of Lemma B: The Taylor expansion of $\mathcal{L}'_n(\theta_n)$ at θ is $0 = \mathcal{L}'_n(\theta_n) = \mathcal{L}'_n(\theta) + \mathcal{L}''_n(\theta)(\theta_n - \theta) + \mathcal{R}_1(\theta_n^*)$, where $\|\theta_n^* - \theta\| \leq \|\theta_n - \theta\|$, where $\|\cdot\|$ is the Frobenius norm and $\theta_n^* = \begin{pmatrix} \text{vec}(\boldsymbol{\eta}_n^*) \\ \lambda_n^* \end{pmatrix}$. Next, we will give explicit expressions of $\mathcal{L}'_n(\theta)$, $\mathcal{L}''_n(\theta)$ and $\mathcal{R}_1(\theta_n^*)$.

With simple calculation, $\mathcal{L}'_n(\theta) = \begin{pmatrix} \mathcal{V}'_n(\boldsymbol{\eta}) + (I_d \otimes \hat{\Sigma}_{\mathbf{X}}\boldsymbol{\eta})(I_{d^2} + I_{(d,d)})\lambda_0 \\ \text{vec}(\boldsymbol{\eta}^T \hat{\Sigma}_{\mathbf{X}}\boldsymbol{\eta}) - \text{vec}(I_d) \end{pmatrix}$

$$\mathcal{L}''_n(\theta) = \begin{pmatrix} \mathcal{V}''_n(\boldsymbol{\eta}) + \hat{L} & (I_d \otimes \hat{\Sigma}_{\mathbf{X}}\boldsymbol{\eta})(I_{d^2} + I_{(d,d)}) \\ (I_{d^2} + I_{(d,d)}^T)(I_d \otimes \boldsymbol{\eta}^T \hat{\Sigma}_{\mathbf{X}}) & 0 \end{pmatrix}.$$

where $\hat{L} = (\text{vec}(\hat{L}_{11}), \text{vec}(\hat{L}_{21}), \dots, \text{vec}(\hat{L}_{p1}), \dots, \text{vec}(\hat{L}_{1d}), \text{vec}(\hat{L}_{2d}), \dots, \text{vec}(\hat{L}_{pd}))^T$ and $\hat{L}_{ij} = \hat{\Sigma}_{\mathbf{X}}^T I_p(:, i) \lambda_0^T (I_{d^2} + I_{(d,d)}^T) (I_d(:, j) \otimes I_d)$. It is obvious that $\hat{L} \xrightarrow{a.s.} L$, where $L = (\text{vec}(L_{11}), \text{vec}(L_{21}), \dots, \text{vec}(L_{p1}), \dots, \text{vec}(L_{1d}), \text{vec}(L_{2d}), \dots, \text{vec}(L_{pd}))^T$ and $L_{ij} = \Sigma_X^T I_p(:, i) \lambda_0^T (I_{d^2} + I_{(d,d)}^T) (I_d(:, j) \otimes I_d)$. Here $i = 1, \dots, p$ and $j = 1, \dots, d$.

The remainder term $\mathcal{R}_1(\theta_n^*)$ involves the third derivative of $\mathcal{L}(\zeta)$ at θ_n^* . Let $T_n = \mathcal{L}'''_n(\theta_n^*)$, where T_n is a $(pd + d^2) \times (pd + d^2) \times (pd + d^2)$ array and each $T_n(j, :, :)$, $j = 1, \dots, pd + d^2$, is a $(pd + d^2) \times (pd + d^2)$ matrix. Therefore, the form of $\mathcal{R}_1(\theta_n^*)$ can be written as

$$\mathcal{R}_1(\theta_n^*) = \frac{1}{2} \begin{pmatrix} (\theta_n - \theta)^T T_n(1, :, :)(\theta_n - \theta) \\ (\theta_n - \theta)^T T_n(2, :, :)(\theta_n - \theta) \\ \vdots \\ (\theta_n - \theta)^T T_n(pd + d^2, :, :)(\theta_n - \theta) \end{pmatrix}.$$

Based on the above explicit expression of $\mathcal{L}'_n(\theta)$, $\mathcal{L}''_n(\theta)$ and $\mathcal{R}_1(\theta_n^*)$, the Taylor expansion of $\mathcal{L}'_n(\theta_n)$ at θ can be written as

$$\begin{aligned} 0 &= \begin{pmatrix} \mathcal{V}'_n(\boldsymbol{\eta}) + (I_d \otimes \hat{\Sigma}_{\mathbf{X}}\boldsymbol{\eta})(I_{d^2} + I_{(d,d)})\lambda_0 \\ \text{vec}(\boldsymbol{\eta}^T \hat{\Sigma}_{\mathbf{X}}\boldsymbol{\eta}) - \text{vec}(I_d) \end{pmatrix} \\ &+ \begin{pmatrix} \mathcal{V}''_n(\boldsymbol{\eta}) + \hat{L} & (I_d \otimes \hat{\Sigma}_{\mathbf{X}}\boldsymbol{\eta})(I_{d^2} + I_{(d,d)}) \\ (I_{d^2} + I_{(d,d)}^T)(I_d \otimes \boldsymbol{\eta}^T \hat{\Sigma}_{\mathbf{X}}) & 0 \end{pmatrix} \begin{pmatrix} \text{vec}(\boldsymbol{\eta}_n) - \text{vec}(\boldsymbol{\eta}) \\ \lambda_n - \lambda_0 \end{pmatrix} \\ &+ \frac{1}{2} \begin{pmatrix} (\theta_n - \theta)^T T_n(1, :, :)(\theta_n - \theta) \\ (\theta_n - \theta)^T T_n(2, :, :)(\theta_n - \theta) \\ \vdots \\ (\theta_n - \theta)^T T_n(pd + d^2, :, :)(\theta_n - \theta) \end{pmatrix}. \end{aligned}$$

