Supplement to "A Fast and Accurate Estimator for Large Scale Linear Model via Data Averaging"

Rui Wang
Center for Applied Statistics
and School of Statistics
Renmin University of China
Beijing 100872, China 446100240@qq.com

Yanyan Ouyang
Center for Applied Statistics
and School of Statistics
Renmin University of China
Beijing 100872, China staoyyy@ruc.edu.cn

Panpan Yu
NavInfo
Beijing 100094, China
yupanpan@navinfo.com

Wangli Xu
Center for Applied Statistics
and School of Statistics
Renmin University of China
Beijing 100872, China wlxu@ruc.edu.cn

In this Supplementary Material, we provide numerical results and proofs.

A Additional numerical results

A.1 Simulation results

Tables A.1-A.3 list additional simulation results which are mentioned in the main text.

A.2 Real data example

In this section, we illustrate the performance of the proposed estimator based on a real dataset, namely Beijing multi-site air-quality data collected by Zhang et al. [2017]. This dataset contains hourly air pollutants data collected by 12 air quality monitoring sites in Beijing from March 1st, 2013 to February 28th, 2017. After discarding the observations with mising values, there are $382,168$ observations in total.

One important index of air quality is PM2.5, the concentration of particulate matter with aerodynamic diameter of no more than $2.5\,\mu\mathrm{m}$. We would like to fit linear models to predict PM2.5 concentration ($\mu\mathrm{g\,m^{-3}}$) under various conditions. Our model includes the following 8 continuous predictors: PM10 ($\mu\mathrm{g\,m^{-3}}$); SO2 ($\mu\mathrm{g\,m^{-3}}$); NO2 ($\mu\mathrm{g\,m^{-3}}$); CO ($\mu\mathrm{g\,m^{-3}}$); O3 ($\mu\mathrm{g\,m^{-3}}$); temperature ($^\circ\mathrm{C}$) pressure (hPa); dew point temperature ($^\circ\mathrm{C}$). The proposed algorithm is designed for continuous predictors. To break the ties of the predictors, we add independent $\mathcal{N}(0,10^{-16})$ noises to all predictors. For large scale real data, it is rare that the data follows a linear regression model strictly. Hence we only use the predictors in this real data example. And y_i is generated as follows: We compute the least square estimator based on the original full data, and use it as the ground truth of $\boldsymbol{\beta}$. Then we define $y_i = X_i^\top \boldsymbol{\beta} + \epsilon_i$ where $\epsilon_i \sim \mathcal{N}(0,1)$. We randomly select half of the observations to fit the model. For competing sketching methods, we take $n = \lfloor N/p \rfloor$ where N is the sample size for model fitting. The above procedure is independently repeated $1,000$ times. The resulting empirical mean squared errors for NEW, VDA, UNI, SRHT, LEV, IBOSS and FULL methods are 782.516, $4,902,110$, $1,149.22$, $1,224.03$, $1,255.47$, 920.734 and 84.2843, respectively. The results show that the proposed method has better statistical performance than the competing sketching methods.

37th Conference on Neural Information Processing Systems (NeurIPS 2023).

Table A.1: Empirical mean squared errors (multiplied by 10^3) of various algorithms with $N = 8\times10^4$ and $\varepsilon_1 \sim (\chi^2(1)-1)/\sqrt{2}$.

	p	NEW	VDA	UNI	SRHT	LEV	IBOSS	FULL
Case 1	50	176.753	16358.3	490.582	490.759	488.958	498.165	9.42386
	100	666.326	15133.9	2164.76	2175.63	2079.18	2125.49	18.1874
	200	2587.92	14346.7	14410.1	16360.5	15540.7	15653.2	38.9494
Case 2	50	7.1498	1076.89	33.6416	33.3228	33.534	25.283	0.626021
	100	22.4851	1010.59	145.947	146.255	145.517	121.719	1.2829
	200	69.4755	994.376	1018.37	1023.84	1023.54	915.626	2.48547
Case 3	50	4.8162	806.811	23.3352	27.2343	36.3333	11.866	0.592169
	100	10.6642	727.343	131.172	105.01	177.892	55.106	0.929276
	200	19.472	827.801	924.311	722.072	782.354	218.543	2.00839
Case 4	50	1.96697	349.656	11.988	11.6585	11.6786	2.28837	0.222879
	100	4.67671	339.694	54.7945	49.4988	50.7473	7.90413	0.434104
	200	11.1357	336.907	383.744	339.625	355.49	40.2336	0.864697
Case 5	50	19.5341	2013.35	64.4977	63.4975	65.4009	55.5417	1.25583
	100	61.9263	2012.47	275.38	286.032	271.544	265.213	2.47587
	200	194.656	2007.65	1946.9	1962.62	2037.76	1913.67	5.05518
Case 6	50	20.2014	2001.41	64.3295	63.2757	62.5513	56.1434	1.27484
	100	63.9794	2077.95	287.962	283.532	284.435	267.821	2.44736
	200	197.052	2049.1	2054.34	2007.75	2017.92	1927.66	4.91178

B Technical lemmas

Lemma B.1. *For $x > 0$, we have*

$$1-\Phi(x) \le \exp(-x^2/2), \tag{B.1}$$

$$(x^{-1}-x^{-3})\varphi(x) \le 1-\Phi(x) \le x^{-1}\varphi(x). \tag{B.2}$$

The inequality (B.1) can be proved by the Cramér-Chernoff bound; see Boucheron et al. [2013], Section 2.2. See Durrett [2019], Theorem 1.2.6 for a proof of the inequality (B.2).

Lemma B.2. *(Weyl's inequality) Let $\mathbf{B}_1$ and $\mathbf{B}_2$ be two symmetric $m \times m$ matrices. Then* $\max_{j\in\{1,\ldots,p\}} |\lambda_j(\mathbf{B}_1) - \lambda_j(\mathbf{B}_2)| \le \|\mathbf{B}_1 - \mathbf{B}_2\|$.

See, e.g., Theorem 8.1 of Bhatia [2007b].

