Biometrika (2021), **xx**, x, *pp*. 1–18
Printed in Great Britain

Supplementary material for "Significance testing for canonical correlation analysis in high dimensions"

BY IAN W. MCKEAGUE

Department of Biostatistics, Columbia University, Room R639, 722 West 168th Street, New York, NY 10032, USA

im2131@columbia.edu

AND XIN ZHANG

Department of Statistics, Florida State University, 214 OSB, 117 N. Woodward Ave. Tallahassee, FL 32306, USA

xzhang8@fsu.edu

S.1. GREEDY ALGORITHM

S.1.1. *The increments in the sample Pillai trace*

Recall that, Fig. 1 gives the results from a toy example showing that the proposed greedy algorithm for finding the maximal sample root-Pillai trace under varying sparsity constraints provides almost perfect agreement with the full search. The data were generated from Model A1 used in the simulation study (in Section 4), with $p = q = 10$, the true numbers of active variables $s_x^\star = s_y^\star = 3$, $\tau_{\max} = 0.8$, the true number of non-zero canonical correlations $K = 1$, and $n = 500$. The result of the greedy search agrees with the full search at all sparsity levels except at $s = s_x = s_y = 2$.

The left panel of Fig. S.1 shows the successive increments in the sample Pillai trace when adding one variable at a time (either X_k or Y_j), using the same simulation model as the first panel except with $p = q = 5000$. This plot is analogous to the scree plot used in principal component analysis and factor analysis, providing intuition and graphical diagnostics for how sparse the true model might be. It is clear from the plot that $s_x + s_y = 6$ gives the most reasonable terminating point; also, at that point we have $s_x = s_y = 3$ (not shown in the plot), agreeing with the true numbers of active variables. The right panel of Fig. S.1 shows the corresponding values of the studentized $\widehat{\tau}_{\max}$ used as the test statistic for the proposed test. As the sparsity level s increases, the test statistic monotonically increases until s reaches the true value $s_x^\star = s_y^\star = 3$, but the accuracy of the sample coherence matrix $C_{X_\mathcal{K} Y_\mathcal{J}}$ used in the test statistic decreases as its dimension ($s_x \times s_y$) grows, causing a decrease in power at larger values of s. In our experience, the proposed test performs well for relatively small sparsity levels, as in this simulation example, but degrades when $s_x^\star$ and $s_y^\star$ become large (say $s_x^\star = s_y^\star = 20$).

Since the true sparsity levels are $s_x^\star = s_y^\star = 5$, we have satisfactory results (p-value less than 0.01) when $s \geq 4$. When fewer variables are selected, $s \leq 3$, the test procedure is not powerful enough to reject the null at $\alpha = 0.05$ level. It is also interesting to note that slightly over-specified $s = 6$ or 7 produces more significant test results than at the true sparsity level.

S.1.2. *Submodularity*

To gain some further insight into the performance the proposed greedy search algorithm, we show in a simulation example that the root-Pillai trace comes close to satisfying the submodular

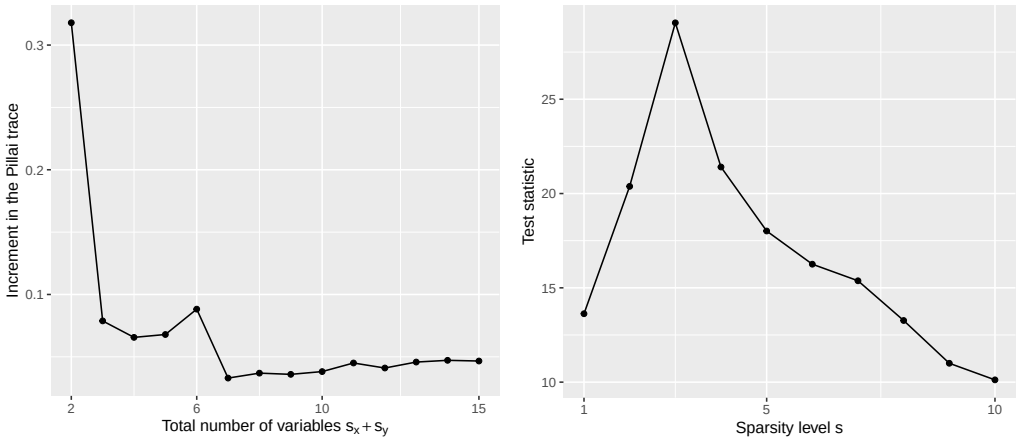

Fig. S.1. Results based on single samples generated under Model A1 with $n = 500$, $s_x^\star = s_y^\star = 3$, $\tau_{\max} = 0.8$. Left: scree plot of the increment in the sample Pillai trace, $\|C_{E_{j|\mathcal{J}} X_{\mathcal{K}}}\|_F^2$ or $\|C_{Y_{\mathcal{J}} R_{k|\mathcal{K}}}\|_F^2$, when adding one variable at a time in the greedy search, for $p = q = 5000$. Right: corresponding values of the studentized $\widehat{\tau}_{\max}$ (test statistic for $\tau_{\max} = 0$) as a function of s.

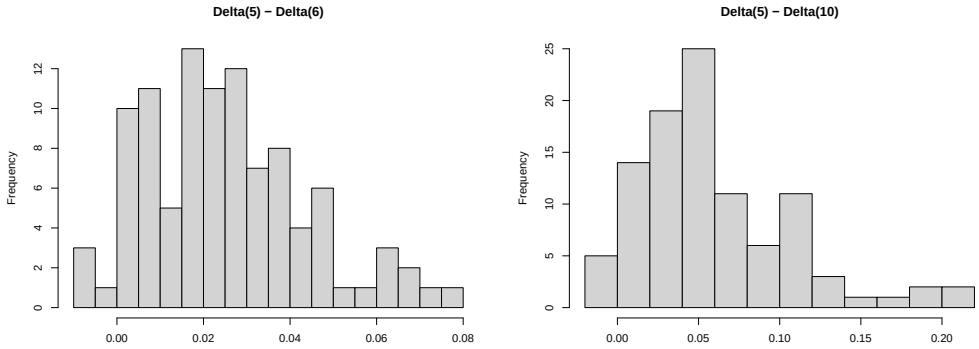

Fig. S.2. Histograms used for checking submodularity of the root-Pillai trace as a function of $S = (\mathcal{J}, \mathcal{K})$ for the GBM data set. The plotted values of $\Delta(e_i \mid S_1) - \Delta(e_i \mid S_2)$ should be non-negative if the function is submodular. The histograms are based on 100 randomly sampled elements $e_i \notin S_2$. The set S_2 is randomly chosen with $|\mathcal{K}| = |\mathcal{J}| = 6$ on the left, and with $|\mathcal{K}| = |\mathcal{J}| = 10$ on the right. For $S_1 \subset S_2$ we take the first 5 elements of S_2.

property (which, if true, would give a guarantee of finding the maximum to within a factor of $1/e$). Maximization of the root-Pillai trace over $\mathcal{K}$ and $\mathcal{J}$ is a discrete combinatorial optimization problem for a set function $f : 2^V \mapsto \mathbb{R}$, where the finite set $V = \{1, \ldots, p\} \times \{1, \ldots, q\}$ and the utility function is $f(S) = \|C_{X_{\mathcal{K}} Y_{\mathcal{J}}}\|_F$ for $S = (\mathcal{J}, \mathcal{K}) \subseteq V$. Submodularity in set functions is the discrete analogy of convexity in continuous functions. Fast greedy algorithms with theoretical guarantees have been developed for submodular function maximization (see Nemhauser & Wolsey, 1978; Krause & Golovin, 2014; Khim et al., 2016, for example), when the util-

ity function is also monotonic. Specifically, the set function f is monotonic if for any two sets $S_1 \subseteq S_2 \subseteq V$, we have $f(S_1) \leq f(S_2)$. Based on Lemma 1, it is not difficult to see that $f(S) = \|C_{X_\mathcal{K} Y_\mathcal{J}}\|_F$ is monotonic.

We now briefly review the definition of submodularity and then numerically demonstrate that our set function f can be close to submodular, although not exactly so. A key concept related to submodularity is the discrete derivative. The discrete derivative of f at S with respect to a new element $e \in V$ is defined as $\Delta(e \mid S) = f(S \cup \{e\}) - f(S)$. Then f is called submodular if, for any $S_1 \subseteq S_2 \subset V$ and $e \in V \setminus S_2$, we have $\Delta(e \mid S_1) \geq \Delta(e \mid S_2)$. We consider a simple numerical experiment using the real data in Section 5, where $(p, q, n) = (1000, 534, 397)$. First, we randomly sampled $S_2 \subset V$ with size $s_x = s_y = 6$ (or 10) and defined the first 5 elements in S_2 to be S_1. Then we computed the histogram of the differences $\Delta(e_i \mid S_1) - \Delta(e_i \mid S_2)$ for 100 randomly selected elements $e_i \notin S_2$. The results displayed in Fig. S.2 indicate a close approximation to submodularity since there are very few negative values in each histogram. In contrast, the Pillai trace is readily seen to violate the submodular property: the scree plot in Fig. S.1 is not monotonically decreasing.

S.2. Computation Time

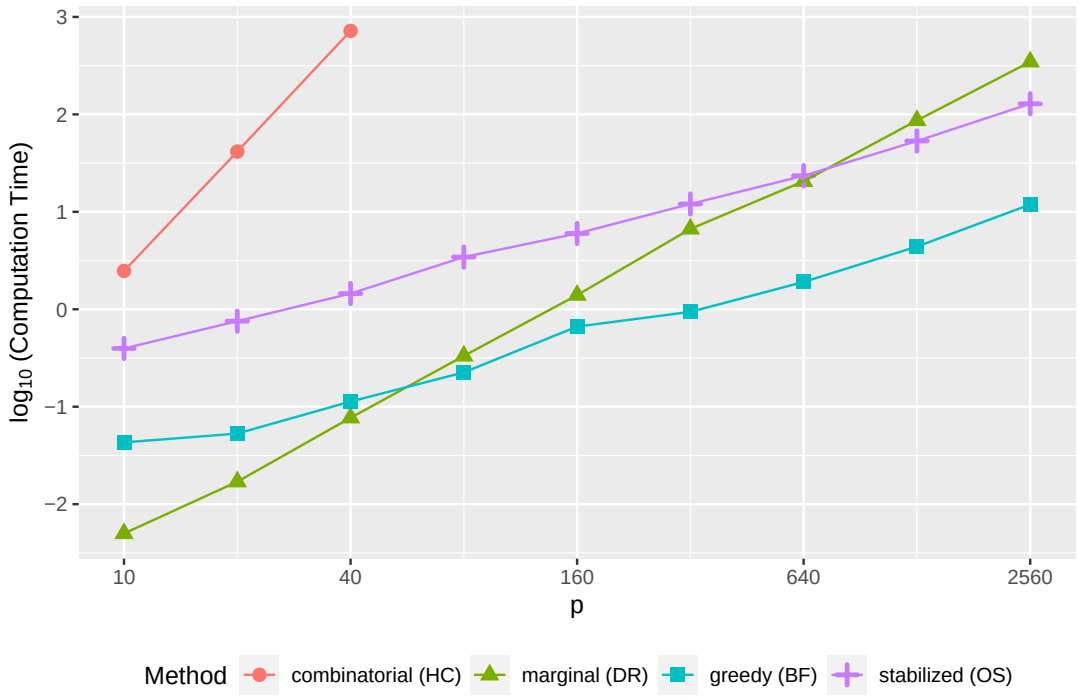

Fig. S.3. Computation time comparison. B-DR and BH-DR have roughly the same computation time and we only report BH-DR.

In this section, we report the computation time of each of the following methods: Higher Criticism (HC) (Donoho & Jin, 2004, 2015) based on p-values computed from the F-test for all $\binom{p}{s_x}\binom{q}{s_y}$ combinations of variables; the normal approximation test (B-DR or BH-DR) described in DiCiccio & Romano (2017) and applied to all pq pairs of marginal correlations $\mathrm{corr}(X_k, Y_j)$;

the F-test on selected variables from the greedy search algorithm and with Bonferroni correction (BF); the proposed testing procedure using the stabilized one-step estimator (OS) with the greedy search algorithm.