From the above Taylor expansion of $\mathcal{L}'_n(\theta_n)$ at θ , we get

$$\begin{aligned} & - \begin{pmatrix} \mathcal{V}''_n(\boldsymbol{\eta}) + \hat{L} & (I_d \otimes \hat{\Sigma}_{\mathbf{X}}\boldsymbol{\eta})(I_{d^2} + I_{(d,d)}) \\ (I_{d^2} + I_{(d,d)}^T)(I_d \otimes \boldsymbol{\eta}^T \hat{\Sigma}_{\mathbf{X}}) & 0 \end{pmatrix}^{-1} \times \\ & \sqrt{n} \begin{pmatrix} \mathcal{V}'_n(\boldsymbol{\eta}) + (I_d \otimes \hat{\Sigma}_{\mathbf{X}}\boldsymbol{\eta})(I_{d^2} + I_{(d,d)})\lambda_0 \\ \text{vec}(\boldsymbol{\eta}^T \hat{\Sigma}_{\mathbf{X}}\boldsymbol{\eta}) - \text{vec}(I_d) \end{pmatrix} = \\ & [I_{pd+d^2} + \frac{1}{2} \begin{pmatrix} \mathcal{V}''_n(\boldsymbol{\eta}) + \hat{L} & (I_d \otimes \hat{\Sigma}_{\mathbf{X}}\boldsymbol{\eta})(I_{d^2} + I_{(d,d)}) \\ (I_{d^2} + I_{(d,d)}^T)(I_d \otimes \boldsymbol{\eta}^T \hat{\Sigma}_{\mathbf{X}}) & 0 \end{pmatrix}^{-1} \times \end{aligned}$$

$$\begin{pmatrix} (\boldsymbol{\theta}_n - \boldsymbol{\theta})^T T_n(1, :, :) \\ (\boldsymbol{\theta}_n - \boldsymbol{\theta})^T T_n(2, :, :) \\ \vdots \\ (\boldsymbol{\theta}_n - \boldsymbol{\theta})^T T_n(pd + d^2, :, :) \end{pmatrix}] \sqrt{n}(\boldsymbol{\theta}_n - \boldsymbol{\theta}).$$

Next, we will prove two parts:

$$\text{Part 1: } \begin{pmatrix} \mathcal{V}_n''(\boldsymbol{\eta}) + \hat{L} & (I_d \otimes \hat{\Sigma}_{\mathbf{X}}\boldsymbol{\eta})(I_{d^2} + I_{(d,d)}) \\ (I_{d^2} + I_{(d,d)}^T)(I_d \otimes \boldsymbol{\eta}^T \hat{\Sigma}_{\mathbf{X}}) & 0 \end{pmatrix}^{-1} \times \\ \sqrt{n} \begin{pmatrix} \mathcal{V}_n'(\boldsymbol{\eta}) + (I_d \otimes \hat{\Sigma}_{\mathbf{X}}\boldsymbol{\eta})(I_{d^2} + I_{(d,d)})\lambda_0 \\ \text{vec}(\boldsymbol{\eta}^T \hat{\Sigma}_{\mathbf{X}}\boldsymbol{\eta}) - \text{vec}(I_d) \end{pmatrix} \longrightarrow N(0, \mathbf{V}).$$

$$\text{Part 2: } \sqrt{n}(\boldsymbol{\theta}_n - \boldsymbol{\theta}) \stackrel{\mathcal{D}}{=} \\ [I_{pd+d^2} + \frac{1}{2} \begin{pmatrix} \mathcal{V}_n''(\boldsymbol{\eta}) + \hat{L} & (I_d \otimes \hat{\Sigma}_{\mathbf{X}}\boldsymbol{\eta})(I_{d^2} + I_{(d,d)}) \\ (I_{d^2} + I_{(d,d)}^T)(I_d \otimes \boldsymbol{\eta}^T \hat{\Sigma}_{\mathbf{X}}) & 0 \end{pmatrix}^{-1} \times \\ \begin{pmatrix} (\boldsymbol{\theta}_n - \boldsymbol{\theta})^T T_n(1, :, :) \\ (\boldsymbol{\theta}_n - \boldsymbol{\theta})^T T_n(2, :, :) \\ \vdots \\ (\boldsymbol{\theta}_n - \boldsymbol{\theta})^T T_n(pd + d^2, :, :) \end{pmatrix}] \sqrt{n}(\boldsymbol{\theta}_n - \boldsymbol{\theta}).$$

Proof of part 1 : We will show that both $\mathcal{V}_n'(\boldsymbol{\eta}) + (I_d \otimes \hat{\Sigma}_{\mathbf{X}}\boldsymbol{\eta})(I_{d^2} + I_{(d,d)})\lambda_0$ and $\text{vec}(\boldsymbol{\eta}^T \hat{\Sigma}_{\mathbf{X}}\boldsymbol{\eta}) - \text{vec}(I_d)$ are linear combinations of U-statistics and the asymptotic distribution can be achieved by the asymptotic property of U-statistics.