Lemma B.3. *Let $\mathbf{B}_1$ and $\mathbf{B}_2$ be two matrices. Then*

$$\|\mathbf{B}_1\mathbf{B}_1^\top - \mathbf{B}_2\mathbf{B}_2^\top\| \le (2\|\mathbf{B}_2\| + \|\mathbf{B}_1-\mathbf{B}_2\|)\|\mathbf{B}_1-\mathbf{B}_2\|.$$

Proof.

$$\begin{aligned}\|\mathbf{B}_1\mathbf{B}_1^\top - \mathbf{B}_2\mathbf{B}_2^\top\| =& \|\mathbf{B}_1\mathbf{B}_1^\top - \mathbf{B}_1\mathbf{B}_2^\top + \mathbf{B}_1\mathbf{B}_2^\top - \mathbf{B}_2\mathbf{B}_2^\top\| \\ \le& (\|\mathbf{B}_1\| + \|\mathbf{B}_2\|)\|\mathbf{B}_1-\mathbf{B}_2\| \\ \le& (2\|\mathbf{B}_2\| + \|\mathbf{B}_1-\mathbf{B}_2\|)\|\mathbf{B}_1-\mathbf{B}_2\|.\end{aligned}$$

□

Lemma B.4. *Let $\mathbf{B}_1$ and $\mathbf{B}_2$ be two $m \times m$ symmetric matrices and $\mathbf{B}_2$ is positive definite. Suppose $\|\mathbf{B}_2^{-1}\|\|\mathbf{B}_1-\mathbf{B}_2\| < 1$. Then $\mathbf{B}_1$ is positive definite and*

$$\|\mathbf{B}_1^{-1} - \mathbf{B}_2^{-1}\| \le \frac{\|\mathbf{B}_2^{-1}\|^2\|\mathbf{B}_1-\mathbf{B}_2\|}{(1-\|\mathbf{B}_2^{-1}\|\|\mathbf{B}_1-\mathbf{B}_2\|)}.$$

Table A.2: Empirical mean squared errors (multiplied by 10^3) of various algorithms with $N = 6.4 \times 10^5$ and $\varepsilon_1 \sim \mathcal{N}(0, 1)$.

	p	NEW	VDA	UNI	SRHT	LEV	IBOSS	FULL
Case 1	100	81.9272	15770.9	239.974	230.171	241.03	225.048	2.41947
	200	324.856	15365.4	1020.77	974.136	1000.16	945.015	4.6511
	400	1255.91	15436.7	5093.1	4941.02	5110.64	5033.94	9.97678
Case 2	100	2.73287	991.157	16.3687	16.2829	16.0081	13.6101	0.161021
	200	8.76771	1022.71	67.5424	67.0934	67.4559	61.8691	0.313399
	400	29.7099	994.329	334.472	332.531	328.551	315.948	0.633274
Case 3	100	1.95374	735.115	10.2444	12.6629	22.6209	4.81436	0.129786
	200	3.78109	745.755	57.8812	58.1879	52.8285	22.6675	0.274122
	400	8.09969	724.26	267.593	283.055	278.49	78.4843	0.483496
Case 4	100	0.603808	343.914	5.59336	5.46564	5.58407	0.758669	0.0531629
	200	1.53723	344.041	23.4118	22.8189	22.85	2.69151	0.107645
	400	3.6276	337.205	117.696	111.846	110.7	10.7265	0.212023
Case 5	100	7.71805	2018.53	31.5943	31.5477	31.3756	29.4129	0.308866
	200	24.4975	2015.42	132.541	132.42	132.778	130.563	0.611402
	400	80.2474	2004.28	667.938	661.453	660.191	646.256	1.25611
Case 6	100	7.98704	1974.66	31.5267	31.4872	31.1946	29.2168	0.308048
	200	24.5327	1981.99	132.369	134.215	133.191	128.187	0.630886
	400	81.4014	1997.83	672.68	672.385	676.376	658.253	1.24475

Table A.3: Empirical mean squared errors (multiplied by 10^3) of various algorithms with $N = 6.4 \times 10^5$ and $\varepsilon_1 \sim (\chi^2(1) - 1)/\sqrt{2}$.

	p	NEW	VDA	UNI	SRHT	LEV	IBOSS	FULL
Case 1	100	82.5572	15073.2	238.427	237.482	227.898	227.843	2.33521
	200	335.829	14593	999.549	985.468	1029.66	997.179	4.7406
	400	1242.44	14931.6	4948.67	5190.31	5127.05	5031.33	9.22342
Case 2	100	2.76676	1020.6	15.4655	16.1021	16.0937	14.0472	0.160336
	200	9.03052	1001.22	66.944	66.5285	66.1458	59.9584	0.308678
	400	29.4599	996.876	333.189	337.005	338.663	315.418	0.619093
Case 3	100	1.15453	663.173	15.0485	12.8247	16.1937	5.18835	0.116118
	200	3.64909	836.29	50.7558	64.229	53.7309	19.939	0.305421
	400	7.93885	794.178	316.611	229.983	262.708	100.936	0.501982
Case 4	100	0.612663	346.966	5.60731	5.46467	5.40086	0.804744	0.0538384
	200	1.53891	340.407	23.3167	22.7071	22.8374	2.73583	0.106413
	400	3.56156	334.397	118.763	112.567	111.384	10.943	0.210416
Case 5	100	7.92053	1969.97	31.8714	31.5303	31.5133	29.3753	0.315144
	200	24.7695	2034.25	138.096	134.546	134.965	128.695	0.624047
	400	80.556	1990.13	672.685	668.159	655.674	631.659	1.25941
Case 6	100	7.82976	2053.86	32.4805	31.3492	31.3302	28.4122	0.309881
	200	24.2661	1999.75	136.008	131.905	133.408	125.27	0.620036
	400	80.16	2019.01	664.25	667.027	659.313	658.417	1.24739