For the simulation example of Model N, Figure S.3 shows the computation time for one data replication. We set $s_x = s_y = 2$ so that HC is feasible for $p = q = 40$. All computation was done on a MacOS laptop computer with 3.1 GHz Intel Core i5 processor and 16 GB memory. Recall that $n = 500$ and we vary $p = q \in \{10, 20, \ldots, 2560\}$. From this plot, we clearly see that the proposed greedy search algorithm, applied to either the F-test on selected variables with Bonferroni correction or the stabilized one-step estimator, scales better with high dimensionality than methods based on combinatorial full search over $\binom{p}{s_x}\binom{q}{s_y}$ test statistics or based on marginal selection over all pq pairs of variables. For very large p and q, we find that the methods based on our greedy search algorithm (i.e., BF and OS) are computationally tractable while other methods are not.

S.3. GLIOBLASTOMA MULTIFORME (GBM) DATA ANALYSIS

Table S.1 lists the most correlated variables under sparsity level $s = 3$. Interestingly, the first two microRNA measurements (*hsa.miR.219* and *hsa.miR.222*) also appear in a reported dependency network of important microRNAs obtained by precision matrix estimation (Wang, 2015, Figure 1). In Fig. S.5, we repeat the analysis in the paper for 10 random re-orderings of the data, and compare the results to the ones without random re-ordering. The random ordering of the samples has little effects on the significance test result. The two plots are very similar, and it is clear that the number of important variables should be specified as less than 25. In Table S.2, we report the selected top-25 variables for X (gene expression) and Y (microRNA). The selection was based on the proposed greedy algorithm (Algorithm 1).

For $s = s_x = s_y = 1$, we calculated the 95% confidence intervals based on 100 random re-ordering of the original data. The endpoints of the CIs are displayed in Fig. S.4, where each point in the scatterplot represents one CI. All the CIs have very similar widths, and are far away from zero, which is consistent with the finding of very small p-values.

Table S.1. *GBM data analysis. Marginal correlations between the selected genes (each row) and microRNAs (each column) at sparsity level* $s = s_x = s_y = 3$. *The stabilized one-step estimator* $\widehat{\tau}_{\max} = 0.931$ *with standard error* 0.085

	hsa.miR.219	hsa.miR.222	hsa.miR.138
SPIRE2	0.76	−0.27	0.50
FGF9	0.76	−0.06	0.33
ZNF553	−0.24	0.61	−0.17

S.4. CANONICAL GRADIENT

S.4.1. *Some matrix notation and properties*

The Frobenius norm of a matrix $A \in \mathbb{R}^{p\times q}$ is $\|A\|_F = (\sum_{i,j} A_{ij}^2)^{1/2}$. The maximum norm $\|A\|_{\max} = \max_{i,j} |A_{ij}|$ and the ℓ_1-norm $\|A\|_1 = \max_j \sum_i |A_{ij}|$. The ℓ_2-norm (or operator norm) $\|A\|$ is its largest singular value, and the ℓ_∞-norm is $\|A\|_\infty = \max_i \sum_j |A_{ij}|$. For a symmetric matrix A, $\|A\|_1 = \|A\|_\infty$. For any matrices $A \in \mathbb{R}^{p\times q}$ and $B \in \mathbb{R}^{r\times s}$,

Table S.2. *Top-25 selected variables based on Algorithm 1 for GBM Data Set. The sequentially selected variables are ordered from top left to bottom right*

	Selected Variables				
gene expression	SPIRE2	FGF9	ZNF553	CPD	UCK2
	KIAA1632	SCGB3A2	SAMD9	KBTBD5	AURKB
	ERCC1	FAM113B	CCL19	DLX4	GCG
	LRRTM3	SCN10A	SMC1A	SBNO2	CENTG2
	RHOV	COX7A1	ROGDI	SUHW2	SNX1
microRNA	hsa-miR-219	hsa-miR-222	hsa-miR-138	hsa-miR-124a	hsa-miR-130b
	hsa-miR-92b	hsa-miR-423	hsa-miR-9*	hsa-miR-199a*	hsa-miR-22
	hsa-miR-195	hsa-miR-21	hsa-miR-204	hsa-miR-30e-3p	hsa-miR-181d
	hsa-miR-186	hsa-miR-484	hsa-miR-210	kshv-miR-K12-12	hsa-miR-376a
	hsa-miR-106b	hsa-miR-9	d1.hsa-miR-15b	hsa-miR-181a	hsa-miR-217

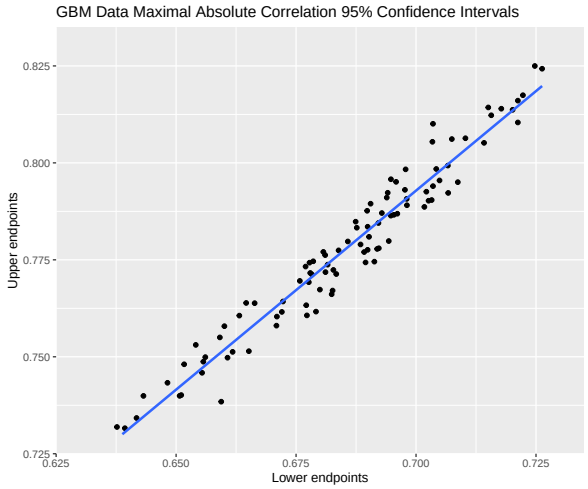

Fig. S.4. GBM data analysis. The lower and upper endpoints of the 95% confidence interval for $\tau_{\max}$ when $s = 1$, for 100 random re-orderings of the data.

we have $\|A\| \le \|A\|_F$, $p^{-1/2}\|A\|_1 \le \|A\| \le q^{1/2}\|A\|_1$, $\|AB\| \le \|A\|\|B\| = \|A \otimes B\|$ and $\|A\| \le (\|A\|_1\|A\|_\infty)^{1/2}$. For a symmetric matrix $A \in \mathbb{R}^{p\times p}$, we have $\|A\| \le \|A\|_1 = \|A\|_\infty \le p\|A\|_{\max}$ and also $\|A\| \le \|A\|_F \le p\|A\|_{\max}$.

We will also be using the following matrix inequalities repeatedly in the proofs below. For matrices A, B, X, Y with compatible dimensions, we have

$$\begin{aligned}
|\mathrm{tr}(AXBY)| &= |\mathrm{vec}^{\mathrm{T}}(Y)(B^{\mathrm{T}} \otimes A)\mathrm{vec}(X)| \\
&\le \|\mathrm{vec}(Y)\| \cdot \|(B^{\mathrm{T}} \otimes A)\mathrm{vec}(X)\| \\
&\le \|\mathrm{vec}(Y)\| \cdot \|(B^{\mathrm{T}} \otimes A)\| \cdot \|\mathrm{vec}(X)\| \\
&= \|A\|\|B\| \cdot \|\mathrm{vec}(X)\| \cdot \|\mathrm{vec}(Y)\| \\
&= \|A\|\|B\| \cdot \|X\|_F \cdot \|Y\|_F.
\end{aligned}$$

The operator norm (largest singular value of a matrix) satisfies $\|AX\| \le \|A\|\|X\| = \|A \otimes X\|$.

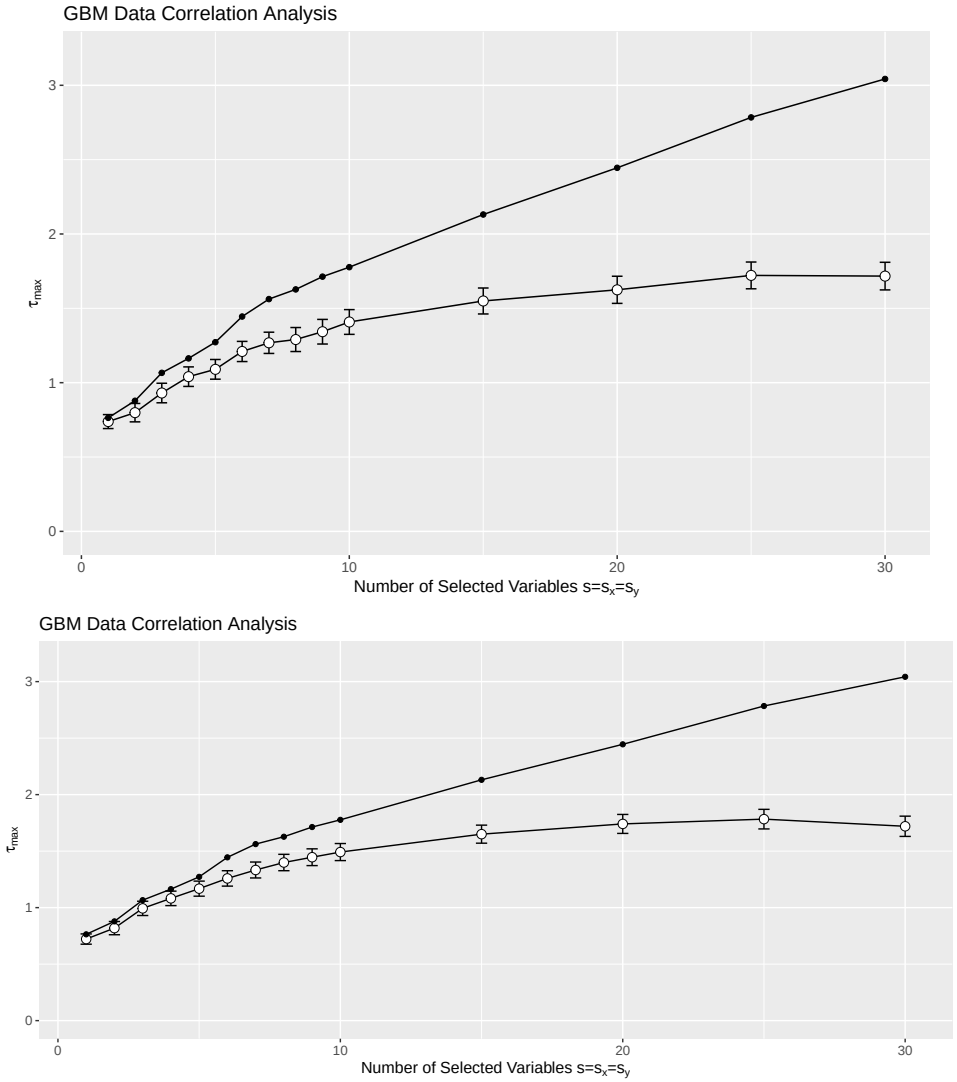

Fig. S.5. GBM data analysis. Estimates of the maximal root-Pillai trace $\tau_{\max}$ are plotted against $s = s_x = s_y$ varying from 1–30, with the selected variables at each step found using Algorithm 1. The black dots are estimated values without adjusting for post-selection. The white dots and associated 95% confidence intervals are based on the stabilized one-step estimator. The top panel is without random re-ordering of the data; the bottom panel uses 10 random re-orderings and reports the averaged point estimates and averaged lower/upper CI endpoints over these re-orderings.