Based on the equation (2.18) in Székely, Rizzo and Bakirov (2007), $\mathcal{V}_n(\boldsymbol{\eta}) = S_1(\boldsymbol{\eta}) + S_2(\boldsymbol{\eta}) - 2S_3(\boldsymbol{\eta})$, where

$$S_1(\boldsymbol{\eta}) = \frac{1}{n^2} \sum_{k,l=1}^n |\boldsymbol{\eta}^T(X_k - X_l)| |Y_k - Y_l|, \\ S_2(\boldsymbol{\eta}) = \frac{1}{n^2} \sum_{k,l=1}^n |\boldsymbol{\eta}^T(X_k - X_l)| \frac{1}{n^2} \sum_{k,l=1}^n |Y_k - Y_l|, \\ S_3(\boldsymbol{\eta}) = \frac{1}{n^3} \sum_{k=1}^n \sum_{l,m=1}^n |\boldsymbol{\eta}^T(X_k - X_l)| |Y_k - Y_m|.$$

Therefore, $\mathcal{V}_n'(\boldsymbol{\eta}) + (I_d \otimes \hat{\Sigma}_{\mathbf{X}}\boldsymbol{\eta})(I_{d^2} + I_{(d,d)})\lambda_0 = S_1'(\boldsymbol{\eta}) + S_2'(\boldsymbol{\eta}) - 2S_3'(\boldsymbol{\eta}) + (I_d \otimes \hat{\Sigma}_{\mathbf{X}}\boldsymbol{\eta})(I_{d^2} +$

$I_{(d,d)}\lambda_0$, where

$$\begin{aligned}
S'_1(\boldsymbol{\eta}) &= \frac{1}{n^2} \sum_{k,l=1}^n \frac{(I_d \otimes (X_k - X_l))(I_d \otimes (X_k - X_l)^T) \text{vec}(\boldsymbol{\eta})}{|(I_d \otimes (X_k - X_l)^T) \text{vec}(\boldsymbol{\eta})|} |Y_k - Y_l|, \\
S'_2(\boldsymbol{\eta}) &= \frac{1}{n^2} \sum_{k,l=1}^n \frac{(I_d \otimes (X_k - X_l))(I_d \otimes (X_k - X_l)^T) \text{vec}(\boldsymbol{\eta})}{|(I_d \otimes (X_k - X_l)^T) \text{vec}(\boldsymbol{\eta})|} \frac{1}{n^2} \sum_{k,l=1}^n |Y_k - Y_l|, \\
S'_3(\boldsymbol{\eta}) &= \frac{1}{n^3} \sum_{k=1}^n \sum_{l,m=1}^n \frac{(I_d \otimes (X_k - X_l))(I_d \otimes (X_k - X_l)^T) \text{vec}(\boldsymbol{\eta})}{|(I_d \otimes (X_k - X_l)^T) \text{vec}(\boldsymbol{\eta})|} |Y_k - Y_m|, \\
\hat{\Sigma}_{\mathbf{X}} &= \frac{1}{n} \sum_{i=1}^n X_i X_i^T - \bar{\mathbf{X}} \bar{\mathbf{X}}^T = \frac{1}{n} \sum_{i=1}^n X_i X_i^T - \frac{1}{n^2} \sum_{i,j=1}^n X_i X_j^T.
\end{aligned}$$

Here $S'_1(\boldsymbol{\eta})$, $S'_2(\boldsymbol{\eta})$, $S'_3(\boldsymbol{\eta})$ and $\hat{\Sigma}_{\mathbf{X}}$ are V-statistics, which can be written as linear combinations of U-statistics. Let

$$\begin{aligned}
\mathbf{U}_{1n} &= \binom{n}{2}^{-1} \sum_{1 \leq k < l \leq n} \frac{(I_d \otimes (X_k - X_l))(I_d \otimes (X_k - X_l)^T) \text{vec}(\boldsymbol{\eta})}{|(I_d \otimes (X_k - X_l)^T) \text{vec}(\boldsymbol{\eta})|} |Y_k - Y_l|, \\
\mathbf{U}_{2n} &= \binom{n}{3}^{-1} \sum_{1 \leq k < l < m \leq n} \left\{ \frac{1}{6} \sum_{P_1} \frac{(I_d \otimes (X_{i_1} - X_{i_2}))(I_d \otimes (X_{i_1} - X_{i_2})^T) \text{vec}(\boldsymbol{\eta})}{|(I_d \otimes (X_{i_1} - X_{i_2})^T) \text{vec}(\boldsymbol{\eta})|} |Y_{i_1} - Y_{i_3}| \right\}, \\
\mathbf{U}_{3n} &= \binom{n}{4}^{-1} \sum_{1 \leq k_1 < l_1 < k_2 < l_2 \leq n} \left\{ \frac{1}{24} \sum_{P_2} \frac{(I_d \otimes (X_{i_1} - X_{i_2}))(I_d \otimes (X_{i_1} - X_{i_2})^T) \text{vec}(\boldsymbol{\eta})}{|(I_d \otimes (X_{i_1} - X_{i_2})^T) \text{vec}(\boldsymbol{\eta})|} |Y_{i_3} - Y_{i_4}| \right\}, \\
\mathbf{U}_{4n} &= \frac{1}{n} \sum_{i=1}^n (I_d \otimes X_i X_i^T \boldsymbol{\eta}) (I_{d^2} + I_{d,d}^T) \lambda_0, \\
\mathbf{U}_{5n} &= \binom{n}{2}^{-1} \sum_{i < j} \frac{1}{2} (I_d \otimes (X_i X_j^T + X_j X_i^T) \boldsymbol{\eta}) (I_{d^2} + I_{d,d}^T) \lambda_0.
\end{aligned}$$

Here $\sum_{P_1}$ denotes summation over 3! permutations (i_1, i_2, i_3) of (k, l, m) , and $\sum_{P_2}$ denotes summation over 4! permutations (i_1, i_2, i_3, i_4) of (k_1, l_1, k_2, l_2) .

Through some tedious calculations, we can get $\mathcal{V}'_n(\boldsymbol{\eta}) + (I_d \otimes \hat{\Sigma}_{\mathbf{X}} \boldsymbol{\eta}) (I_{d^2} + I_{d,d}^T) \lambda_0 = \frac{(n-1)(n^2-2n+2)}{n^3} \mathbf{U}_{1n} - \frac{2(n-1)(n-2)^2}{n^3} \mathbf{U}_{2n} + \frac{(n-1)(n-2)(n-3)}{n^3} \mathbf{U}_{3n} + \frac{n-1}{n} \mathbf{U}_{4n} - \frac{n-1}{n} \mathbf{U}_{5n}$.