Proof. Since $\mathbf{B}_2$ is positive definite, we have $\|\mathbf{B}_2^{-1}\| = 1/\lambda_m(\mathbf{B}_2)$. From Lemma B.2, $|\lambda_m(\mathbf{B}_1) - \lambda_m(\mathbf{B}_2)| \le \|\mathbf{B}_1 - \mathbf{B}_2\| < \lambda_m(\mathbf{B}_2)$. Consequently, $\lambda_m(\mathbf{B}_1) > 0$ and $\mathbf{B}_1$ is positive definite. We have

$$\begin{aligned}\|\mathbf{B}_1^{-1} - \mathbf{B}_2^{-1}\| =& \|\mathbf{B}_1^{-1}(\mathbf{B}_1 - \mathbf{B}_2)\mathbf{B}_2^{-1}\| \\ \le& \|\mathbf{B}_1^{-1}\|\|\mathbf{B}_2^{-1}\|\|\mathbf{B}_1 - \mathbf{B}_2\| \\ \le& (\|\mathbf{B}_1^{-1} - \mathbf{B}_2^{-1}\| + \|\mathbf{B}_2^{-1}\|)\|\mathbf{B}_2^{-1}\|\|\mathbf{B}_1 - \mathbf{B}_2\|.\end{aligned}$$

Consequently, $(1 - \|\mathbf{B}_2^{-1}\|\|\mathbf{B}_1 - \mathbf{B}_2\|)\|\mathbf{B}_1^{-1} - \mathbf{B}_2^{-1}\| \le \|\mathbf{B}_2^{-1}\|^2\|\mathbf{B}_1 - \mathbf{B}_2\|$. And the conclusion follows. □

Lemma B.5. *Let $\mathbf{A}$ be an $n \times p$ random matrix whose entries are independent standard normal random variables. Then for any $t \ge 0$,*

$$\Pr\left(\|\frac{1}{n}\mathbf{A}^\top\mathbf{A} - \mathbf{I}_p\| > 3\max\left((p/n)^{1/2} + (2t/n)^{1/2},\, \left((p/n)^{1/2} + (2t/n)^{1/2}\right)^2\right)\right) \le 2\exp(-t).$$

Lemma B.5 is a direct corollary of Vershynin [2010], Corollary 5.35 and Lemma 5.36.

Lemma B.6. *Let ξ be a random variable with density function with respect to the Lebesgue measure. Let $d > 0$ be a fixed number. Define set $\mathscr{A} = \{|\xi| > d\}$. Then for any set $\mathscr{B}$ with $\Pr(\mathscr{B}) \le \Pr(\mathscr{A})$, we have*

$$\mathrm{E}(|\xi|\mathbf{1}_{\mathscr{B}}) \le \mathrm{E}(|\xi|\mathbf{1}_{\mathscr{A}}).$$

Proof. We have

$$\mathrm{E}(|\xi|\mathbf{1}_{\mathscr{B}}) = \mathrm{E}(|\xi|\mathbf{1}_{\mathscr{B}\cap\mathscr{A}}) + \mathrm{E}(|\xi|\mathbf{1}_{\mathscr{B}\cap\mathscr{A}^\complement}) \le \mathrm{E}(|\xi|\mathbf{1}_{\mathscr{B}\cap\mathscr{A}}) + d\Pr(\mathscr{B}\cap\mathscr{A}^\complement).$$

But

$$\begin{aligned}d\Pr(\mathscr{B}\cap\mathscr{A}^\complement) =& d\left(\Pr(\mathscr{B}) - \Pr(\mathscr{B}\cap\mathscr{A})\right) \\ \le& d\left(\Pr(\mathscr{A}) - \Pr(\mathscr{B}\cap\mathscr{A})\right) = d\Pr(\mathscr{A}\cap\mathscr{B}^\complement) \le \mathrm{E}(|\xi|\mathbf{1}_{\mathscr{A}\cap\mathscr{B}^\complement}).\end{aligned}$$

Combining the above two inequalities leads to the conclusion. □

Lemma B.7. *Suppose $\xi_1, \ldots, \xi_n$ are i.i.d. random variables with continuous distribution function $F_\xi(x)$. Let $\xi_{(1)} \le \cdots \le \xi_{(n)}$ be the order statistics of $\xi_1, \ldots, \xi_n$. Then for $k \in \{1, \ldots, n\}$, $F_\xi(\xi_{(k)}) \sim \mathrm{Beta}(k, n-k+1)$. And for $1 \le k_1 < k_2 \le n$, $F_\xi(\xi_{(k_2)}) - F_\xi(\xi_{(k_1)}) \sim \mathrm{Beta}(k_2 - k_1, n - (k_2 - k_1) + 1)$.*

Proof. Since $F_\xi(x)$ is continuous, $F_\xi(\xi_1), \ldots, F_\xi(\xi_n)$ are i.i.d. random variables uniformly distributed on the interval $(0, 1)$. Then the conclusion follows from the fact that the order statistics of uniform random variables and their differences have beta distributions; see, e.g., Arnold et al. [2008], Section 2.5. □

C Conditional distributions of the selected observations

We summarize the IBOSS algorithm of Wang et al. [2019] in Algorithm 1. In this section, we study the conditional distributions of the selected observations in Algorithm 1. Note that selection procedure of Algorithm 1 is equivalent to the selection procedure of Algorithm 1 with $n = N$. Hence the results in this section also hold for Algorithm 1. In Algorithm 1, the distributions of the selected covariates $\{Z_i\}_{i\in\mathscr{I}}$ are typically not independent nor identically distributed. Nevertheless, the conditional distribution of $\{Z_i\}_{i\in\mathscr{I}}$ given the thresholds $\{\gamma_{1,j}, \gamma_{2,j}\}_{j=1}^p$ has a relatively simple characterization. To state this characterization, we need some further notations. For $j = 1, \ldots, p$, let $s(1,j)$ and $s(2,j)$ be the indices of $\gamma_{1,j}$ and $\gamma_{2,j}$ in $\{z_{i,j}\}_{i=1}^N$, respectively. That is, $z_{s(1,j),j} = \gamma_{1,j}$ and $z_{s(2,j),j} = \gamma_{2,j}$. Let $Z_{1,j}^L, \ldots, Z_{r-1,j}^L$ be a uniformly random alignment of $\{Z_i : i \in \mathscr{L}_{r,j},\, i \ne s(1,j)\}$, $Z_{1,j}^R, \ldots, Z_{r-1,j}^R$ be a uniformly random alignment of $\{Z_i : i \in \mathscr{R}_{r,j},\, i \ne s(1,j)\}$, and $Z_{1,j}^M, \ldots, Z_{N-2rj,j}^M$ be a uniformly random alignment of $\{Z_i : i \notin \bigcup_{\ell=1}^j \mathscr{L}_{r,\ell} \cup \mathscr{R}_{r,\ell}\}$, $j = 1, \ldots, p$. Note that $Z_{1,j}^M, \ldots, Z_{N-2rj,j}^M$ are the observations left by the IBOSS algorithm in step j. We have the following theorem.