S.4.2. *Proof of Theorem 1 (Derivation of the canonical gradient)*

In the following derivation, we write

$$\Phi(P) = \Phi^d(P) = \|\Lambda_{X_{\mathcal{K}}Y_{\mathcal{J}}}\|_F^2 = \|\Lambda_{XY}\|_F^2 = \mathrm{tr}(\Lambda_{XY}\Lambda_{XY}^{\mathrm{T}}) = \mathrm{tr}(\Sigma_X^{-1}\Sigma_{XY}\Sigma_Y^{-1}\Sigma_{XY}^{\mathrm{T}}).$$

Using standard results from matrix calculus,

$$\begin{aligned} d\Phi(P) &= d\mathrm{tr}(\Sigma_X^{-1}\Sigma_{XY}\Sigma_Y^{-1}\Sigma_{XY}^{\mathrm{T}}) \\ &= \frac{\partial \mathrm{tr}(\Sigma_X^{-1}\Sigma_{XY}\Sigma_Y^{-1}\Sigma_{XY}^{\mathrm{T}})}{\partial \Sigma_X}\cdot d\Sigma_X + \frac{\partial \mathrm{tr}(\Sigma_X^{-1}\Sigma_{XY}\Sigma_Y^{-1}\Sigma_{XY}^{\mathrm{T}})}{\partial \Sigma_Y}\cdot d\Sigma_Y \\ &\quad + \frac{\partial \mathrm{tr}(\Sigma_X^{-1}\Sigma_{XY}\Sigma_Y^{-1}\Sigma_{XY}^{\mathrm{T}})}{\partial \Sigma_{XY}}\cdot d\Sigma_{XY} \\ &= -\Sigma_X^{-1}\Sigma_{XY}\Sigma_Y^{-1}\Sigma_{XY}^{\mathrm{T}}\Sigma_X^{-1}\cdot d\Sigma_X - \Sigma_Y^{-1}\Sigma_{XY}^{\mathrm{T}}\Sigma_X^{-1}\Sigma_{XY}\Sigma_Y^{-1}\cdot d\Sigma_Y \\ &\quad + 2\Sigma_X^{-1}\Sigma_{XY}\Sigma_Y^{-1}\cdot d\Sigma_{XY}, \end{aligned}$$

where the dot indicates inner product of two matrices of same dimension: $A \cdot dB = \langle A, dB\rangle = \mathrm{vec}^{\mathrm{T}}(A)\mathrm{vec}(dB) = \mathrm{tr}(A^{\mathrm{T}}dB)$. Hence

$$\begin{aligned} \frac{d\Phi(P_\epsilon)}{d\epsilon} &= -\mathrm{tr}\left\{\Sigma_X^{-1}\Sigma_{XY}\Sigma_Y^{-1}\Sigma_{XY}^{\mathrm{T}}\Sigma_X^{-1}\frac{d\Sigma_X(P_\epsilon)}{d\epsilon}\right\} \\ &\quad -\mathrm{tr}\left\{\Sigma_Y^{-1}\Sigma_{XY}^{\mathrm{T}}\Sigma_X^{-1}\Sigma_{XY}\Sigma_Y^{-1}\frac{d\Sigma_Y(P_\epsilon)}{d\epsilon}\right\} + 2\mathrm{tr}\left\{\Sigma_X^{-1}\Sigma_{XY}\Sigma_Y^{-1}\frac{d\Sigma_{XY}(P_\epsilon)}{d\epsilon}\right\}, \end{aligned}$$

recalling that $P_\epsilon = (1-\epsilon)P + \epsilon\delta_o$, where δ_o is the Dirac measure at $o = (x^{\mathrm{T}}, y^{\mathrm{T}})^{\mathrm{T}}$. Expressing $\Sigma_X(P_\epsilon)$ as

$$(1-\epsilon)P(XX^{\mathrm{T}}) + \epsilon xx^{\mathrm{T}} - (1-\epsilon)^2P(X)P(X^{\mathrm{T}}) - \epsilon(1-\epsilon)\{P(X)x^{\mathrm{T}} + xP(X^{\mathrm{T}})\} - \epsilon^2 xx^{\mathrm{T}},$$

by direct calculation we then have

$$\left.\frac{d\Sigma_X(P_\epsilon)}{d\epsilon}\right|_{\epsilon=0} = \lim_{\epsilon\downarrow 0}\frac{\Sigma_X(P_\epsilon) - \Sigma_X(P)}{\epsilon} = \{x - P(X)\}\{x^{\mathrm{T}} - P(X^{\mathrm{T}})\}.$$

Similarly,

$$\left.\frac{d\Sigma_{XY}(P_\epsilon)}{d\epsilon}\right|_{\epsilon=0} = \{x - P(X)\}\{y^{\mathrm{T}} - P(Y^{\mathrm{T}})\}, \quad \left.\frac{d\Sigma_Y(P_\epsilon)}{d\epsilon}\right|_{\epsilon=0} = \{y - P(Y)\}\{y^{\mathrm{T}} - P(Y^{\mathrm{T}})\}.$$

Plugging-in these expressions, and including the relevant sets of indices $\mathcal{K}$ and $\mathcal{J}$, we obtain canonical gradient of $\Phi^d(P)$ as stated in (11):

$$\begin{aligned} \left.\frac{d\Phi^d(P_\epsilon)}{d\epsilon}\right|_{\epsilon=0} &= -\{x_{\mathcal{K}} - \mathrm{E}_P(X_{\mathcal{K}})\}^{\mathrm{T}}\Sigma_{X_{\mathcal{K}}}^{-1}\Sigma_{X_{\mathcal{K}}Y_{\mathcal{J}}}\Sigma_{Y_{\mathcal{J}}}^{-1}\Sigma_{Y_{\mathcal{J}}X_{\mathcal{K}}}\Sigma_{X_{\mathcal{K}}}^{-1}\{x_{\mathcal{K}} - \mathrm{E}_P(X_{\mathcal{K}})\} \\ &\quad -\{y_{\mathcal{J}} - \mathrm{E}_P(Y_{\mathcal{J}})\}^{\mathrm{T}}\Sigma_{Y_{\mathcal{J}}}^{-1}\Sigma_{Y_{\mathcal{J}}X_{\mathcal{K}}}\Sigma_{X_{\mathcal{K}}}^{-1}\Sigma_{X_{\mathcal{K}}Y_{\mathcal{J}}}\Sigma_{Y_{\mathcal{J}}}^{-1}\{y_{\mathcal{J}} - \mathrm{E}_P(Y_{\mathcal{J}})\} \\ &\quad +2\{y_{\mathcal{J}} - \mathrm{E}_P(Y_{\mathcal{J}})\}^{\mathrm{T}}\Sigma_{Y_{\mathcal{J}}}^{-1}\Sigma_{Y_{\mathcal{J}}X_{\mathcal{K}}}\Sigma_{X_{\mathcal{K}}}^{-1}\{x_{\mathcal{K}} - \mathrm{E}_P(X_{\mathcal{K}})\}. \end{aligned}$$

Then, as discussed in Section 2.1, the canonical gradient of $\Psi^d(P) = \|\Lambda_{X_{\mathcal{K}}Y_{\mathcal{J}}}\|_F$ is obtained via the relationship $\{\Psi^d(P)\}^2 = \Phi^d(P)$, resulting in

$$D^d(P)(o) = \left.\frac{d\Psi^d(P_\epsilon)}{d\epsilon}\right|_{\epsilon=0} = \frac{1}{2\Psi^d(P)}\left.\frac{d\Phi^d(P_\epsilon)}{d\epsilon}\right|_{\epsilon=0}$$

when $\Psi^d(P) \neq 0$.

To prove $\mathrm{E}_P\{D^d(P)(O)\} = 0$ it suffices to show that the canonical gradient of $\Phi^d(P)$ has zero mean. We apply the following property of trace and expectation operators. For any random vector $X \in \mathbb{R}^p$ with entries having finite second moments and any non-stochastic matrix $M \in$

$\mathbb{R}^{p\times p}$,

$$\mathrm{E}(X^{\mathrm{T}}MX) = \mathrm{E}\{\mathrm{tr}(X^{\mathrm{T}}MX)\} = \mathrm{E}\{\mathrm{tr}(MXX^{\mathrm{T}})\} = \mathrm{tr}\{M\mathrm{E}(XX^{\mathrm{T}})\}. \tag{S.1}$$

Therefore, direct calculation shows that

$$\begin{aligned}
\mathrm{E}\left\{ \left.\frac{d\Phi^d(P_\epsilon)}{d\epsilon}\right|_{\epsilon=0}\right\} = & -\mathrm{tr}\left(\Sigma_{X_\mathcal{K}}^{-1}\Sigma_{X_\mathcal{K}Y_\mathcal{J}}\Sigma_{Y_\mathcal{J}}^{-1}\Sigma_{Y_\mathcal{J}X_\mathcal{K}}\Sigma_{X_\mathcal{K}}^{-1}\cdot \mathrm{E}_P\left[\{X_\mathcal{K}-\mathrm{E}_P(X_\mathcal{K})\}\{X_\mathcal{K}-\mathrm{E}_P(X_\mathcal{K})\}^{\mathrm{T}}\right]\right) \\
& -\mathrm{tr}\left(\Sigma_{Y_\mathcal{J}}^{-1}\Sigma_{Y_\mathcal{J}X_\mathcal{K}}\Sigma_{X_\mathcal{K}}^{-1}\Sigma_{X_\mathcal{K}Y_\mathcal{J}}\Sigma_{Y_\mathcal{J}}^{-1}\cdot \mathrm{E}_P\left[\{Y_\mathcal{J}-\mathrm{E}_P(Y_\mathcal{J})\}\{Y_\mathcal{J}-\mathrm{E}_P(Y_\mathcal{J})\}^{\mathrm{T}}\right]\right) \\
& +2\mathrm{tr}\left(\Sigma_{Y_\mathcal{J}}^{-1}\Sigma_{Y_\mathcal{J}X_\mathcal{K}}\Sigma_{X_\mathcal{K}}^{-1}\cdot \mathrm{E}_P\left[\{X_\mathcal{K}-\mathrm{E}_P(X_\mathcal{K})\}\{Y_\mathcal{J}-\mathrm{E}_P(Y_\mathcal{J})\}^{\mathrm{T}}\right]\right) \\
= & -\Phi^d(P)-\Phi^d(P)+2\Phi^d(P)=0.
\end{aligned}$$

When $\Psi^d(P)=0$, using instead the expression (12) for the canonical gradient and again applying (S.1) we obtain

$$\begin{aligned}
\mathrm{E}_P\{D^d(P)(O)\} &= \mathrm{E}_P\left[\{Y_\mathcal{J}-\mathrm{E}_P(Y_\mathcal{J})\}^{\mathrm{T}}\Sigma_{Y_\mathcal{J}}^{-1/2}L\Sigma_{X_\mathcal{K}}^{-1/2}\{X_\mathcal{K}-\mathrm{E}_P(X_\mathcal{K})\}\right] \\
&= \mathrm{tr}\left(L\Sigma_{X_\mathcal{K}}^{-1/2}\mathrm{E}_P\left[\{X_\mathcal{K}-\mathrm{E}_P(X_\mathcal{K})\}\{Y_\mathcal{J}-\mathrm{E}_P(Y_\mathcal{J})\}^{\mathrm{T}}\right]\Sigma_{Y_\mathcal{J}}^{-1/2}\right) \\
&= \mathrm{tr}(L\Lambda_{X_\mathcal{K}Y_\mathcal{J}}) = 0
\end{aligned}$$

since $\Lambda_{X_\mathcal{K}Y_\mathcal{J}}=0$ when $\Psi^d(P)=0$.

S.4.3. *Derivation of the remainder term*

The second order remainder term is defined as,

$$\mathrm{Rem}(P) = \Psi^d(P) - \Psi^d(P_0) + \int D^d(P)(o)dP_0(o), \tag{S.2}$$

where we later use P to denote the sample empirical distribution and P_0 to denote the population distribution. In the following calculations, we omit the superscript d since it is fixed. To simplify the calculation, it is convenient to study $\Phi(P)=\mathrm{tr}(S_X^{-1}S_{XY}S_Y^{-1}S_{XY}^{\mathrm{T}})$ and $\Phi(P_0)=\mathrm{tr}(\Sigma_X^{-1}\Sigma_{XY}\Sigma_Y^{-1}\Sigma_{XY}^{\mathrm{T}})$ instead of doing the calculation in the $\Psi=\Phi^{1/2}$ scale. As such, we have

$$\begin{aligned}
\int D(P)(o)dP_0(o) = & -\mathrm{tr}\left\{S_X^{-1}S_{XY}S_Y^{-1}S_{YX}S_X^{-1}P_0\left[\{x-P(X)\}\{x^{\mathrm{T}}-P(X^{\mathrm{T}})\}\right]\right\} \\
& -\mathrm{tr}\left\{S_Y^{-1}S_{YX}S_X^{-1}S_{XY}S_Y^{-1}P_0\left[\{y-P(Y)\}\{y^{\mathrm{T}}-P(Y^{\mathrm{T}})\}\right]\right\} \\
& +2\mathrm{tr}\left\{S_Y^{-1}S_{YX}S_X^{-1}P_0\left[\{x-P(X)\}\{y^{\mathrm{T}}-P(Y^{\mathrm{T}})\}\right]\right\}.
\end{aligned}$$