$\text{vec}(\boldsymbol{\eta}^T \hat{\Sigma}_{\mathbf{X}} \boldsymbol{\eta})$ is also a linear combination of U-statistics, let

$$\begin{aligned}
\mathbf{U}_{6n} &= \frac{1}{n} \sum_{i=1}^n \text{vec}(\boldsymbol{\eta}^T X_i X_i^T \boldsymbol{\eta}) \\
\mathbf{U}_{7n} &= \binom{n}{2}^{-1} \sum_{i < j} \frac{1}{2} \text{vec}(\boldsymbol{\eta}^T (X_i X_j^T + X_j X_i^T) \boldsymbol{\eta}),
\end{aligned}$$

then $\text{vec}(\boldsymbol{\eta}^T \hat{\Sigma}_{\mathbf{X}} \boldsymbol{\eta}) = \frac{n-1}{n} \mathbf{U}_{6n} - \frac{n-1}{n} \mathbf{U}_{7n}$.

let

$$\begin{aligned}
\mu_1 &= E \frac{(I_d \otimes (X - X'))(I_d \otimes (X - X')^T) \text{vec}(\boldsymbol{\eta})}{|(I_d \otimes (X - X')^T) \text{vec}(\boldsymbol{\eta})|} |Y - Y'|, \\
\mu_2 &= E \frac{(I_d \otimes (X - X'))(I_d \otimes (X - X')^T) \text{vec}(\boldsymbol{\eta})}{|(I_d \otimes (X - X')^T) \text{vec}(\boldsymbol{\eta})|} |Y - Y''|, \\
\mu_3 &= E \frac{(I_d \otimes (X - X'))(I_d \otimes (X - X')^T) \text{vec}(\boldsymbol{\eta})}{|(I_d \otimes (X - X')^T) \text{vec}(\boldsymbol{\eta})|} |Y'' - Y'''|, \\
\mu_4 &= E(I_d \otimes XX^T \boldsymbol{\eta})(I_{d^2} + I_{(d,d)}^T) \lambda_0, \\
\mu_5 &= (I_d \otimes (EX)(EX)^T \boldsymbol{\eta})(I_{d^2} + I_{(d,d)}^T) \lambda_0, \\
\mu_6 &= \text{vec}(\boldsymbol{\eta}^T (EXX^T) \boldsymbol{\eta}), \\
\mu_7 &= \text{vec}(\boldsymbol{\eta}^T (EX)(EX)^T \boldsymbol{\eta}).
\end{aligned}$$

Here $(\mathbf{X}, Y), (\mathbf{X}', Y'), (\mathbf{X}'', Y''), (\mathbf{X}''', Y''')$ are i.i.d copies.

According to Theorem 6.1.6 (Lehmann, 1999, Ch.6),

$$\sqrt{n} \begin{pmatrix} \mathbf{U}_{1n} - \mu_1 \\ \mathbf{U}_{2n} - \mu_2 \\ \mathbf{U}_{3n} - \mu_3 \\ \mathbf{U}_{4n} - \mu_4 \\ \mathbf{U}_{5n} - \mu_5 \\ \mathbf{U}_{6n} - \mu_6 \\ \mathbf{U}_{7n} - \mu_7 \end{pmatrix} \xrightarrow{\mathcal{D}} N(0, \Sigma),$$

$$\text{where } \Sigma = \begin{pmatrix} \Sigma_{11} & \Sigma_{12} & \Sigma_{13} & \Sigma_{14} & \Sigma_{15} & \Sigma_{16} & \Sigma_{17} \\ \cdot & \Sigma_{22} & \Sigma_{23} & \Sigma_{24} & \Sigma_{25} & \Sigma_{26} & \Sigma_{27} \\ \cdot & \cdot & \Sigma_{33} & \Sigma_{34} & \Sigma_{35} & \Sigma_{36} & \Sigma_{37} \\ \cdot & \cdot & \cdot & \Sigma_{44} & \Sigma_{45} & \Sigma_{46} & \Sigma_{47} \\ \cdot & \cdot & \cdot & \cdot & \Sigma_{55} & \Sigma_{56} & \Sigma_{57} \\ \cdot & \cdot & \cdot & \cdot & \cdot & \Sigma_{66} & \Sigma_{67} \\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \Sigma_{77} \end{pmatrix},$$

$\Sigma_{ij} = a_{ij} \text{Cov}(\phi^{(i)}, \phi^{(j)})$. Here a_{ij} is a constant, which equals to the number of inputs of $\phi^{(i)}$ multiplies the number of inputs of $\phi^{(j)}$.

Let $\mathbf{B} = \begin{pmatrix} I_{pd} & (-2)I_{pd} & I_{pd} & I_{pd} & I_{pd} & \mathbf{0} & \mathbf{0} \\ \mathbf{0}^T & \mathbf{0}^T & \mathbf{0}^T & \mathbf{0}^T & \mathbf{0}^T & I_{d^2 \times d^2} & -I_{d^2 \times d^2} \end{pmatrix}$, where $\mathbf{0}$ is a $pd \times d^2$ zero matrix, then

$$\sqrt{n}\mathbf{B} \begin{pmatrix} \mathbf{U}_{1n} - \mu_1 \\ \mathbf{U}_{2n} - \mu_2 \\ \mathbf{U}_{3n} - \mu_3 \\ \mathbf{U}_{4n} - \mu_4 \\ \mathbf{U}_{5n} - \mu_5 \\ \mathbf{U}_{6n} - \mu_6 \\ \mathbf{U}_{7n} - \mu_7 \end{pmatrix} = \sqrt{n} \begin{pmatrix} \mathbf{U}_{1n} - 2\mathbf{U}_{2n} + \mathbf{U}_{3n} + \mathbf{U}_{4n} - \mathbf{U}_{5n} \\ \mathbf{U}_{6n} - \mathbf{U}_{7n} - \text{vec}(I_d) \end{pmatrix}.$$