Algorithm 1: The IBOSS algorithm in Wang et al. [2019]

Input: Observations $\{Z_i, y_i\}_{i=1}^N$, covariate dimension p, subdata sample size n
Output: Estimator of $\boldsymbol{\beta}$
$r = \lfloor \frac{n}{2p} \rfloor$
$\mathscr{I} \leftarrow \emptyset$
for $j \in \{1, ..., p\}$ **do**
$\gamma_{1,j} \leftarrow$ the rth smallest element of $\{z_{i,j} : i \in \{1, \ldots, N\} \backslash \big(\bigcup_{\ell=1}^{j-1} (\mathscr{L}_{r,\ell} \cup \mathscr{R}_{r,\ell}) \big) \}$
$\gamma_{2,j} \leftarrow$ the rth largest element of $\{z_{i,j} : i \in \{1, \ldots, N\} \backslash \big(\bigcup_{\ell=1}^{j-1} (\mathscr{L}_{r,\ell} \cup \mathscr{R}_{r,\ell}) \big) \}$
$\mathscr{L}_{r,j} \leftarrow \{i \in \{1, \ldots, N\} \backslash \big(\bigcup_{\ell=1}^{j-1} (\mathscr{L}_{r,\ell} \cup \mathscr{R}_{r,\ell}) \big) : z_{i,j} \le \gamma_{1,j} \}$
$\mathscr{R}_{r,j} \leftarrow \{i \in \{1, \ldots, N\} \backslash \big(\bigcup_{\ell=1}^{j-1} (\mathscr{L}_{r,\ell} \cup \mathscr{R}_{r,\ell}) \big) : z_{i,j} \ge \gamma_{2,j} \}$
$\mathscr{I}_j \leftarrow \mathscr{L}_{r,j} \cup \mathscr{R}_{r,j}$
$\mathscr{I} \leftarrow \mathscr{I} \cup \mathscr{I}_j$
$\hat{\boldsymbol{\beta}}_{\mathrm{I}} \leftarrow (\sum_{i \in \mathscr{I}} X_i X_i^\top)^{-1} (\sum_{i \in \mathscr{I}} X_i y_i)$
return $\hat{\boldsymbol{\beta}}_{\mathrm{I}}$

Theorem C.1. *Suppose Assumption 1 holds. Given $\{\gamma_{1,j}, \gamma_{2,j}\}_{j=1}^p$, the conditional distributions of the random vectors $Z_{i,j}^L$, $Z_{i,j}^R$, $i = 1, \ldots, r-1$, $j = 1, \ldots, p$, and $Z_{i,p}^M$, $i = 1, \ldots, N - 2rp$, are mutually independent. The conditional density of $Z_{i,j}^L$ is*

$$f_{Z_{i,j}^L | \{\gamma_{1,j}, \gamma_{2,j}\}_{j=1}^p}(Z) = \frac{f(z_1, \ldots, z_p) (\prod_{\ell=1}^{j-1} \mathbf{1}_{\{\gamma_{1,\ell} < z_\ell < \gamma_{2,\ell}\}}) \mathbf{1}_{\{z_j < \gamma_{1,j}\}}}{\int_{\mathbb{R}^p} f(z_1, \ldots, z_p) (\prod_{\ell=1}^{j-1} \mathbf{1}_{\{\gamma_{1,\ell} < z_\ell < \gamma_{2,\ell}\}}) \mathbf{1}_{\{z_j < \gamma_{1,j}\}} \, \mathrm{d}Z}; \tag{C.3}$$

the conditional density of $Z_{i,j}^R$ is

$$f_{Z_{i,j}^R | \{\gamma_{1,j}, \gamma_{2,j}\}_{j=1}^p}(Z) = \frac{f(z_1, \ldots, z_p) (\prod_{\ell=1}^{j-1} \mathbf{1}_{\{\gamma_{1,\ell} < z_\ell < \gamma_{2,\ell}\}}) \mathbf{1}_{\{z_j > \gamma_{2,j}\}}}{\int_{\mathbb{R}^p} f(z_1, \ldots, z_p) (\prod_{\ell=1}^{j-1} \mathbf{1}_{\{\gamma_{1,\ell} < z_\ell < \gamma_{2,\ell}\}}) \mathbf{1}_{\{z_j > \gamma_{2,j}\}} \, \mathrm{d}Z}; \tag{C.4}$$

and the conditional density of $Z_{i,p}^M$ is

$$f_{Z_{i,p}^M | \{\gamma_{1,j}, \gamma_{2,j}\}_{j=1}^p}(Z) = \frac{f(z_1, \ldots, z_p) \prod_{\ell=1}^{p} \mathbf{1}_{\{\gamma_{1,\ell} < z_\ell < \gamma_{2,\ell}\}}}{\int_{\mathbb{R}^p} f(z_1, \ldots, z_p) \prod_{\ell=1}^{p} \mathbf{1}_{\{\gamma_{1,\ell} < z_\ell < \gamma_{2,\ell}\}} \, \mathrm{d}Z}. \tag{C.5}$$

To prove Theorem C.1, we need some preparations. It is known that, for independent and identically distributed random variables, given certain order statistics, the distributions of the remaining variables are the original distribution truncated at the given order statistics; see, e.g., Arnold et al. [2008], Section 2.4. A similar result also holds for random vectors, and will be the basis of the proof of Theorem C.1. To state this result, we need to introduce some additional notations.