We can substitute the $P_0[\cdots]$ terms as follows,

$$P_0\left[\{x-P(X)\}\{y^{\mathrm{T}}-P(Y^{\mathrm{T}})\}\right] = \Sigma_{XY} - (\mu_{X,0}-\mu_X)(\mu_{Y,0}-\mu_Y)^{\mathrm{T}},$$

where we simplified notation as $\mu_X=P(X)$, $\mu_Y=P(Y)$, $\mu_{X,0}=P_0(X)$ and $\mu_{Y,0}=P_0(Y)$. We decompose the term $\int D(P)(o)dP_0(o)=T_1+T_2$ into two parts accordingly. The first part is $O_P(1/n)$ if μ_X and μ_Y are close to $\mu_{X,0}$ and $\mu_{Y,0}$ in the order of $O_p(n^{-1/2})$:

$$\begin{aligned}
T_1 = & 2\mathrm{tr}\left\{S_Y^{-1}S_{YX}S_X^{-1}(\mu_{X,0}-\mu_X)(\mu_{Y,0}-\mu_Y)^{\mathrm{T}}\right\} \\
& -\mathrm{tr}\left\{S_X^{-1}S_{XY}S_Y^{-1}S_{YX}S_X^{-1}(\mu_{X,0}-\mu_X)(\mu_{X,0}-\mu_X)^{\mathrm{T}}\right\} \\
& -\mathrm{tr}\left\{S_Y^{-1}S_{YX}S_X^{-1}S_{XY}S_Y^{-1}(\mu_{Y,0}-\mu_Y)(\mu_{Y,0}-\mu_Y)^{\mathrm{T}}\right\}.
\end{aligned}$$

The second part is small when Σ's are close to Σ_0's, but not small enough (i.e. $O_P(n^{-1/2})$ instead of $O_P(n^{-1})$) and will be combined with $\Phi(P) - \Phi(P_0)$ later:

$$\begin{aligned}
T_2 = & 2\mathrm{tr}\left(S_Y^{-1}S_{YX}S_X^{-1}\Sigma_{XY}\right) \\
& -\mathrm{tr}\left(S_X^{-1}S_{XY}S_Y^{-1}S_{YX}S_X^{-1}\Sigma_X\right) - \mathrm{tr}\left(S_Y^{-1}S_{YX}S_X^{-1}S_{XY}S_Y^{-1}\Sigma_Y\right) \\
= & 2\mathrm{tr}\left\{S_Y^{-1}S_{YX}S_X^{-1}(\Sigma_{XY} - S_{XY})\right\} + \mathrm{tr}\left\{S_X^{-1}S_{XY}S_Y^{-1}S_{YX}S_X^{-1}(S_X - \Sigma_X)\right\} \\
& +\mathrm{tr}\left\{S_Y^{-1}S_{YX}S_X^{-1}S_{XY}S_Y^{-1}(S_Y - \Sigma_Y)\right\}.
\end{aligned}$$

In $\mathrm{Rem}(P)$, we express $\Phi(P) - \Phi(P_0)$ as follows,

$$\begin{aligned}
\Phi(P) - \Phi(P_0) = & \mathrm{tr}(S_X^{-1}S_{XY}S_Y^{-1}S_{YX} - \Sigma_X^{-1}\Sigma_{XY}\Sigma_Y^{-1}\Sigma_{YX}) \\
= & \mathrm{tr}\left\{\Sigma_{XY}\Sigma_Y^{-1}\Sigma_{YX}(S_X^{-1} - \Sigma_X^{-1})\right\} + \mathrm{tr}\left\{\Sigma_{YX}S_X^{-1}\Sigma_{XY}(S_Y^{-1} - \Sigma_Y^{-1})\right\} \\
& +\mathrm{tr}\left\{S_X^{-1}\Sigma_{XY}S_Y^{-1}(S_{YX} - \Sigma_{YX})\right\} + \mathrm{tr}\left\{S_Y^{-1}S_{YX}S_X^{-1}(S_{XY} - \Sigma_{XY})\right\},
\end{aligned}$$

where the first two part can be re-organized by noticing $(S_X^{-1} - \Sigma_X^{-1}) = S_X^{-1}(\Sigma_X - S_X)\Sigma_X^{-1}$. Hence, we have

$$\begin{aligned}
\Phi(P) - \Phi(P_0) = & \mathrm{tr}\left\{\Sigma_X^{-1}\Sigma_{XY}\Sigma_Y^{-1}\Sigma_{YX}S_X^{-1}(\Sigma_X - S_X)\right\} \\
& +\mathrm{tr}\left\{S_Y^{-1}\Sigma_{YX}S_X^{-1}\Sigma_{XY}\Sigma_Y^{-1}(\Sigma_Y - S_Y)\right\} \\
& +\mathrm{tr}\left\{S_X^{-1}\Sigma_{XY}S_Y^{-1}(S_{YX} - \Sigma_{YX})\right\} + \mathrm{tr}\left\{S_Y^{-1}S_{YX}S_X^{-1}(S_{XY} - \Sigma_{XY})\right\}.
\end{aligned}$$

Adding the various terms, we have

$$\begin{aligned}
\mathrm{Rem}(P) = & T_1 + T_2 + \Phi(P) - \Phi(P_0) \\
= & 2\mathrm{tr}\left\{S_Y^{-1}S_{YX}S_X^{-1}(\mu_{X,0} - \mu_X)(\mu_{Y,0} - \mu_Y)^{\mathrm{T}}\right\} \\
& -\mathrm{tr}\left\{S_X^{-1}S_{XY}S_Y^{-1}S_{YX}S_X^{-1}(\mu_{X,0} - \mu_X)(\mu_{X,0} - \mu_X)^{\mathrm{T}}\right\} \\
& -\mathrm{tr}\left\{S_Y^{-1}S_{YX}S_X^{-1}S_{XY}S_Y^{-1}(\mu_{Y,0} - \mu_Y)(\mu_{Y,0} - \mu_Y)^{\mathrm{T}}\right\} \\
& +\mathrm{tr}\left\{S_Y^{-1}(\Sigma_{YX} - S_{YX})S_X^{-1}(S_{XY} - \Sigma_{XY})\right\} \\
& +\mathrm{tr}\left\{\left(\Sigma_X^{-1}\Sigma_{XY}\Sigma_Y^{-1}\Sigma_{YX} - S_X^{-1}S_{XY}S_Y^{-1}S_{YX}\right)S_X^{-1}(\Sigma_X - S_X)\right\} \\
& +\mathrm{tr}\left\{S_Y^{-1}\left(\Sigma_{YX}S_X^{-1}\Sigma_{XY}\Sigma_Y^{-1} - S_{YX}S_X^{-1}S_{XY}S_Y^{-1}\right)(\Sigma_Y - S_Y)\right\},
\end{aligned}$$

which leads to the following final expression after expanding the last two terms,

$$\begin{aligned}
\mathrm{Rem}(P) &= T_1 + T_2 + \Phi(P) - \Phi(P_0) \\
&= 2\mathrm{tr}\left\{S_Y^{-1}S_{YX}S_X^{-1}(\mu_{X,0}-\mu_X)(\mu_{Y,0}-\mu_Y)^{\mathrm{T}}\right\} \\
&\quad -\mathrm{tr}\left\{S_X^{-1}S_{XY}S_Y^{-1}S_{YX}S_X^{-1}(\mu_{X,0}-\mu_X)(\mu_{X,0}-\mu_X)^{\mathrm{T}}\right\} \\
&\quad -\mathrm{tr}\left\{S_Y^{-1}S_{YX}S_X^{-1}S_{XY}S_Y^{-1}(\mu_{Y,0}-\mu_Y)(\mu_{Y,0}-\mu_Y)^{\mathrm{T}}\right\} \\
&\quad +\mathrm{tr}\left\{S_Y^{-1}(\Sigma_{YX}-S_{YX})S_X^{-1}(S_{XY}-\Sigma_{XY})\right\} \\
&\quad +\mathrm{tr}\left\{\Sigma_X^{-1}\Sigma_{XY}\Sigma_Y^{-1}(\Sigma_{YX}-S_{YX})S_X^{-1}(\Sigma_X-S_X)\right\} \\
&\quad +\mathrm{tr}\left\{\Sigma_X^{-1}\Sigma_{XY}\Sigma_Y^{-1}(S_Y-\Sigma_Y)S_Y^{-1}S_{YX}S_X^{-1}(\Sigma_X-S_X)\right\} \\
&\quad +\mathrm{tr}\left\{\Sigma_X^{-1}(\Sigma_{XY}-S_{XY})S_Y^{-1}S_{YX}S_X^{-1}(\Sigma_X-S_X)\right\} \\
&\quad +\mathrm{tr}\left\{\Sigma_X^{-1}(S_X-\Sigma_X)S_X^{-1}S_{XY}S_Y^{-1}S_{YX}S_X^{-1}(\Sigma_X-S_X)\right\} \\
&\quad +\mathrm{tr}\left\{S_Y^{-1}\Sigma_{YX}S_X^{-1}\Sigma_{XY}\Sigma_Y^{-1}(S_Y-\Sigma_Y)S_Y^{-1}(\Sigma_Y-S_Y)\right\} \\
&\quad +\mathrm{tr}\left\{S_Y^{-1}\Sigma_{YX,0}S_X^{-1}(\Sigma_{XY}-S_{XY})S_Y^{-1}(\Sigma_Y-S_Y)\right\} \\
&\quad +\mathrm{tr}\left\{S_Y^{-1}(\Sigma_{YX}-S_{YX})S_X^{-1}S_{XY}S_Y^{-1}(\Sigma_Y-S_Y)\right\}. \qquad \text{(S.3)}
\end{aligned}$$

S.4.4. *Proof of Lemma* 1

Let $\mathbb{X} \in \mathbb{R}^{n\times s_x}$, $\mathbb{Y} \in \mathbb{R}^{n\times s_y}$ and $\mathbb{Z} \in \mathbb{R}^{n\times 1}$ be the centered data matrices of $X_{\mathcal{K}}$, $Y_{\mathcal{J}}$ and X_k. That is, the i-th row of $\mathbb{X}$ is $(X_{\mathcal{K},i} - \overline{X}_{\mathcal{K}})^{\mathrm{T}}$. Note that $S_{X_{\mathcal{K}}} = \mathbb{X}\mathbb{X}^{\mathrm{T}}/n$ is the sample covariance matrix of $X_{\mathcal{K}}$, which is assumed to be positive definite. Then we can write $C_{Y_{\mathcal{J}}X_{\mathcal{K}}} = n^{-1}S_{Y_{\mathcal{J}}}^{-1/2}\mathbb{Y}^{\mathrm{T}}\mathbb{X}S_{X_{\mathcal{K}}}^{-1/2}$ and $C_{Y_{\mathcal{J}}X_{\mathcal{K}}}C_{Y_{\mathcal{J}}X_{\mathcal{K}}}^{\mathrm{T}} = n^{-1}S_{Y_{\mathcal{J}}}^{-1/2}\mathbb{Y}^{\mathrm{T}}\mathbb{X}(\mathbb{X}^{\mathrm{T}}\mathbb{X})^{-1}\mathbb{X}^{\mathrm{T}}\mathbb{Y}S_{Y_{\mathcal{J}}}^{-1/2}$, where $\mathbb{X}(\mathbb{X}^{\mathrm{T}}\mathbb{X})^{-1}\mathbb{X}^{\mathrm{T}} \equiv P_{\mathbb{X}} \in \mathbb{R}^{n\times n}$ is the projection matrix onto the s_x-dimensional subspace spanned by the columns of $\mathbb{X}$. Then,

$$\|C_{Y_{\mathcal{J}}X_{\mathcal{K}}}\|_F^2 = \mathrm{tr}(C_{Y_{\mathcal{J}}X_{\mathcal{K}}}C_{Y_{\mathcal{J}}X_{\mathcal{K}}}^{\mathrm{T}}) = n^{-1}\mathrm{tr}(S_{Y_{\mathcal{J}}}^{-1/2}\mathbb{Y}^{\mathrm{T}}P_{\mathbb{X}}\mathbb{Y}S_{Y_{\mathcal{J}}}^{-1/2}), \qquad \text{(S.4)}$$

$$\|C_{Y_{\mathcal{J}}X_{\mathcal{K}\cup k}}\|_F^2 - \|C_{Y_{\mathcal{J}}X_{\mathcal{K}}}\|_F^2 = n^{-1}\mathrm{tr}\{S_{Y_{\mathcal{J}}}^{-1/2}\mathbb{Y}^{\mathrm{T}}(P_{(\mathbb{X},\mathbb{Z})} - P_{\mathbb{X}})\mathbb{Y}S_{Y_{\mathcal{J}}}^{-1/2}\}, \qquad \text{(S.5)}$$