Note that

$$\begin{aligned} & \sqrt{n} \begin{pmatrix} \mathcal{V}'_n(\boldsymbol{\eta}) + (I_d \otimes \hat{\Sigma}_{\mathbf{X}}\boldsymbol{\eta})(I_{d^2} + I_{(d,d)})\lambda_0 \\ \text{vec}(\boldsymbol{\eta}^T \hat{\Sigma}_{\mathbf{X}}\boldsymbol{\eta}) - \text{vec}(I_d) \end{pmatrix} = \\ & \sqrt{n} \begin{pmatrix} \frac{(n-1)(n^2-2n+2)}{n^3}\mathbf{U}_{1n} - \frac{2(n-1)(n-2)^2}{n^3}\mathbf{U}_{2n} + \frac{(n-1)(n-2)(n-3)}{n^3}\mathbf{U}_{3n} + \frac{n-1}{n}\mathbf{U}_{4n} - \frac{n-1}{n}\mathbf{U}_{5n} \\ \frac{n-1}{n}\mathbf{U}_{6n} - \frac{n-1}{n}\mathbf{U}_{7n} - \text{vec}(I_d) \end{pmatrix}, \\ & \text{under assumption A,} \\ & \sqrt{n} \begin{pmatrix} \frac{(n-1)(n^2-2n+2)}{n^3}\mathbf{U}_{1n} - \frac{2(n-1)(n-2)^2}{n^3}\mathbf{U}_{2n} + \frac{(n-1)(n-2)(n-3)}{n^3}\mathbf{U}_{3n} + \frac{n-1}{n}\mathbf{U}_{4n} - \frac{n-1}{n}\mathbf{U}_{5n} \\ \frac{n-1}{n}\mathbf{U}_{6n} - \frac{n-1}{n}\mathbf{U}_{7n} - \text{vec}(I_d) \end{pmatrix} - \\ & \sqrt{n} \begin{pmatrix} \mathbf{U}_{1n} - 2\mathbf{U}_{2n} + \mathbf{U}_{3n} + \mathbf{U}_{4n} - \mathbf{U}_{5n} \\ \mathbf{U}_{6n} - \mathbf{U}_{7n} - \text{vec}(I_d) \end{pmatrix} \xrightarrow{P} 0, \text{ therefore by Slutsky's theorem,} \end{aligned}$$

$$\sqrt{n} \begin{pmatrix} \mathcal{V}'_n(\boldsymbol{\eta}) + (I_d \otimes \hat{\Sigma}_{\mathbf{X}}\boldsymbol{\eta})(I_{d^2} + I_{(d,d)})\lambda_0 \\ \text{vec}(\boldsymbol{\eta}^T \hat{\Sigma}_{\mathbf{X}}\boldsymbol{\eta}) - \text{vec}(I_d) \end{pmatrix} \stackrel{D}{=} \sqrt{n}\mathbf{B} \begin{pmatrix} \mathbf{U}_{1n} - \mu_1 \\ \mathbf{U}_{2n} - \mu_2 \\ \mathbf{U}_{3n} - \mu_3 \\ \mathbf{U}_{4n} - \mu_4 \\ \mathbf{U}_{5n} - \mu_5 \\ \mathbf{U}_{6n} - \mu_6 \\ \mathbf{U}_{7n} - \mu_7 \end{pmatrix}.$$

$$\text{Let } A_n = \begin{pmatrix} \mathcal{V}''_n(\boldsymbol{\eta}) + \hat{L} & (I_d \otimes \hat{\Sigma}_{\mathbf{X}}\boldsymbol{\eta})(I_{d^2} + I_{(d,d)}) \\ (I_{d^2} + I_{(d,d)}^T)(I_d \otimes \boldsymbol{\eta}^T \hat{\Sigma}_{\mathbf{X}}) & 0 \end{pmatrix}^{-1},$$

$$A = \begin{pmatrix} \mathcal{V}''(\boldsymbol{\eta}) + L & (I_d \otimes \Sigma_{\mathbf{X}}\boldsymbol{\eta})(I_{d^2} + I_{(d,d)}) \\ (I_{d^2} + I_{(d,d)}^T)(I_d \otimes \boldsymbol{\eta}^T \Sigma_{\mathbf{X}}) & 0 \end{pmatrix}^{-1},$$

under assumption B and our definition of second derivative of $\mathcal{V}_n(\boldsymbol{\eta})$, by SLLN of U-statistics,

$A_n \xrightarrow{a.s.} A$, therefore

$$\begin{aligned} & \begin{pmatrix} \mathcal{V}''_n(\boldsymbol{\eta}) + \hat{L} & (I_d \otimes \hat{\Sigma}_{\mathbf{X}}\boldsymbol{\eta})(I_{d^2} + I_{(d,d)}) \\ (I_{d^2} + I_{(d,d)}^T)(I_d \otimes \boldsymbol{\eta}^T \hat{\Sigma}_{\mathbf{X}}) & 0 \end{pmatrix}^{-1} \times \\ & \sqrt{n} \begin{pmatrix} \mathcal{V}'_n(\boldsymbol{\eta}) + (I_d \otimes \hat{\Sigma}_{\mathbf{X}}\boldsymbol{\eta})(I_{d^2} + I_{(d,d)})\lambda_0 \\ \text{vec}(\boldsymbol{\eta}^T \hat{\Sigma}_{\mathbf{X}}\boldsymbol{\eta}) - \text{vec}(I_d) \end{pmatrix} \stackrel{D}{=} \sqrt{n}\mathbf{A}\mathbf{B} \begin{pmatrix} \mathbf{U}_{1n} - \mu_1 \\ \mathbf{U}_{2n} - \mu_2 \\ \mathbf{U}_{3n} - \mu_3 \\ \mathbf{U}_{4n} - \mu_4 \\ \mathbf{U}_{5n} - \mu_5 \\ \mathbf{U}_{6n} - \mu_6 \\ \mathbf{U}_{7n} - \mu_7 \end{pmatrix} \longrightarrow N(0, \mathbf{V}), \text{ where } \mathbf{V} = \\ & \mathbf{A}\mathbf{B}\Sigma\mathbf{B}^T\mathbf{A}^T. \end{aligned}$$