Fix an integer $j \in \{1, \ldots, p\}$. Let $\tau(i) = \sum_{\ell=1}^N \mathbf{1}_{\{z_{\ell,j} \le z_{i,j}\}}$ be the rank of $z_{i,j}$ in $\{z_{i,j}\}_{i=1}^N$. It can be seen that $z_{(\tau(i)),j} = z_{i,j}$, $i = 1, \ldots, N$. Consequently, $z_{(i),j} = z_{\tau^{-1}(i),j}$, $i = 1, \ldots, N$, where τ^{-1} is the inverse map of τ. Thus, $Z_{\tau^{-1}(1)}, \ldots, Z_{\tau^{-1}(N)}$ is a rearrangement of $Z_1, \ldots, Z_N$ such that their jth elements are in increasing order.

On the other hand, let $k \in \{1, \ldots, N\}$ and $1 \le m_1 < \cdots < m_k \le N$, $m_0 = 0$, $m_{k+1} = N + 1$. Let $\{\mathscr{G}_\ell\}_{\ell=1}^{k+1}$ be permutation groups on $\{1, \ldots, N\}$ where $\mathscr{G}_\ell$ is the collection of permutation π_ℓ such that π_ℓ maps the set $\{m_{\ell-1} + 1, \ldots, m_\ell - 1\}$ onto itself and $\pi_\ell(i) = i$ for $i \notin \{m_{\ell-1} + 1, \ldots, m_\ell - 1\}$. We define group $\mathscr{G} = \{\pi_1 \circ \cdots \circ \pi_k : \pi_\ell \in \mathscr{G}_\ell, \ell = 1, \ldots, k\}$. Thus, for any permutation π in $\mathscr{G}$, we have $\pi(m_\ell) = m_\ell$, $\ell = 1, \ldots, k$, and π maps the set $\{m_{\ell-1} + 1, \ldots, m_\ell - 1\}$ onto itself. The following lemma gives the distribution of $Z_{\tau^{-1} \circ \pi(1)}, \ldots, Z_{\tau^{-1} \circ \pi(N)}$ where π is a random permutation uniformly sampled from $\mathscr{G}$.

Lemma C.8. *Suppose Assumption 1 holds, π is a random permutation uniformly sampled from $\mathscr{G}$. Then given $Z_{\tau^{-1}(m_1)}, \ldots Z_{\tau^{-1}(m_k)}$, the random vectors $\{Z_{\tau^{-1} \circ \pi(i)}\}_{i \in \{1, \ldots, N\} \backslash \{m_1, \ldots, m_k\}}$ are independent, and for $m_{\ell-1} < i < m_\ell$, the conditional distribution of $Z_{\tau^{-1} \circ \pi(i)}$ given $Z_{\tau^{-1}(m_1)}$, $\ldots$, $Z_{\tau^{-1}(m_k)}$ has density function*

$$f_{Z_{\tau^{-1} \circ \pi(i)} | Z_{\tau^{-1}(m_1)}, \ldots, Z_{\tau^{-1}(m_k)}}(Z) = \frac{1}{F_j(z_{(m_\ell),j}) - F_j(z_{(m_{\ell-1}),j})} f(Z) \mathbf{1}_{(z_{(m_{\ell-1}),j}, z_{(m_\ell),j})}(z_j),$$

where $F_j(x) := \int_{\{\mathbf{a} \in \mathbb{R}^p : a_j \le x\}} f(\mathbf{a}) \, \mathrm{d}\mathbf{a}$ is the distribution function of the jth coordinate of Z_1.

Proof. The density function of $Z_{\tau^{-1}(1)}, \ldots, Z_{\tau^{-1}(N)}$ at $\mathbf{a}_1, \ldots, \mathbf{a}_N$ is

$$f_{Z_{\tau^{-1}(1)},\ldots,Z_{\tau^{-1}(N)}}(\mathbf{a}_1, \ldots, \mathbf{a}_N) = N!\mathbf{1}_{\{a_{1,j}<\cdots<a_{N,j}\}} \prod_{i=1}^{N} f(\mathbf{a}_i),$$

where $\mathbf{a}_1, \ldots, \mathbf{a}_N \in \mathbb{R}^p$ and $a_{i,j}$ is the jth element of $\mathbf{a}_i$. Then the density function of $Z_{\tau^{-1}\circ\pi(1)}$, $\ldots$, $Z_{\tau^{-1}\circ\pi(N)}$ is

$$\begin{aligned}
&f_{Z_{\tau^{-1}\circ\pi(1)},\ldots,Z_{\tau^{-1}\circ\pi(N)}}(\mathbf{a}_1, \ldots, \mathbf{a}_N) \\
=&\sum_{\pi\in\mathscr{G}} \frac{1}{\prod_{\ell=1}^{k+1}(m_\ell - m_{\ell-1} - 1)!} f_{Z_{\tau^{-1}(1)},\ldots,Z_{\tau^{-1}(N)}}(\mathbf{a}_{\pi^{-1}(1)}, \ldots, \mathbf{a}_{\pi^{-1}(N)}) \\
=&\sum_{\pi_1\in\mathscr{G}_1} \cdots \sum_{\pi_{k+1}\in\mathscr{G}_{k+1}} \frac{1}{\prod_{\ell=1}^{k+1}(m_\ell - m_{\ell-1} - 1)!} \cdot \\
&f_{Z_{\tau^{-1}(1)},\ldots,Z_{\tau^{-1}(N)}}(\mathbf{a}_{\pi_1^{-1}(1)}, \ldots, \mathbf{a}_{\pi_1^{-1}(m_1-1)}, \mathbf{a}_{m_1}, \mathbf{a}_{\pi_2^{-1}(m_1+1)}, \ldots, \mathbf{a}_{m_k}, \mathbf{a}_{\pi_{k+1}^{-1}(m_k+1)}, \ldots, \mathbf{a}_{\pi_{k+1}^{-1}(N)}) \\
=&\frac{N!}{\prod_{\ell=1}^{k+1}(m_\ell - m_{\ell-1} - 1)!} \Big(\prod_{\ell=1}^{k+1} \prod_{i=m_{\ell-1}+1}^{m_\ell-1} (\mathbf{1}_{\{a_{m_{\ell-1},j}<a_{i,j}<a_{m_\ell,j}\}} f(\mathbf{a}_i)) \Big) \Big(\prod_{\ell=1}^{k} f(\mathbf{a}_{m_\ell}) \Big) \mathbf{1}_{\{a_{m_1,j}<\cdots<a_{m_k,j}\}}.
\end{aligned}$$