where $P_{(\mathbb{X},\mathbb{Z})} \in \mathbb{R}^{n\times n}$ is the projection matrix onto the (s_x+1)-dimensional subspace of $\mathbb{R}^n$ spanned by the columns of $(\mathbb{X},\mathbb{Z})$. That is,

$$\begin{aligned}
P_{(\mathbb{X},\mathbb{Z})} &= \begin{pmatrix}\mathbb{X} & \mathbb{Z}\end{pmatrix}\begin{pmatrix}\mathbb{X}^{\mathrm{T}}\mathbb{X} & \mathbb{X}^{\mathrm{T}}\mathbb{Z} \\ \mathbb{Z}^{\mathrm{T}}\mathbb{X} & \mathbb{Z}^{\mathrm{T}}\mathbb{Z}\end{pmatrix}^{-1}\begin{pmatrix}\mathbb{X}^{\mathrm{T}} \\ \mathbb{Z}^{\mathrm{T}}\end{pmatrix} \\
&= n^{-1}\begin{pmatrix}\mathbb{X} & \mathbb{Z}\end{pmatrix}\begin{pmatrix}S_{X_{\mathcal{K}}}^{-1} + S_{X_{\mathcal{K}}}^{-1}S_{X_{\mathcal{K}}X_k}S_{R_k}^{-1}S_{X_kX_{\mathcal{K}}}S_{X_{\mathcal{K}}}^{-1} & -S_{X_{\mathcal{K}}}^{-1}S_{X_{\mathcal{K}}X_k}S_{R_k}^{-1} \\ -S_{R_k}^{-1}S_{X_kX_{\mathcal{K}}}S_{X_{\mathcal{K}}}^{-1} & S_{R_k}^{-1}\end{pmatrix}\begin{pmatrix}\mathbb{X}^{\mathrm{T}} \\ \mathbb{Z}^{\mathrm{T}}\end{pmatrix} \\
&= n^{-1}\mathbb{X}S_{X_{\mathcal{K}}}^{-1}\mathbb{X}^{\mathrm{T}} + n^{-1}(\mathbb{Z}-\mathbb{X}S_{X_{\mathcal{K}}}^{-1}S_{X_{\mathcal{K}}X_k})S_{R_k}^{-1}(\mathbb{Z}^{\mathrm{T}} - S_{X_kX_{\mathcal{K}}}S_{X_{\mathcal{K}}}^{-1}\mathbb{X}^{\mathrm{T}}), \\
&= P_{\mathbb{X}} + P_{\mathbb{Z}|\mathbb{X}},
\end{aligned}$$

where $S_{R_k} = n^{-1}\{\mathbb{Z}^{\mathrm{T}}\mathbb{Z} - \mathbb{Z}^{\mathrm{T}}\mathbb{X}(\mathbb{X}^{\mathrm{T}}\mathbb{X})^{-1}\mathbb{X}^{\mathrm{T}}\mathbb{Z}\}$ is the sample covariance of the fitted residual $(\mathbb{Z}-\mathbb{X}S_{X_{\mathcal{K}}}^{-1}S_{X_{\mathcal{K}}X_k})$. Specifically, $(\mathbb{Z}-\mathbb{X}S_{X_{\mathcal{K}}}^{-1}S_{X_{\mathcal{K}}X_k})^{\mathrm{T}}(\mathbb{Z}-\mathbb{X}S_{X_{\mathcal{K}}}^{-1}S_{X_{\mathcal{K}}X_k}) = nS_{R_k}$. Therefore, $P_{\mathbb{Z}|\mathbb{X}} \in \mathbb{R}^{n\times n}$ is indeed the projection matrix onto the one-dimensional subspace spanned by $(\mathbb{Z}-\mathbb{X}S_{X_{\mathcal{K}}}^{-1}S_{X_{\mathcal{K}}X_k})$. The conclusion follows from (S.4), (S.5) and $P_{(\mathbb{X},\mathbb{Z})} = P_{\mathbb{X}} + P_{\mathbb{Z}|\mathbb{X}}$.

S.5. STABILIZED ONE-STEP ESTIMATOR

S.5.1. *Recursive computation of the one-step estimator*

In the implementation, given a new observation O_{j+1}, we need to update the weight $w_j = \overline{\sigma}_n/\widehat{\sigma}_j$ and the variance of the canonical gradient $\widehat{\sigma}_j^2$ in order to update the current estimate ψ_j of the target parameter $\tau_{\max}$. Since we only need to know these quantities when all observations are included, an efficient approach is to exploit the following recursive properties of $\psi_j/\overline{\sigma}_j$ and $\overline{\sigma}_j^{-1}$:

$$\frac{\psi_{j+1}}{\overline{\sigma}_{j+1}} = \frac{1}{j+1}\left\{ j \cdot \frac{\psi_j}{\overline{\sigma}_j} + \frac{\Psi^{d_{nj}}(P_j) + D^{d_j}(P_j)(O_{j+1})}{\widehat{\sigma}_{nj}} \right\}, \quad \overline{\sigma}_{j+1}^{-1} = \frac{1}{j+1}\left(j \cdot \overline{\sigma}_j^{-1} + \widehat{\sigma}_{nj}^{-1} \right).$$

We then obtain the final estimate when $j = n$:

$$\widehat{\tau}_{\max} \equiv \psi_n = \frac{\overline{\sigma}_n}{n-\ell_n} \sum_{j=\ell_n}^{n-1} \left\{ \frac{\Psi^{d_{nj}}(P_j) + D^{d_j}(P_j)(O_{j+1})}{\widehat{\sigma}_{nj}} \right\}, \quad \overline{\sigma}_n^{-1} = \frac{1}{n-\ell_n} \sum_{j=\ell_n}^{n-1} \widehat{\sigma}_{nj}^{-1}.$$

S.5.2. *Proofs of Theorems* 2 *and* 3

The proofs are analogous to those of Theorems A.10 and A.11 in Luedtke & van der Laan (2018), which will henceforth be abbreviated to LL2018, establishing the asymptotic validity of the confidence interval (extending Theorem 1 of LL2018). In what follows, we first state various lemmas that together give all the ingredients we need.

Recall that in Theorems 2 and 3 we assume each variable X_k, $k = 1, \ldots, p$, or Y_j, $j = 1, \ldots, q$, takes values within $[-1, 1]$. The other conditions are as follows. The canonical gradient satisfies for some constant $\gamma > 0$

$$\inf_{n \geq 2} \min_{d \in \mathcal{D}_n} \mathrm{var}_{P_0}\{D^d(P_0)(O)\} \geq \gamma.$$

To ensure that the canonical gradient is uniformly bounded for all d, we assume that, for some $\delta > 0$,

$$\sup_{d \in \mathcal{D}_n} \max\{\|\Sigma_{X_\mathcal{K}}^{-1}\|, \|\Sigma_{Y_\mathcal{J}}^{-1}\|\} < \delta^{-1},$$

where $\|M\|$ denotes the operator norm. We treat δ, γ and $s = s_x = s_y$ as fixed when $n \to \infty$, and thus omit the dependence on δ, γ and s in the asymptotic statements. On the other hand, we allow both the dimensions p and q to grow quickly with the sample size n. We define $\beta_n^2 = n^{-1/2} \cdot \log \max\{n, p, q\}$ and, for any $\epsilon \in (0, 2)$, assume

$$\frac{\log \max(n,p,q)}{\ell_n} \to 0, \ \beta_n^2 \log \frac{n}{\ell_n} \to 0, \ \limsup_{n\to\infty} \frac{\ell_n}{n} < 1, \tag{S.6}$$

where

$$\ell_n = \max\left\{ (\log \max(n,p,q))^{1+\epsilon}, n \exp(-\beta_n^{-2+\epsilon}) \right\}.$$

For Theorem 3, we further assume the following margin condition. For some sequence $t_n \to \infty$, there exists a sequence of nonempty subsets $\mathcal{D}_n^\star \subseteq \mathcal{D}_n$ such that, for all n,

$$\sup_{d_1, d_2 \in \mathcal{D}_n^\star} \left\{ \Psi^{d_1}(P_0) - \Psi^{d_2}(P_0) \right\} = o(n^{-1/2}), \ \inf_{d \in \mathcal{D}_n^\star} \Psi^d(P_0) - \sup_{d \in \mathcal{D}_n \setminus \mathcal{D}_n^\star} \Psi^d(P_0) \geq t_n n^{-1/2} \beta_n.$$

Replacing X and Y above by $X_\mathcal{K}$ and $Y_\mathcal{J}$, the remainder term $\mathrm{Rem}^d(P)$ given by (S.3) can be bounded as follows.

LEMMA S.1. *For $d \in \mathcal{D}_n$, define*

$$M_1(P,d) \equiv \max_{j\in\mathcal{J},k\in\mathcal{K}}\{|\mathrm{E}_P(Y_j)-\mathrm{E}_{P_0}(Y_j)|,|\mathrm{E}_P(X_k)-\mathrm{E}_{P_0}(X_k)|\};$$
$$M_2(P,d) \equiv \max_{j\in\mathcal{J},k\in\mathcal{K}}\{|\mathrm{var}_P(Y_j)-\mathrm{var}_{P_0}(Y_j)|,|\mathrm{var}_P(X_k)-\mathrm{var}_{P_0}(X_k)|,|\mathrm{cov}_P(X_k,Y_j)-\mathrm{cov}_{P_0}(X_k,Y_j)|\}.$$

Suppose $\widetilde{\delta}^{-1}$ and λ are some positive constants such that, for all $d \in \mathcal{D}_n$,

$$\max\left(\|S_{X_\mathcal{K}}^{-1}\|,\|S_{Y_\mathcal{J}}^{-1}\|\right) < \widetilde{\delta}^{-1}, \quad \max\left\{\|\Sigma_{Y_\mathcal{J}}^{1/2}S_{Y_\mathcal{J}}^{-1}\Sigma_{Y_\mathcal{J}}^{1/2}\|,\|\Sigma_{X_\mathcal{K}}^{1/2}S_{X_\mathcal{K}}^{-1}\Sigma_{X_\mathcal{K}}^{1/2}\|\right\} < \lambda. \tag{S.7}$$

Then the second-order remainder term (S.2) *is bounded by*

$$\begin{aligned}|\mathrm{Rem}^d(P)| &< \widetilde{\delta}^{-1}2(s_x+s_y)M_1(P,d) + \{2\widetilde{\delta}^{-1}+(4+2\lambda^2)(\delta\widetilde{\delta})^{-1}\}s_xs_yM_2^2(P,d)\\ &\le \widetilde{\delta}^{-1}4sM_1(P,d) + \{2\widetilde{\delta}^{-1}+(4+2\lambda^2)(\delta\widetilde{\delta})^{-1}\}s^2M_2^2(P,d).\end{aligned}$$

Let $\mathcal{F}_n$ denote the following class of functions mapping $\mathbb{R}^\infty \times \mathbb{R}^\infty$ into the real line:

$$\left\{(x,y)\mapsto x_k^a y_j^b:\ 0\le a,b\le 4,\ a+b\le 4,\ k\in\{1,\ldots,p\},\ j\in\{1,\ldots,q\}\right\}. \tag{S.8}$$

Then $|\mathcal{F}_n| \lesssim pq$, where $\lesssim$ denotes less than or equal to up to a universal multiplicative constant.

We define the empirical process for $f \in \mathcal{F}_n$ and based on the first $j \in \{1,\ldots,n\}$ observations as

$$\mathbb{G}_{nj} \equiv j^{-1/2}\sum_{i=1}^{j}\{f(O_i)-P_0f\} = j^{1/2}(P_j-P_0)f,$$

where P_j is the empirical distribution of $O_1,\ldots,O_j$ and $Pf = \mathrm{E}_P\{f(O)\}$. Let $\|\mathbb{G}_{nj}\|_{\mathcal{F}_n} \equiv \sup_{f\in\mathcal{F}_n}|\mathbb{G}_{nj}|$, then by van der Vaart & Wellner (1996, Theorem 2.14.1)

$$\mathrm{E}\|\mathbb{G}_{nj}\|_{\mathcal{F}_n} \lesssim (\log pq)^{1/2} = (\log p+\log q)^{1/2} \le (2\log\max\{p,q\})^{1/2}.$$

For all $j=1,\ldots,n$, define the events

$$\mathcal{A}_{nj} = \left\{\max_{f\in\mathcal{F}_n}|(P_j-P_0)f| \le CK_{nj}\right\},\ \mathcal{A}_n = \bigcap_{j=1}^{n}\mathcal{A}_{nj}, \quad K_{nj} = j^{-1/2}(\log\max\{n,p,q\})^{1/2}, \tag{S.9}$$

where C in the definition is the smallest universal constant satisfying

$$\mathrm{E}\|\mathbb{G}_{nj}\|_{\mathcal{F}_n} \le (C-1)(2\log\max\{p,q\})^{1/2}.$$

LEMMA S.2. *For any sample size n, the event $\mathcal{A}_n$ occurs with probability at least $1-1/n$.*

Next we show that the various estimators are close to their population versions when $\mathcal{A}_n$ occurs.