Proof of part 2 :

Under assumption B and Definition 1,

$$\begin{aligned}
& [I_{pd+d^2} + \frac{1}{2} \begin{pmatrix} \mathcal{V}_n''(\boldsymbol{\eta}) + \hat{L} & (I_d \otimes \hat{\Sigma}_{\mathbf{X}} \boldsymbol{\eta})(I_{d^2} + I_{(d,d)}) \\ (I_{d^2} + I_{(d,d)}^T)(I_d \otimes \boldsymbol{\eta}^T \hat{\Sigma}_{\mathbf{X}}) & 0 \end{pmatrix}^{-1} \\
& \times \begin{pmatrix} (\boldsymbol{\theta}_n - \boldsymbol{\theta})^T T_n(1, :, :) \\ (\boldsymbol{\theta}_n - \boldsymbol{\theta})^T T_n(2, :, :) \\ \vdots \\ (\boldsymbol{\theta}_n - \boldsymbol{\theta})^T T_n(pd + d^2, :, :) \end{pmatrix}] \xrightarrow{P} I_{pd+d^2}, \text{ therefore by Slutsky's theorem,} \\
& \sqrt{n}(\boldsymbol{\theta}_n - \boldsymbol{\theta}) \stackrel{\mathcal{D}}{=} \\
& [I_{pd+d^2} + \frac{1}{2} \begin{pmatrix} \mathcal{V}_n''(\boldsymbol{\eta}) + \begin{pmatrix} \text{vec}^T(\hat{L}_{11}) \\ \vdots \\ \text{vec}^T(\hat{L}_{pd}) \end{pmatrix} & (I_d \otimes \hat{\Sigma}_{\mathbf{X}} \boldsymbol{\eta})(I_{d^2} + I_{(d,d)}) \\ (I_{d^2} + I_{(d,d)}^T)(I_d \otimes \boldsymbol{\eta}^T \hat{\Sigma}_{\mathbf{X}}) & 0 \end{pmatrix}^{-1} \times \\
& \begin{pmatrix} (\boldsymbol{\theta}_n - \boldsymbol{\theta})^T T_n(1, :, :) \\ (\boldsymbol{\theta}_n - \boldsymbol{\theta})^T T_n(2, :, :) \\ \vdots \\ (\boldsymbol{\theta}_n - \boldsymbol{\theta})^T T_n(pd + d^2, :, :) \end{pmatrix}] \sqrt{n}(\boldsymbol{\theta}_n - \boldsymbol{\theta}). \\
& \text{Therefore } \sqrt{n}(\boldsymbol{\theta}_n - \boldsymbol{\theta}) \xrightarrow{\mathcal{D}} N(0, \mathbf{V}), \text{ or in other words, } \boldsymbol{\theta}_n \text{ is } \sqrt{n}\text{-consistent estimation of } \boldsymbol{\theta}.
\end{aligned}$$

In the above proof, without loss of generality we assume that $\mathbf{Q} = I_d$. Note that with an orthogonal matrix $\mathbf{Q}$, $\mathcal{V}_n^2(\mathbf{Q}^T \boldsymbol{\beta}^T \mathbf{X}, \mathbf{Y}) = \mathcal{V}_n^2(\boldsymbol{\beta}^T \mathbf{X}, \mathbf{Y})$ and $\mathcal{V}^2(\mathbf{Q}^T \boldsymbol{\beta}^T X, Y) = \mathcal{V}^2(\boldsymbol{\beta}^T X, Y)$ (Székely and Rizzo 2009). If define $\boldsymbol{\eta}_{\mathbf{Q}} = \boldsymbol{\eta} \mathbf{Q}$, without assuming $\mathbf{Q} = I_d$, then Lemma B holds by using $\mathbf{V}(\boldsymbol{\eta}_{\mathbf{Q}})$ which is obtained by replacing every $\boldsymbol{\eta}$ in $\mathbf{V}$ with $\boldsymbol{\eta}_{\mathbf{Q}}$. (of course, then $\mathbf{V}(\boldsymbol{\eta}_{I_d}) = \mathbf{V}$ in the proof). $\square$

B.3 Proof of Proposition 4

Proof of Proposition 4: Let $G = (I_{pd}, 0)$ be a $pd \times (pd + d^2)$ matrix, where I_{pd} is a $pd \times pd$ identity matrix. Then $\text{vec}(\boldsymbol{\eta}_n) = G\boldsymbol{\theta}_n$ and $\text{vec}(\boldsymbol{\eta} \mathbf{Q}) = G\boldsymbol{\theta}$. By Lemma B, we have $\sqrt{n}(\text{vec}(\boldsymbol{\eta}_n) - \text{vec}(\boldsymbol{\eta} \mathbf{Q})) = \sqrt{n}G(\boldsymbol{\theta}_n - \boldsymbol{\theta}) \xrightarrow{\mathcal{D}} N(0, \mathbf{V}_{11}(\boldsymbol{\eta}_{\mathbf{Q}}))$, or in other word, $\sqrt{n}[\text{vec}(\boldsymbol{\eta}_n) - \text{vec}(\boldsymbol{\eta} \mathbf{Q})] \xrightarrow{\mathcal{D}} N(0, \mathbf{V}_{11}(\boldsymbol{\eta}_{\mathbf{Q}}))$, where $\mathbf{V}_{11}(\boldsymbol{\eta}_{\mathbf{Q}}) = G\mathbf{V}(\boldsymbol{\eta}_{\mathbf{Q}})G^T$. $\square$