The density function of $Z_{\tau^{-1}(m_1)}, \ldots, Z_{\tau^{-1}(m_k)}$ is

$$\begin{aligned}
&f_{Z_{\tau^{-1}(m_1)},\ldots,Z_{\tau^{-1}(m_k)}}(\mathbf{a}_{m_1}, \ldots, \mathbf{a}_{m_k}) \\
=&\int_{-\infty}^{\infty} f_{Z_{\tau^{-1}(1)},\ldots,Z_{\tau^{-1}(N)}}(\mathbf{a}_1, \ldots, \mathbf{a}_n)\, \mathrm{d}\mathbf{a}_1 \cdots \mathrm{d}\mathbf{a}_{m_1-1} \mathrm{d}\mathbf{a}_{m_1+1} \cdots \mathrm{d}\mathbf{a}_{m_k-1} \mathrm{d}\mathbf{a}_{m_k+1} \cdots \mathrm{d}\mathbf{a}_N \\
=&N! \left(\prod_{\ell=1}^{k+1} \frac{\big(F_j(a_{m_\ell,j}) - F_j(a_{m_{\ell-1},j})\big)^{m_\ell - m_{\ell-1}-1}}{(m_\ell - m_{\ell-1} - 1)!} \right) \Big(\prod_{\ell=1}^{k} f(\mathbf{a}_{m_\ell}) \Big) \mathbf{1}_{\{a_{m_1,j}<\cdots<a_{m_k,j}\}}.
\end{aligned}$$

Hence the conditional density function of

$$Z_{\tau^{-1}\circ\pi(1)}, \ldots, Z_{\tau^{-1}\circ\pi(m_1-1)}, Z_{\tau^{-1}\circ\pi(m_1+1)}, \ldots, Z_{\tau^{-1}\circ\pi(m_k-1)}, Z_{\tau^{-1}\circ\pi(m_k+1)}, \ldots, Z_{\tau^{-1}\circ\pi(N)}$$

given $Z_{\tau^{-1}(m_1)}, \ldots, Z_{\tau^{-1}(m_k)}$ is

$$\begin{aligned}
&\frac{f_{Z_{\tau^{-1}\circ\pi(1)},\ldots,Z_{\tau^{-1}\circ\pi(N)}}(\mathbf{a}_1, \ldots, \mathbf{a}_N)}{f_{Z_{\tau^{-1}(m_1)},\ldots,Z_{\tau^{-1}(m_k)}}(\mathbf{a}_{m_1}, \ldots, \mathbf{a}_{m_k})} \\
=&\mathbf{1}_{\{a_{m_1,j}<\cdots<a_{m_k,j}\}} \prod_{\ell=1}^{k+1} \prod_{m_{\ell-1}<i<m_\ell} \frac{f(\mathbf{a}_i)\mathbf{1}_{\{a_{m_{\ell-1},j}<a_{i,j}<a_{m_\ell,j}\}}}{F_j(a_{m_\ell,j}) - F_j(a_{m_{\ell-1},j})}.
\end{aligned}$$

And the conclusion follows. □

Now we are ready to prove Theorem C.1.

Proof of Theorem C.1. We prove the conclusion by induction on j. Precisely, we prove that for any $1 \le j^* \le p$, given $\{\gamma_{1,j}, \gamma_{2,j}\}_{j=1}^{j^*}$, the conditional distributions of the random vectors $Z_{i,j}^L$, $Z_{i,j}^R$, $i = 1, \ldots, r-1$, $j = 1, \ldots, j^*$, and Z_{i,j^*}^M, $i = 1, \ldots, N - 2j^*r$, are mutually independent and their densities are given by (C.3), (C.4) and

$$f_{Z_{i,j^*}^M | \{\gamma_{1,\ell}, \gamma_{2,\ell}\}_{\ell=1}^{j^*}}(Z) = \frac{f(z_1, \ldots, z_p) \prod_{\ell=1}^{j^*} \mathbf{1}_{\{\gamma_{1,\ell}<z_\ell<\gamma_{2,\ell}\}}}{\int_{\mathbb{R}^p} f(z_1, \ldots, z_p) \prod_{\ell=1}^{j^*} \mathbf{1}_{\{\gamma_{1,\ell}<z_\ell<\gamma_{2,\ell}\}}\, \mathrm{d}Z}. \tag{C.6}$$

The case for $j^* = 1$ follows from Lemma C.8. Suppose the above statement is valid for $1 \le j^* < p$. Note that γ_{1,j^*+1} and γ_{2,j^*+1} only rely on $Z_{1,j^*}^M, \ldots, Z_{N-2j^*r,j^*}^M$. Hence given $\{\gamma_{1,j}, \gamma_{2,j}\}_{j=1}^{j^*}$, the random variables γ_{1,j^*+1} and γ_{2,j^*+1} are independent of $Z_{i,j}^L$, $Z_{i,j}^R$, $i = 1, \ldots, r-1$, $j = 1, \ldots, j^*$. Consequently, given $\{\gamma_{1,j}, \gamma_{2,j}\}_{j=1}^{j^*+1}$, the conditional distributions of $Z_{i,j}^L$, $Z_{i,j}^R$, $i = 1, \ldots, r-1$, $j = 1, \ldots, j^*$, also have density functions (C.3) and (C.4), and are independent of $Z_{1,j^*}^M, \ldots, Z_{N-2j^*r,j^*}^M$.