LEMMA S.3. *For any fixed sample size $n \ge 2$, the event $\mathcal{A}_n$ implies the following statements for $j=2,\ldots,n$, and for any $d \in \mathcal{D}_n$,*

$$\max\{M_1(P_j,d),M_2(P_j,d)\} \lesssim K_{nj}, \tag{S.10}$$

*where $M_1(P,d)$ and $M_2(P,d)$ are defined in Lemma S.*1.

LEMMA S.4. *Let C be the smallest universal constant in* (S.10) *and let n be any natural number satisfying $n \ge \lceil 4C^2\delta^{-2}s^2\log\max\{n,p,q\}\rceil \equiv J(n,\delta)$. The occurrence of event $\mathcal{A}_n$*

implies that, for all $j = J(n,\delta), \ldots, n$*, for any* $d \in \mathcal{D}_n$*,*

$$\max\left\{\|\Sigma_{Y_\mathcal{J}}^{1/2} S_{Y_\mathcal{J}}^{-1} \Sigma_{Y_\mathcal{J}}^{1/2}\|, \|\Sigma_{X_\mathcal{K}}^{1/2} S_{X_\mathcal{K}}^{-1} \Sigma_{X_\mathcal{K}}^{1/2}\|\right\} \le 2; \tag{S.11}$$

$$\max\left(\|S_{X_\mathcal{K}}^{-1}\|, \|S_{Y_\mathcal{J}}^{-1}\|, \|S_{X_\mathcal{K}}^{-1} S_{X_\mathcal{K} Y_\mathcal{J}} S_{Y_\mathcal{J}}^{-1}\|, \|S_{X_\mathcal{K}}^{-1} S_{X_\mathcal{K} Y_\mathcal{J}} S_{Y_\mathcal{J}}^{-1} S_{Y_\mathcal{J} X_\mathcal{K}} S_{X_\mathcal{K}}^{-1}\|\right) \le 2\delta^{-1}; \tag{S.12}$$

$$|\widehat{D}_{nj}(o)| \le 2(s_x^{1/2} + s_y^{1/2})^2 \delta^{-1} \le 8\delta^{-1} s; \tag{S.13}$$

and the estimator $\tau^2(P_j)$ *of the Pillai trace satisfies*

$$|\tau^2 - \tau^2(P_j)| \equiv |\Psi_n(P_0) - \Psi^{d_{nj}}(P_j)| \lesssim \delta^{-2} s^2 K_{nj}. \tag{S.14}$$

LEMMA S.5. *Under the same conditions as in Lemma S.*4*, the occurrence of event* $\mathcal{A}_n$ *implies that, for all* $j = J(n,\delta), \ldots, n$*, for any* $d \in \mathcal{D}_n$*,*

$$|\widehat{\sigma}_{nj}^2 - \sigma_{nj}^2| \le \delta^{-2} s^2 K_{nj}; \tag{S.15}$$

$$|\widehat{\sigma}_{nj}^2 - \mathrm{var}_{P_0}\{D_{nj}(P_0)(O)\}| \le \delta^{-4} s^2 K_{nj} + \left(\widehat{\mathrm{Rem}}_{nj}\right)^2; \tag{S.16}$$

$$|\widehat{\mathrm{Rem}}_{nj}| \equiv |\mathrm{Rem}^{d_n}(P_j)| \lesssim \delta^{-1}(s_x + s_y) K_{nj}^2 + \delta^{-2} s_x s_y K_{nj}^2 \lesssim \left(\delta^{-1} s + \delta^{-2} s^2\right) K_{nj}^2. \tag{S.17}$$

LEMMA S.6. *Let* $\gamma > 0$ *be as defined in* (14)*. For a constant* $C(\delta, \gamma, s) > 0$ *relying on* γ*,* δ *and* s *only, the occurrence of event* $\mathcal{A}_n$ *implies that, for all* $j = \lceil C(\delta,\gamma,s) \log\max(n,p,q)\rceil, \ldots, n$*,*

$$\widehat{\sigma}_{nj}^2 \ge \gamma/2. \tag{S.18}$$

In what follows, we treat δ, γ and s as fixed and omit them in the asymptotic statements. Recall that

$$\beta_n^2 = n^{-1/2}\log(p+q), \quad \ell_n = \max\left\{(\log\max(n,p,q))^{1+\epsilon}, n\exp(-\beta_n^{-2+\epsilon})\right\}.$$

LEMMA S.7. *The following properties hold as* $n \to \infty$

(C1) There exists $M < \infty$ *such that*

$$\frac{1}{n-\ell_n}\sum_{j=\ell_n}^{n-1} P_0\left(\frac{|\widehat{D}_{nj}(O)|}{\widehat{\sigma}_{nj}^2} < M \mid O_0, \ldots, O_{j-1}\right) \to 1$$

in probability.

(C2) $(n-\ell_n)^{-1}\sum_{j=\ell_n}^{n-1}\left|\sigma_{nj}^2/\widehat{\sigma}_{nj}^2 - 1\right| \to 0$ *in probability.*

(C3) $(n-\ell_n)^{-1/2}\sum_{j=\ell_n}^{n-1}\widehat{\sigma}_{nj}^{-1}\widehat{\mathrm{Rem}}_{nj} \to 0$ *in probability, where* $\widehat{\mathrm{Rem}}_{nj}$ *is from Lemma S.*5*.*

S.5.3. *Proof of Lemma S.*1

Under the assumption (15), we have the following immediate implications for any $\mathcal{K}$ and $\mathcal{J}$:

1. $\|\Sigma_{X_\mathcal{K}}^{-1}\| < \delta^{-1}$ and $\|\Sigma_{Y_\mathcal{J}}^{-1}\| < \delta^{-1}$;
2. $\|\Sigma_{X_\mathcal{K}}^{-1/2}\| < \delta^{-1/2}$ and $\|\Sigma_{Y_\mathcal{J}}^{-1/2}\| < \delta^{-1/2}$;
3. $\|\Sigma_{X_\mathcal{K}}^{-1}\Sigma_{X_\mathcal{K} Y_\mathcal{J}}\Sigma_{Y_\mathcal{J}}^{-1}\| < \delta^{-1}$;

4. $\|\Sigma_{Y_\mathcal{J}}^{-1}\Sigma_{Y_\mathcal{J}X_\mathcal{K}}\Sigma_{X_\mathcal{K}}^{-1}\Sigma_{X_\mathcal{K}Y_\mathcal{J}}\Sigma_{Y_\mathcal{J}}^{-1}\| < \delta^{-1}$.

For the last two statements, we notice that $\|\Sigma_{X_\mathcal{K}}^{-1/2}\Sigma_{X_\mathcal{K}Y_\mathcal{J}}\Sigma_{Y_\mathcal{J}}^{-1/2}\| \leq 1$ because the singular values of $\Lambda_{XY} = \Sigma_X^{-1/2}\Sigma_{XY}\Sigma_Y^{-1/2}$ are no larger than 1.

The first three terms in (S.3), and the remainder term $\mathrm{Rem}^d(P)$, can be bounded as follows. For notational convenience, we omit the subscripts in $X_\mathcal{K}$ and $Y_\mathcal{J}$. Let $\delta_X = S_X^{-1/2}(\mu_{X,0} - \mu_X)$ and $\delta_Y = S_Y^{-1/2}(\mu_{Y,0} - \mu_Y)$, and note that the singular values of $C_{XY} = S_X^{-1/2}S_{XY}S_Y^{-1/2}$ are less than 1. Then the absolute value of the first three terms in (S.3) becomes

$$\begin{aligned}|2\delta_X^{\mathrm{T}}C_{XY}\delta_Y - \delta_Y^{\mathrm{T}}C_{XY}^{\mathrm{T}}C_{XY}\delta_Y - \delta_X^{\mathrm{T}}C_{XY}C_{XY}^{\mathrm{T}}\delta_X| &\leq 2|\delta_X^{\mathrm{T}}\delta_Y| + \|\delta_Y\|^2 + \|\delta_X\|^2\\ &\leq 2\|\delta_X\|\|\delta_Y\| + \|\delta_Y\|^2 + \|\delta_X\|^2\\ &< \widetilde{\delta}^{-1}\cdot(s_x + s_y + 2s_x^{1/2}s_y^{1/2})\cdot\{M_1(P,d)\}^2\\ &\leq \widetilde{\delta}^{-1}\cdot 4s\cdot\{M_1(P,d)\}^2.\end{aligned}$$

To bound the fourth term in (S.3),

$$\begin{aligned}\left|\mathrm{tr}\left\{S_Y^{-1}(\Sigma_{YX} - S_{YX})S_X^{-1}(S_{XY} - \Sigma_{XY})\right\}\right| &\leq \|S_X^{-1}\|\|S_Y^{-1}\|\|S_{XY} - \Sigma_{XY}\|_F^2\\ &< \widetilde{\delta}^{-2}s_xs_y\cdot\{M_2(P,d)\}^2\\ &\leq \widetilde{\delta}^{-2}s^2\cdot\{M_2(P,d)\}^2.\end{aligned}$$

Similarly, for the other terms in (S.3), we have

$$\begin{aligned}\|S_X^{-1}\otimes\Sigma_X^{-1}\Sigma_{XY}\Sigma_Y^{-1}\| &< (\widetilde{\delta}\delta)^{-1},\\ \|S_Y^{-1}S_{YX}S_X^{-1}\otimes\Sigma_X^{-1}\Sigma_{XY}\Sigma_Y^{-1}\| &< (\widetilde{\delta}\delta)^{-1},\\ \|S_Y^{-1}S_{YX}S_X^{-1}\otimes\Sigma_X^{-1}\| &< (\widetilde{\delta}\delta)^{-1},\\ \|S_X^{-1}S_{XY}S_Y^{-1}S_{YX}S_X^{-1}\otimes\Sigma_X^{-1}\| &< (\widetilde{\delta}\delta)^{-1},\\ \|S_X^{-1}S_{XY}S_Y^{-1}\otimes S_Y^{-1}\| &< \widetilde{\delta}^{-2}.\end{aligned}$$

Further, $\|S_Y^{-1}\otimes S_Y^{-1}\Sigma_{YX}S_X^{-1}\Sigma_{XY}\Sigma_Y^{-1}\|$ is bounded above by $\widetilde{\delta}^{-1}\|S_Y^{-1}\Sigma_{YX}S_X^{-1}\Sigma_{XY}\Sigma_Y^{-1}\|$, which equals

$$\begin{aligned}\widetilde{\delta}^{-1}\|\Sigma_Y^{-1/2}S_Y^{-1}\Sigma_Y^{1/2}\cdot\Sigma_Y^{-1/2}\Sigma_{YX}\Sigma_X^{-1/2}\cdot\Sigma_X^{1/2}S_X^{-1}\Sigma_X^{1/2}\cdot\Sigma_X^{-1/2}\Sigma_{XY}\Sigma_Y^{-1/2}\|&\\ \leq \widetilde{\delta}^{-1}\|\Sigma_Y^{-1/2}S_Y^{-1}\Sigma_Y^{1/2}\|\cdot\|\Sigma_X^{1/2}S_X^{-1}\Sigma_X^{1/2}\|&\\ \leq \widetilde{\delta}^{-1}\delta^{-1}\|\Sigma_Y^{1/2}S_Y^{-1}\Sigma_Y^{1/2}\|\cdot\|\Sigma_X^{1/2}S_X^{-1}\Sigma_X^{1/2}\|&\\ < (\widetilde{\delta}\delta)^{-1}\lambda^2.&\end{aligned}$$

The term $\|S_Y^{-1}\otimes S_Y^{-1}\Sigma_{YX}S_X^{-1}\|$ is bounded above by

$$\begin{aligned}\widetilde{\delta}^{-1}\|S_Y^{-1}\Sigma_Y^{1/2}\cdot\Sigma_Y^{-1/2}\Sigma_{YX}\Sigma_X^{-1/2}\cdot\Sigma_X^{1/2}S_X^{-1}\| &\leq \widetilde{\delta}^{-1}\|S_Y^{-1}\Sigma_Y^{1/2}\|\cdot\|\Sigma_X^{1/2}S_X^{-1}\|\\ &\leq \widetilde{\delta}^{-1}\delta^{-1}\|\Sigma_Y^{1/2}S_Y^{-1}\Sigma_Y^{1/2}\|\cdot\|\Sigma_X^{1/2}S_X^{-1}\Sigma_X^{1/2}\|\\ &< (\widetilde{\delta}\delta)^{-1}\lambda^2.\end{aligned}$$

Adding all the terms together completes the proof.