C Proof of Corollary 1

Proof of Corollary 1. In our proof of the $\sqrt{n}$ -consistency, without loss of generality we assume that $\mathbf{Q} = I_d$. Note that with an orthogonal matrix $\mathbf{Q}$, $\mathcal{V}_n^2(\mathbf{Q}^T \boldsymbol{\beta}^T \mathbf{X}, \mathbf{Y}) = \mathcal{V}_n^2(\boldsymbol{\beta}^T \mathbf{X}, \mathbf{Y})$ and $\mathcal{V}^2(\mathbf{Q}^T \boldsymbol{\beta}^T X, Y) = \mathcal{V}^2(\boldsymbol{\beta}^T X, Y)$ (Székely and Rizzo 2009). If define $\boldsymbol{\eta}_{\mathbf{Q}} = \boldsymbol{\eta} \mathbf{Q}$, then Proposition 4 holds by using $\mathbf{V}_{11}(\boldsymbol{\eta}_{\mathbf{Q}})$ which is obtained by replacing every $\boldsymbol{\eta}$ in $\mathbf{V}_{11}$ with

$\eta_{\mathbf{Q}}$. (of course, then $\mathbf{V}_{11}(\eta_{I_d}) = \mathbf{V}_{11}$ in the proof).

To simplify the proof, here we still use η by assuming $\mathbf{Q} = I_d$. Let $A_{11} = \mathcal{V}_n''(\eta) + \hat{L}$, where $\hat{L}$ is given in the proof of Lemma B in the section B.2, $A_{12} = (I_d \otimes \hat{\Sigma}_{\mathbf{X}}\eta)(I_{d^2} + I_{(d,d)})$, $A_{21} = A_{12}^T$ and $A_{22} = 0$ (A_{22} is a $d^2 \times d^2$ zero matrix), $A_{22.1} = -A_{21}A_{11}^{-1}A_{12}$ then $D = -(A_{11}^{-1} + A_{11}^{-1}A_{12}A_{22.1}^{-1}A_{21}A_{11}^{-1})$ and $F = A_{11}^{-1}A_{12}A_{22.1}^{-1}$

Without loss of generality, we can expand η_n as

$$\eta_n = \eta + E_n\{A^*\} + o_p(n^{-1/2}),$$

and we can expand $\hat{\Sigma}$ as

$$\hat{\Sigma} = \Sigma + E_n\{\Sigma^*\} + o_p(n^{-1/2}).$$

Then we can get the asymptotic expansion of $\eta_n\eta_n^T\Sigma$ as

$$\eta_n\eta_n^T\hat{\Sigma} = \eta\eta^T\Sigma + E_n\{A^*\eta^T\Sigma\} + E_n\{\eta(A^*)^T\Sigma\} + E_n\{\eta\eta^T\Sigma^*\} + o_p(n^{-1/2}). \quad (1)$$

therefore, $\text{vec}(\eta_n\eta_n^T\hat{\Sigma}) - \text{vec}(\eta\eta^T\Sigma) = [(\Sigma\eta \otimes I_p) + (\Sigma \otimes \eta)I_{(d,p)}]\text{vec}(E_n\{A^*\}) + (I_p \otimes \eta\eta^T)\text{vec}(E_n\{\Sigma^*\}) + o_p(n^{-1/2})$, where $\text{vec}(E_n\{A^*\}) = D[\mathcal{V}_n'(\eta) + (I_d \otimes \hat{\Sigma}_{\mathbf{X}}\eta)(I_{d^2} + I_{(d,d)})\lambda_0] + F[\text{vec}(\eta^T\hat{\Sigma}_{\mathbf{X}}\eta) - \text{vec}(I_d)]$ and $\text{vec}(E_n\{\Sigma^*\}) = \frac{1}{n} \sum_{i=1}^n \text{vec}(X_i - \mu)(X_i - \mu)^T - \text{vec}(E(X - \mu)(X - \mu)^T)$, where $\mu = E(X)$.

Let $C = [(\Sigma\eta \otimes I_p) + (\Sigma \otimes \eta)I_{(d,p)}]$, $H = I_p \otimes \eta\eta^T$, $U_{8n} = \frac{1}{n} \sum_{i=1}^n \text{vec}(X_i - \mu)(X_i - \mu)^T$, $\mu_8 = \text{vec}(E(X - \mu)(X - \mu)^T)$, then $\text{vec}(\eta_n\eta_n^T\hat{\Sigma}) - \text{vec}(\eta\eta^T\Sigma) = \frac{(n-1)(n^2-2n+2)}{n^3}CDU_{1n} - \frac{2(n-1)(n-2)^2}{n^3}CDU_{2n} + \frac{(n-1)(n-2)(n-3)}{n^3}CDU_{3n} + \frac{n-1}{n}CDU_{4n} - \frac{n-1}{n}CDU_{5n} + \frac{n-1}{n}CFU_{6n} - \frac{n-1}{n}CFU_{7n} - CF\text{vec}(I_d) + HU_{8n} - H\text{vec}(E(X - \mu)(X - \mu)^T)$,

Let $U_{1n}^* = CDU_{1n}$, $U_{2n}^* = CDU_{2n}$, $U_{3n}^* = CDU_{3n}$, $U_{4n}^* = CDU_{4n}$, $U_{5n}^* = CDU_{5n}$, $U_{6n}^* = CFU_{6n}$, $U_{7n}^* = CFU_{7n}$ and $U_{8n}^* = HU_{8n}$; let $\mu_1^* = CD\mu_1$, $\mu_2^* = CD\mu_2$, $\mu_3^* = CD\mu_3$, $\mu_4^* = CD\mu_4$, $\mu_5^* = CD\mu_5$, $\mu_6^* = CF\mu_6$, $\mu_7^* = CF\mu_7$, $\mu_8^* = H\mu_8$, where U_{1n} , U_{2n} , U_{3n} , U_{4n} , U_{5n} , U_{6n} , U_{7n} , μ_1 , μ_2 , μ_3 , μ_4 , μ_5 , μ_6 , μ_7 are defined in the proof of Lemma B in the section B.2.