Note that given $\{\gamma_{1,j}, \gamma_{2,j}\}_{j=1}^{j^*}$, the random vectors Z_{1,j^*}^M, …, Z_{N-2j^*r,j^*}^M are i.i.d. with density (C.6). Then one can apply Lemma C.8 to Z_{1,j^*}^M, …, Z_{N-2j^*r,j^*}^M conditioning on $\{\gamma_{1,j}, \gamma_{2,j}\}_{j=1}^{j^*}$. It follows that given $\{\gamma_{1,j}, \gamma_{2,j}\}_{j=1}^{j^*+1}$, the random vectors Z_{i,j^*+1}^L, Z_{i,j^*+1}^R, $i = 1, \ldots, r-1$, and Z_{1,j^*+1}^M, …, $Z_{N-2(j^*+1)r,j^*+1}^M$ are independent and have density functions (C.3), (C.4) and (C.6) with j^* replaced by j^*+1. This completes the proof.

□

D Concentration inequalities for the order statistics

The proofs of the theoretical results in the main text heavily rely on the concentration inequalities for the order statistics. Boucheron and Thomas [2012] derived an exponential Efron-Stein inequality for order statistics. However, to apply their results, one needs to bound the expectation of certain function of the spacing statistics, which may be a nontrivial task for concrete distributions. Boucheron and Thomas [2012] remarked that Rényi's representation of the order statistics can be used to derive the concentration inequality of the order statistics. However, they did not provide the general expression of the concentration inequality using this method. Here we give a thorough investigation of concentration inequalities via Rényi's representation. These results will play an important role in the proofs of the main theorems.

Lemma D.9. *Let $\xi_{(1)} \le \xi_{(2)} \le \cdots \le \xi_{(N)}$ be the order statistics of N i.i.d. standard exponential random variables with density function $f(x) = \exp(-x)\mathbf{1}_{[0,+\infty)}(x)$. Then for any $1 \le k \le N$ and $t > 0$,*

$$\Pr\left\{\xi_{(k)} - \sum_{i=1}^{k} \frac{1}{N-i+1} > \Big(\frac{2k}{(N-k+1/2)(N+1/2)}t\Big)^{1/2} + \frac{1}{N-k+1}t\right\} \le \exp(-t),$$

and

$$\Pr\left\{\xi_{(k)} - \sum_{i=1}^{k} \frac{1}{N-i+1} < -\Big(\frac{2k}{(N-k+1/2)(N+1/2)}t\Big)^{1/2}\right\} \le \exp(-t).$$

Proof. From Rényi's representation of the order statistics (see, e.g., Arnold et al. [2008], Theorem 4.6.1), we can write $\xi_{(k)} = \sum_{i=1}^{k} \eta_i/(N-i+1)$, $k = 1, \ldots, N$, where $\eta_1, \ldots, \eta_N$ are independent with standard exponential distributions. We have $\mathrm{E}(\xi_{(k)}) = \sum_{i=1}^{k} 1/(N-i+1)$. From Rényi's representation, it can be seen that for $\lambda < N-k+1$,

$$\log \mathrm{E}(\exp(\lambda(\xi_{(k)} - \mathrm{E}\,\xi_{(k)}))) = -\sum_{i=1}^{k} \log\left(1 - \frac{\lambda}{N-i+1}\right) - \sum_{i=1}^{k} \frac{\lambda}{N-i+1}. \tag{D.7}$$

For $0 < \lambda < N-k+1$, we have

$$\begin{aligned}
\log \mathrm{E}(\exp(\lambda(\xi_{(k)} - \mathrm{E}\,\xi_{(k)}))) \le & \sum_{i=1}^{k} \frac{1}{(N-i+1)^2} \frac{\lambda^2}{2\left(1 - \frac{\lambda}{N-i+1}\right)} \\
\le & \sum_{i=1}^{k} \left(\frac{1}{(N-i+1/2)} - \frac{1}{(N-i+3/2)}\right) \frac{\lambda^2}{2\left(1 - \frac{\lambda}{N-k+1}\right)} \\
= & \frac{k}{(N-k+1/2)(N+1/2)} \frac{\lambda^2}{2\left(1 - \frac{\lambda}{N-k+1}\right)},
\end{aligned}$$

where the first inequality follows from (D.7) and the fact that for $u \in (0,1)$, $-\log(1-u) - u \le u^2/(2(1-u))$. It follows that $\xi_{(k)} - \mathrm{E}\,\xi_{(k)}$ is sub-Gamma on the right tail with variance factor $k/((N-k+1/2)(N+1/2))$ and scale parameter $1/(N-k+1)$; see Boucheron et al. [2013], Section 2.4. Thus, for any $t > 0$,

$$\Pr\left\{\xi_{(k)} - \mathrm{E}\,\xi_{(k)} > \Big(\frac{2k}{(N-k+1/2)(N+1/2)}t\Big)^{1/2} + \frac{1}{N-k+1}t\right\} \le \exp(-t).$$

Hence the first claim holds.

Now we prove the second claim. For $\lambda < 0$,

$$\begin{aligned}\log \mathrm{E}(\exp(\lambda(\xi_{(k)} - \mathrm{E}\,\xi_{(k)}))) \leq & \sum_{i=1}^{k} \frac{1}{(N-i+1)^2} \frac{\lambda^2}{2} \\ \leq & \sum_{i=1}^{k} (\frac{1}{N-i+1/2} - \frac{1}{N-i+3/2}) \frac{\lambda^2}{2} \\ = & \frac{k}{(N-k+1/2)(N+1/2)} \frac{\lambda^2}{2},\end{aligned}$$

where the first inequality follows from (D.7) and the fact that for $u < 0$, $-\log(1-u) - u \leq u^2/2$. It follows that $\xi_{(k)} - \mathrm{E}\,\xi_{(k)}$ is sub-Gamma on the left tail with variance factor $k/((N-k+1/2)(N+1/2))$ and scale parameter 0; see Boucheron et al. [2013], Section 2.4. Thus, for any $t > 0$,

$$\Pr\left\{\xi_{(k)} - \mathrm{E}(\xi_{(k)}) < -\Big(\frac{2k}{(N-k+1/2)(N+1/2)} t\Big)^{1/2}\right\} \leq \exp(-t).$$

This completes the proof of the second claim.