S.5.4. *Proof of Lemma S.2*

Omitted since it is analogous to the proof of Lemma A.4 in LL2018.

S.5.5. *Proof of Lemma S.*3

First recall the definitions of $M_1(P_j, d)$ and $M_2(P_j, d)$ given in Lemma S.1. By definition,

$$\max_{1\le k\le p} |\mathrm{E}_{P_j}(X_k) - \mathrm{E}_{P_0}(X_k)| \le \max_{f\in\mathcal{F}_n} |(P_j - P_0)f| \lesssim K_{nj}$$

holds for all $j = 2, \ldots, n$. Similarly,

$$\max_{1\le i\le q} |\mathrm{E}_{P_j}(Y_i) - \mathrm{E}_{P_0}(Y_i)| \lesssim K_{nj}, \quad \max_{1\le k\le p} |\mathrm{var}_{P_j}(X_k) - \mathrm{var}_{P_0}(X_k)| \lesssim K_{nj}$$

and $\max_{1\le i\le q} |\mathrm{var}_{P_j}(Y_i) - \mathrm{var}_{P_0}(Y_i)| \lesssim K_{nj}$. Therefore $M_1(P_j, d) \lesssim K_{nj}$ for all j and d. To see $M_2(P_j, d) \lesssim K_{nj}$, we notice that $|\mathrm{cov}_{P_j}(X_k, Y_i) - \mathrm{cov}_{P_0}(X_k, Y_i)|$ is bounded above by

$$|\mathrm{E}_{P_j}(X_kY_i) - \mathrm{E}_{P_0}(X_kY_i)| + |\mathrm{E}_{P_j}(X_k)| \cdot |(\mathrm{E}_{P_j} - \mathrm{E}_{P_0})Y_i| + |\mathrm{E}_{P_0}(Y_i)| \cdot |(\mathrm{E}_{P_j} - \mathrm{E}_{P_0})X_k| \lesssim K_{nj}.$$

S.5.6. *Proof of Lemma S.*4

First we show $\|\Sigma_{X_\mathcal{K}}^{1/2} S_{X_\mathcal{K}}^{-1} \Sigma_{X_\mathcal{K}}^{1/2}\| \le 2$. Note that

$$\begin{aligned}
\|\Sigma_{X_\mathcal{K}}^{1/2} S_{X_\mathcal{K}}^{-1} \Sigma_{X_\mathcal{K}}^{1/2}\| &= \| \left\{ \Sigma_{X_\mathcal{K}}^{-1/2} (\Sigma_{X_\mathcal{K}} + S_{X_\mathcal{K}} - \Sigma_{X_\mathcal{K}}) \Sigma_{X_\mathcal{K}}^{-1/2} \right\}^{-1} \| \\
&= \| \left\{ I_{s_n} + \Sigma_{X_\mathcal{K}}^{-1/2} (S_{X_\mathcal{K}} - \Sigma_{X_\mathcal{K}}) \Sigma_{X_\mathcal{K}}^{-1/2} \right\}^{-1} \| \\
&= \left\{ 1 - \|\Sigma_{X_\mathcal{K}}^{-1/2} (S_{X_\mathcal{K}} - \Sigma_{X_\mathcal{K}}) \Sigma_{X_\mathcal{K}}^{-1/2}\| \right\}^{-1} \\
&\le \left\{ 1 - \|\Sigma_{X_\mathcal{K}}^{-1}\| \|S_{X_\mathcal{K}} - \Sigma_{X_\mathcal{K}}\| \right\}^{-1}, \qquad \text{(S.19)}
\end{aligned}$$

where $\|\Sigma_{X_\mathcal{K}}^{-1}\| < \delta^{-1}$ by (15) and $\|S_{X_\mathcal{K}} - \Sigma_{X_\mathcal{K}}\| \le \|S_{X_\mathcal{K}} - \Sigma_{X_\mathcal{K}}\|_1 \le s_n C K_{nj}$ from (S.10). Moreover, $s_n C K_{nj} \le s_n C j^{-1/2} (\log \max\{n, p, q\})^{1/2}$. Then for

$$j \ge J(n, \delta) \ge 4C^2 \delta^{-2} \{\max(s_n, t_n)\}^2 \log \max\{n, p, q\}$$

we have

$$\|\Sigma_{X_\mathcal{K}}^{-1}\| \|S_{X_\mathcal{K}} - \Sigma_{X_\mathcal{K}}\| \le \delta^{-1} s_n C j^{-1/2} (\log \max\{n, p, q\})^{1/2} \le 1/2,$$

which immediately implies $\|\Sigma_{X_\mathcal{K}}^{1/2} S_{X_\mathcal{K}}^{-1} \Sigma_{X_\mathcal{K}}^{1/2}\| \le 2$ by (S.19). A similar argument shows that $\|\Sigma_{Y_\mathcal{J}}^{1/2} S_{Y_\mathcal{J}}^{-1} \Sigma_{Y_\mathcal{J}}^{1/2}\| \le 2$.

Next, to see $\|S_{X_\mathcal{K}}^{-1}\| \le 2\delta^{-1}$, note that

$$\|S_{X_\mathcal{K}}^{-1}\| \le \|S_{X_\mathcal{K}}^{-1} \Sigma_{X_\mathcal{K}}\| \|\Sigma_{X_\mathcal{K}}^{-1}\| = \|\Sigma_{X_\mathcal{K}}^{1/2} S_{X_\mathcal{K}}^{-1} \Sigma_{X_\mathcal{K}}^{1/2}\| \|\Sigma_{X_\mathcal{K}}^{-1}\| \le 2\delta^{-1}.$$

Similarly, $\|S_{Y_\mathcal{J}}^{-1}\| \le 2\delta^{-1}$. Then, $\|S_{X_\mathcal{K}}^{-1} S_{X_\mathcal{K} Y_\mathcal{J}} S_{Y_\mathcal{J}}^{-1}\| \le \|S_{X_\mathcal{K}}^{-1/2}\| \cdot \|S_{X_\mathcal{K}}^{-1/2} S_{X_\mathcal{K} Y_\mathcal{J}} S_{Y_\mathcal{J}}^{-1/2}\| \cdot \|S_{Y_\mathcal{J}}^{-1/2}\| \le \|S_{X_\mathcal{K}}^{-1/2}\| \cdot \|S_{Y_\mathcal{J}}^{-1/2}\| \le 2\delta^{-1}$, and similarly, $\|S_{X_\mathcal{K}}^{-1} S_{X_\mathcal{K} Y_\mathcal{J}} S_{Y_\mathcal{J}}^{-1} S_{Y_\mathcal{J} X_\mathcal{K}} S_{X_\mathcal{K}}^{-1}\| \le 2\delta^{-1}$.

To show $|\widehat{D}_{nj}(o)| \le 2(s_n^{1/2} + t_n^{1/2})^2 \delta^{-1}$, recall from (11) that

$$\begin{aligned}
\widehat{D}_{nj}(o) = &-\{x_\mathcal{K} - \mathrm{E}_P(X_\mathcal{K})\}^\mathrm{T} S_{X_\mathcal{K}}^{-1} S_{X_\mathcal{K} Y_\mathcal{J}} S_{Y_\mathcal{J}}^{-1} S_{Y_\mathcal{J} X_\mathcal{K}} S_{X_\mathcal{K}}^{-1} \{x_\mathcal{K} - \mathrm{E}_P(X_\mathcal{K})\} \\
&-\{y_\mathcal{J} - \mathrm{E}_P(Y_\mathcal{J})\}^\mathrm{T} S_{Y_\mathcal{J}}^{-1} S_{Y_\mathcal{J} X_\mathcal{K}} S_{X_\mathcal{K}}^{-1} S_{X_\mathcal{K} Y_\mathcal{J}} S_{Y_\mathcal{J}}^{-1} \{y_\mathcal{J} - \mathrm{E}_P(Y_\mathcal{J})\} \\
&+2\{y_\mathcal{J} - \mathrm{E}_P(Y_\mathcal{J})\}^\mathrm{T} S_{Y_\mathcal{J}}^{-1} S_{Y_\mathcal{J} X_\mathcal{K}} S_{X_\mathcal{K}}^{-1} \{x_\mathcal{K} - \mathrm{E}_P(X_\mathcal{K})\}.
\end{aligned}$$

Because of the bounded support of X_k and Y_i, we have that, under the event $\mathcal{A}_n$,

$$\begin{aligned}|\widehat{D}_{nj}(o)| &\lesssim s_x\|S_{X_\mathcal{K}}^{-1}S_{X_\mathcal{K}Y_\mathcal{J}}S_{Y_\mathcal{J}}^{-1}S_{Y_\mathcal{J}X_\mathcal{K}}S_{X_\mathcal{K}}^{-1}\| + s_y\|S_{Y_\mathcal{J}}^{-1}S_{Y_\mathcal{J}X_\mathcal{K}}S_{X_\mathcal{K}}^{-1}S_{X_\mathcal{K}Y_\mathcal{J}}S_{Y_\mathcal{J}}^{-1}\| \\ &\quad +2s_x^{1/2}s_y^{1/2}\|S_{Y_\mathcal{J}}^{-1}S_{Y_\mathcal{J}X_\mathcal{K}}S_{X_\mathcal{K}}^{-1}\| \\ &\le \left(s_x + s_y + 2s_x^{1/2}s_y^{1/2}\right)2\delta^{-1} \\ &\le 8\delta^{-1}s.\end{aligned}$$

Turning now to the Pillai trace, we write

$$\begin{aligned}|\tau_0^2 - \tau^2(P_j)| &= |\mathrm{tr}(\Sigma_{X_\mathcal{K}}^{-1}\Sigma_{X_\mathcal{K}Y_\mathcal{J}}\Sigma_{Y_\mathcal{J}}^{-1}\Sigma_{Y_\mathcal{J}X_\mathcal{K}} - S_{X_\mathcal{K}}^{-1}S_{X_\mathcal{K}Y_\mathcal{J}}S_{Y_\mathcal{J}}^{-1}S_{Y_\mathcal{J}X_\mathcal{K}})| \\ &\le |\mathrm{tr}\{\Sigma_{X_\mathcal{K}}^{-1}\Sigma_{X_\mathcal{K}Y_\mathcal{J}}\Sigma_{Y_\mathcal{J}}^{-1}(\Sigma_{Y_\mathcal{J}X_\mathcal{K}} - S_{Y_\mathcal{J}X_\mathcal{K}})\}| \\ &\quad +|\mathrm{tr}\{\Sigma_{X_\mathcal{K}}^{-1}\Sigma_{X_\mathcal{K}Y_\mathcal{J}}(\Sigma_{Y_\mathcal{J}}^{-1} - S_{Y_\mathcal{J}}^{-1})S_{Y_\mathcal{J}X_\mathcal{K}}\}| \\ &\quad +|\mathrm{tr}\{\Sigma_{X_\mathcal{K}}^{-1}(\Sigma_{Y_\mathcal{J}X_\mathcal{K}} - S_{Y_\mathcal{J}X_\mathcal{K}})S_{Y_\mathcal{J}}^{-1}S_{Y_\mathcal{J}X_\mathcal{K}}\}| \\ &\quad +|\mathrm{tr}\{(\Sigma_{X_\mathcal{K}}^{-1} - S_{X_\mathcal{K}}^{-1})S_{X_\mathcal{K}Y_\mathcal{J}}S_{Y_\mathcal{J}}^{-1}S_{Y_\mathcal{J}X_\mathcal{K}}\}|.\end{aligned}$$