According to Theorem 6.1.6 (Lehmann, 1999, Ch.6),

$$\sqrt{n} \begin{pmatrix} \mathbf{U}_{1n}^* - \mu_1^* \\ \mathbf{U}_{2n}^* - \mu_2^* \\ \mathbf{U}_{3n}^* - \mu_3^* \\ \mathbf{U}_{4n}^* - \mu_4^* \\ \mathbf{U}_{5n}^* - \mu_5^* \\ \mathbf{U}_{6n}^* - \mu_6^* \\ \mathbf{U}_{7n}^* - \mu_7^* \\ \mathbf{U}_{8n}^* - \mu_8^* \end{pmatrix} \xrightarrow{\mathcal{D}} N(0, \Sigma^*),$$

$$\text{where } \Sigma^* = \begin{pmatrix} \Sigma_{11}^* & \Sigma_{12}^* & \Sigma_{13}^* & \Sigma_{14}^* & \Sigma_{15}^* & \Sigma_{16}^* & \Sigma_{17}^* & \Sigma_{18}^* \\ \cdot & \Sigma_{22}^* & \Sigma_{23}^* & \Sigma_{24}^* & \Sigma_{25}^* & \Sigma_{26}^* & \Sigma_{27}^* & \Sigma_{28}^* \\ \cdot & \cdot & \Sigma_{33}^* & \Sigma_{34}^* & \Sigma_{35}^* & \Sigma_{36}^* & \Sigma_{37}^* & \Sigma_{38}^* \\ \cdot & \cdot & \cdot & \Sigma_{44}^* & \Sigma_{45}^* & \Sigma_{46}^* & \Sigma_{47}^* & \Sigma_{48}^* \\ \cdot & \cdot & \cdot & \cdot & \Sigma_{55}^* & \Sigma_{56}^* & \Sigma_{57}^* & \Sigma_{58}^* \\ \cdot & \cdot & \cdot & \cdot & \cdot & \Sigma_{66}^* & \Sigma_{67}^* & \Sigma_{68}^* \\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \Sigma_{77}^* & \Sigma_{78}^* \\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \Sigma_{88}^* \end{pmatrix},$$

$\Sigma_{ij}^* = CD\Sigma_{ij}D^TC^T, i, j = 1, \dots, 5, \Sigma_{ij}^* = CD\Sigma_{ij}F^TC^T, i = 1, \dots, 5, j = 6, 7, \Sigma_{ij}^* = CF\Sigma_{ij}F^TC^T, i, j = 6, 7$, where $\Sigma_{ij}, i, j = 1, \dots, 7$ are defined in the proof of lemma B in the supplementary file; $\Sigma_{i8}^* = CD\Sigma_{i8}H^T, i = 1, \dots, 5; \Sigma_{i8}^* = CF\Sigma_{i8}H^T, i = 6, 7; \Sigma_{i8}^* = H\Sigma_{i8}H^T$, where $\Sigma_{i8} = a_{i8}\text{Cov}(\phi^{(i)}, \phi^{(8)})$ $a_{i8} = 2, 4, 2, 2, 3, 1, 1, 1$ corresponding to $i = 1, 2, 3, 4, 5, 6, 7, 8$.

Let $\mathbf{B}^* = \begin{pmatrix} I_{p^2} & (-2)I_{p^2} & I_{p^2} & I_{p^2} & -I_{p^2} & I_{p^2} & -I_{p^2} & I_{p^2} \end{pmatrix}$, then by Slutsky's theorem,

$$\sqrt{n} \begin{pmatrix} \text{vec}(\boldsymbol{\eta}_n \boldsymbol{\eta}_n^T \hat{\Sigma}) - \text{vec}(\boldsymbol{\eta} \boldsymbol{\eta}^T \Sigma) \end{pmatrix} \stackrel{D}{=} \sqrt{n} \mathbf{B}^* \begin{pmatrix} \mathbf{U}_{1n}^* - \mu_1^* \\ \mathbf{U}_{2n}^* - \mu_2^* \\ \mathbf{U}_{3n}^* - \mu_3^* \\ \mathbf{U}_{4n}^* - \mu_4^* \\ \mathbf{U}_{5n}^* - \mu_5^* \\ \mathbf{U}_{6n}^* - \mu_6^* \\ \mathbf{U}_{7n}^* - \mu_7^* \\ \mathbf{U}_{8n}^* - \mu_8^* \end{pmatrix} \longrightarrow N(0, \mathbf{V}_{22}), \text{ where } \mathbf{V}_{22} = \mathbf{B}^* \Sigma^* \mathbf{B}^{*T}.$$

In general, without assuming $\mathbf{Q} = I_d$, we have $\mathbf{V}_{22}(\boldsymbol{\eta}_{\mathbf{Q}}) = \mathbf{B}^* \Sigma^*(\boldsymbol{\eta}_{\mathbf{Q}}) \mathbf{B}^{*T}$, and $\Sigma^*(\boldsymbol{\eta}_{\mathbf{Q}})$ is obtained by replacing every $\boldsymbol{\eta}$ in Σ^* with $\boldsymbol{\eta}_{\mathbf{Q}}$. $\square$

References

- Lehmann, E. L. (1999). *Elements of Large-Sample Theory*. New York: Springer-Verlag.
- Serfling, R. J. (1980). *Approximation Theorems of Mathematical Statistics*. New York: Wiley.
- Székely, G. J., Rizzo, M. L., and Bakirov, N. K. (2007). Measuring and testing dependence by correlation of distances. *The Annals of Statistics*, 35(6), 2769–2794.