We bound each of the four terms on the right hand side of the above inequality. For the first term,

$$\begin{aligned}|\mathrm{tr}\{\Sigma_{X_\mathcal{K}}^{-1}\Sigma_{X_\mathcal{K}Y_\mathcal{J}}\Sigma_{Y_\mathcal{J}}^{-1}(\Sigma_{Y_\mathcal{J}X_\mathcal{K}} - S_{Y_\mathcal{J}X_\mathcal{K}})\}| &\le \|\Sigma_{X_\mathcal{K}}^{-1}\Sigma_{X_\mathcal{K}Y_\mathcal{J}}\Sigma_{Y_\mathcal{J}}^{-1}\|_F \cdot \|\Sigma_{Y_\mathcal{J}X_\mathcal{K}} - S_{Y_\mathcal{J}X_\mathcal{K}}\|_F \\ &\lesssim \min(s_n,t_n)\|\Sigma_{X_\mathcal{K}}^{-1}\Sigma_{X_\mathcal{K}Y_\mathcal{J}}\Sigma_{Y_\mathcal{J}}^{-1}\| \cdot s_n^{1/2}t_n^{1/2}K_{nj} \\ &\le \{\max(s_n,t_n)\}^2\delta^{-1}K_{nj}.\end{aligned}$$

For the second term,

$$\begin{aligned}|\mathrm{tr}\{\Sigma_{X_\mathcal{K}}^{-1}\Sigma_{X_\mathcal{K}Y_\mathcal{J}}(\Sigma_{Y_\mathcal{J}}^{-1} - S_{Y_\mathcal{J}}^{-1})S_{Y_\mathcal{J}X_\mathcal{K}}\}| &= |\mathrm{tr}\{\Sigma_{X_\mathcal{K}}^{-1}\Sigma_{X_\mathcal{K}Y_\mathcal{J}}\Sigma_{Y_\mathcal{J}}^{-1}(S_{Y_\mathcal{J}} - \Sigma_{Y_\mathcal{J}})S_{Y_\mathcal{J}}^{-1}S_{Y_\mathcal{J}X_\mathcal{K}}\}| \\ &\le \|\Sigma_{X_\mathcal{K}}^{-1}\Sigma_{X_\mathcal{K}Y_\mathcal{J}}\Sigma_{Y_\mathcal{J}}^{-1}\| \cdot \|S_{Y_\mathcal{J}}^{-1}\| \cdot \|S_{Y_\mathcal{J}} - \Sigma_{Y_\mathcal{J}}\|_F \cdot \|S_{Y_\mathcal{J}X_\mathcal{K}}\|_F \\ &\lesssim \quad \delta^{-1}\cdot 2\delta^{-1}\cdot t_nK_{nj}\cdot s_n^{1/2}t_n^{1/2}, \\ &\lesssim \quad \delta^{-2}\cdot\{\max(s_n,t_n)\}^2\cdot K_{nj},\end{aligned}$$

where $\|S_{Y_\mathcal{J}}^{-1}\| \le 2\delta^{-1}$ by (S.12), and $\|S_{Y_\mathcal{J}X_\mathcal{K}}\|_F \lesssim s_n^{1/2}t_n^{1/2}$ is by the boundedness of X_k and Y_i. For the third term,

$$\begin{aligned}|\mathrm{tr}\{\Sigma_{X_\mathcal{K}}^{-1}(\Sigma_{Y_\mathcal{J}X_\mathcal{K}} - S_{Y_\mathcal{J}X_\mathcal{K}})S_{Y_\mathcal{J}}^{-1}S_{Y_\mathcal{J}X_\mathcal{K}}\}| &\le \|\Sigma_{X_\mathcal{K}}^{-1}\| \cdot \|S_{Y_\mathcal{J}}^{-1}\| \cdot \|\Sigma_{Y_\mathcal{J}X_\mathcal{K}} - S_{Y_\mathcal{J}X_\mathcal{K}}\|_F \cdot \|S_{Y_\mathcal{J}X_\mathcal{K}}\|_F \\ &\lesssim \delta^{-1}\cdot 2\delta^{-1}\cdot s_n^{1/2}t_n^{1/2}K_{nj}\cdot s_n^{1/2}t_n^{1/2} \\ &\lesssim \delta^{-2}\{\max(s_n,t_n)\}^2K_{nj}.\end{aligned}$$

For the fourth term,

$$\begin{aligned}|\mathrm{tr}\{(\Sigma_{X_\mathcal{K}}^{-1} - S_{X_\mathcal{K}}^{-1})S_{X_\mathcal{K}Y_\mathcal{J}}S_{Y_\mathcal{J}}^{-1}S_{Y_\mathcal{J}X_\mathcal{K}}\}| &= |\mathrm{tr}\{\Sigma_{X_\mathcal{K}}^{-1}(S_{X_\mathcal{K}} - \Sigma_{X_\mathcal{K}})S_{X_\mathcal{K}}^{-1}S_{X_\mathcal{K}Y_\mathcal{J}}S_{Y_\mathcal{J}}^{-1}S_{Y_\mathcal{J}X_\mathcal{K}}\}| \\ &\le \|\Sigma_{X_\mathcal{K}}^{-1}\| \cdot \|S_{X_\mathcal{K}} - \Sigma_{X_\mathcal{K}}\|_F \cdot \|S_{X_\mathcal{K}}^{-1}S_{X_\mathcal{K}Y_\mathcal{J}}S_{Y_\mathcal{J}}^{-1}S_{Y_\mathcal{J}X_\mathcal{K}}\|_F \\ &= \|\Sigma_{X_\mathcal{K}}^{-1}\| \cdot \|S_{X_\mathcal{K}} - \Sigma_{X_\mathcal{K}}\|_F \cdot \tau(P_j) \\ &\lesssim \delta^{-1}\cdot\{\max(s_n,t_n)\}^2\cdot K_{nj},\end{aligned}$$

where the last inequality is by the boundedness of the Pillai trace: $\tau^2(P_j) \le \mathrm{rank}\{C_{XY}(P_j)\} \le \min(s_n,t_n)$. Finally, we have

$$|\tau_0^2 - \tau^2(P_j)| \lesssim (\delta^{-1} + \delta^{-2})\{\max(s_n,t_n)\}^2K_{nj} \lesssim \delta^{-2}\{\max(s_n,t_n)\}^2K_{nj}.$$

S.5.7. *Proof of Lemma S.5*

The proof of $|\widehat{\sigma}^2_{nj} - \sigma^2_{nj}| \leq \delta^{-2}\{\max(s_n, t_n)\}^2 K_{nj}$ follows from the proof of Lemma A.6. in LL2018. The only difference lies in the factor $\{\max(s_n, t_n)\}^2 = \max(s_n t_n, s_n^2, t_n^2)$, which comes from the trace terms in $|(P_j - P_0)\widehat{D}_{nj}|$.

To show $|\widehat{\sigma}^2_{nj} - \mathrm{var}_{P_0}\{D_{nj}(P_0)(O)\}| \leq \delta^{-4}\{\max(s_n, t_n)\}^2 K_{nj} + \left(\widehat{\mathrm{Rem}}_{nj}\right)^2$, notice that $P_0\widehat{D}_{nj} = \tau_0^2 - \tau^2(P_j) + \widehat{\mathrm{Rem}}_{nj}$ and therefore

$$\left(P_0\widehat{D}_{nj}\right)^2 \leq \max_d \left\{\tau_0^2 - \tau^2(P_j)\right\}^2 + \left(\widehat{\mathrm{Rem}}_{nj}\right)^2.$$

The rest of the proof follows from (S.14) and the proof of Lemma A.6. in LL2018. Finally for $|\widehat{\mathrm{Rem}}_{nj}|$, we apply Lemma S.1 and note that $\widetilde{\delta} \geq \delta/2$ (S.11) and $\lambda \leq 2$ (S.12),

$$|\mathrm{Rem}^d(P)| < \delta^{-1}(s_n + t_n)CK_{nj} + (\delta^{-1} + 24\delta^{-2})s_n t_n C^2 K_{nj}^2.$$

S.5.8. *Proof of Lemma S.6*

From Lemma S.5,

$$|\widehat{\sigma}^2_{nj} - \mathrm{var}_{P_0}\{D_{nj}(P_0)(O)\}| \lesssim \delta^{-4}\max(s_n^2, t_n^2)K_{nj} + \delta^{-2}\max(s_n^2, t_n^2)K_{nj}^2. \tag{S.20}$$

It is easy to verify that, for a universal constant $C > 0$, the right hand side of the above inequality is bounded above by $\gamma/2$ for all j such that

$$j > C\gamma^{-1}\delta^{-2}\max\{\max(s_n^2, t_n^2), \gamma^{-2}\delta^{-6}\max(s_n^4, t_n^4)\}\log\max(n, p, q). \tag{S.21}$$

Since s_n and t_n are both greater than 1, we can let $C(\delta, \gamma) = C\gamma^{-1}\delta^{-2}\max\{1, \gamma^{-2}\delta^{-6}\}$. Then, if $j > C(\delta, \gamma)\max(s_n^4, t_n^4)\log\max(n, p, q)$, it must satisfy (S.21) and hence

$$|\widehat{\sigma}^2_{nj} - \mathrm{var}_{P_0}\{D_{nj}(P_0)(O)\}| \leq \gamma/2.$$

The result then follows from (14) and application of the triangle inequality.

S.5.9. *Proof of Lemma S.7*

Because we have fixed δ, γ and s as n grows, (S.6) implies that, for n large enough, ℓ_n is at least $J(n, \delta)$ and is at least $C(\delta, \gamma, s)\log\max(n, p, q)$, where these quantities are defined in Lemmas S.3 and S.6. We next verify properties (C1), (C2) and (C3) in Lemma S.7.

Property (C1): Because ℓ_n is at least $C(\delta, \gamma, s)\log\max(n, p, q)$, we have $\widehat{\sigma}^2_{nj} \geq \gamma/2$ from Lemma S.6. Because ℓ_n is at least $J(n, \delta)$, we have $|\widehat{D}_{nj}(o)| \leq 8\delta^{-1}s$ from Lemma S.4. For $j \geq \ell_n$ and under the event $\mathcal{A}_n$, we have $|\widehat{D}_{nj}(o)|/\widehat{\sigma}_{nj} \lesssim s\delta^{-1}\gamma^{-1/2}$. By Lemma S.2, $\mathcal{A}_n$ occurs with probability at least $1 - 1/n$. Properties (C2) and (C3) can be verified in the same way as in Theorem A.10 of LL2018 and noting that $|\widehat{\mathrm{Rem}}_{nj}| \lesssim K_{nj}^2$ and $|\widehat{\sigma}^2_{nj} - \sigma^2_{nj}| \lesssim K_{nj}$.

REFERENCES

DICICCIO, C. J. & ROMANO, J. P. (2017). Robust permutation tests for correlation and regression coefficients. *Journal of the American Statistical Association* **112**, 1211–1220.

DONOHO, D. & JIN, J. (2004). Higher criticism for detecting sparse heterogeneous mixtures. *The Annals of Statistics* **32**, 962–994.

DONOHO, D. & JIN, J. (2015). Higher criticism for large-scale inference, especially for rare and weak effects. *Statistical Science* **30**, 1–25.

KHIM, J. T., JOG, V. & LOH, P.-L. (2016). Computing and maximizing influence in linear threshold and triggering models. In *Advances in Neural Information Processing Systems*.

KRAUSE, A. & GOLOVIN, D. (2014). Submodular function maximization. In *Tractability: Practical Approaches to Hard Problems*. Cambridge University Press.

LUEDTKE, A. R. & VAN DER LAAN, M. (2018). Parametric-rate inference for one-sided differentiable parameters. *Journal of the American Statistical Association* **113**, 780–788.

NEMHAUSER, G. L. & WOLSEY, L. A. (1978). Best algorithms for approximating the maximum of a submodular set function. *Mathematics of Operations Research* **3**, 177–188.

VAN DER VAART, A. W. & WELLNER, J. A. (1996). *Weak Convergence and Empirical Processes with Applications to Statistics*. New York, NY, USA: Springer.

WANG, J. (2015). Joint estimation of sparse multivariate regression and conditional graphical models. *Statistica Sinica* **25**, 831–851.