□

Lemma D.10. *Let $\xi_{(1)} \leq \xi_{(2)} \leq \cdots \leq \xi_{(N)}$ be the order statistics of N i.i.d. random variables with common distribution function $F(x)$. Suppose $F(x)$ has density function $f(x)$ with respect to the Lebesgue measure and the support of $f(x)$ is an interval (a,b), $-\infty \leq a < b \leq +\infty$. Then $F(x)$ maps (a,b) onto $(0,1)$. Let $F^{-1} : (0,1) \mapsto (a,b)$ be the inverse of $F(x)$. For $x \in (0,\infty)$, define $F^{\dagger}(x) := F^{-1}(1-e^{-x})$; for $x \leq 0$, define $F^{\dagger}(x) = a$. Then for any $1 \leq k \leq N$ and $t > 0$,*

$$\Pr\left\{\xi_{(k)} > F^{\dagger}\Big(\sum_{i=1}^{k} \frac{1}{N-i+1} + \Big(\frac{2kt}{(N-k+1/2)(N+1/2)}\Big)^{1/2} + \frac{t}{N-k+1}\Big)\right\} \leq \exp(-t),$$

and

$$\Pr\left\{\xi_{(k)} < F^{\dagger}\Big(\sum_{i=1}^{k} \frac{1}{N-i+1} - \Big(\frac{2kt}{(N-k+1/2)(N+1/2)}\Big)^{1/2}\Big)\right\} \leq \exp(-t).$$

Proof. It can be seen that $F^{\dagger}(x)$ is a strict increasing function from $(0,\infty)$ onto (a,b). Let $\eta_{(1)}, \ldots, \eta_{(N)}$ be the order statistics of N independent and identically distributed random variables with standard exponential population. Then $(\xi_{(1)}, \ldots, \xi_{(N)})$ has the same distribution as $(F^{\dagger}(\eta_{(1)}), \ldots, F^{\dagger}(\eta_{(N)}))$. Hence Lemma D.9 implies that

$$\begin{aligned}& \Pr\left\{\xi_{(k)} > F^{\dagger}\Big(\sum_{i=1}^{k} \frac{1}{N-i+1} + \Big(\frac{2kt}{(N-k+1/2)(N+1/2)}\Big)^{1/2} + \frac{t}{N-k+1}\Big)\right\} \\ = & \Pr\left\{\eta_{(k)} > \sum_{i=1}^{k} \frac{1}{N-i+1} + \Big(\frac{2kt}{(N-k+1/2)(N+1/2)}\Big)^{1/2} + \frac{t}{N-k+1}\right\} \\ \leq & \exp(-t).\end{aligned}$$

This proves the first claim. The second claim can be proved similarly. □

The following lemma gives a concentration inequality of beta random variable using Lemma D.10.

Lemma D.11. *Let ξ be a random variable with distribution $\mathrm{Beta}(a,b)$, where a and b are two positive integers. Then for any $t > 0$,*

$$\Pr\left\{\xi < \frac{a-1/2}{a+b-1/2} \exp\Big(-\Big(\frac{2b}{(a-1/2)(a+b-1/2)} t\Big)^{1/2} - \frac{1}{a} t\Big)\right\} \leq \exp(-t);$$

$$\Pr\left\{\xi > \frac{a}{a+b} \exp\Big(\Big(\frac{2b}{(a-1/2)(a+b-1/2)} t\Big)^{1/2}\Big)\right\} \leq \exp(-t).$$

Proof. Let $\xi_{(1)} \le \cdots \le \xi_{(a+b-1)}$ be order statistics of independent standard exponential random variables. Then $\exp(-\xi_{(b)}) \sim \mathrm{Beta}(a, b)$. From Lemma D.9, for any $t > 0$,

$$\Pr\Big\{ \exp(-\xi_{(b)}) < \exp\Big(-\sum_{i=1}^{b} \frac{1}{a+b-i} - \Big(\frac{2b}{(a-1/2)(a+b-1/2)} t \Big)^{1/2} - \frac{1}{a} t \Big) \Big\} \le \exp(-t),$$

and

$$\Pr\Big\{ \exp(-\xi_{(b)}) > \exp\Big(-\sum_{i=1}^{b} \frac{1}{a+b-i} + \Big(\frac{2b}{(a-1/2)(a+b-1/2)} t \Big)^{1/2} \Big) \Big\} \le \exp(-t).$$

The above inequalities, combined with the facts

$$\sum_{i=1}^{b} \frac{1}{a+b-i} \ge \int_{a}^{a+b} \frac{1}{x} \, \mathrm{d}x = \log\left(\frac{a+b}{a} \right),$$
$$\sum_{i=1}^{b} \frac{1}{a+b-i} \le \int_{a-1/2}^{a+b-1/2} \frac{1}{x} \, \mathrm{d}x = \log\left(\frac{a+b-1/2}{a-1/2} \right),$$

leads to the conclusion. □

Now we consider the order statistics of i.i.d. standard normal random variables, that is, $F(x)$ is the standard normal distribution $\Phi(x)$. Define $\Phi^{\dagger}(x) = \Phi^{-1}(1-\exp(-x))$ for $x > 0$, and $\Phi^{\dagger}(x) = -\infty$ for $x \le 0$. Then for $x \ge 0$,

$$x = -\log\left(1 - \Phi(\Phi^{\dagger}(x))\right).$$

Lemma D.12. *For $x \ge \log(2)$,*

$$0 \le \frac{\mathrm{d}}{\mathrm{d}x} \Phi^{\dagger}(x) \le (2\pi)^{1/2}.$$

Proof. The inequality (B.1) implies that for any $x > 0$, $\Phi^{-1}(1 - \exp(-x)) \le (2x)^{1/2}$. Hence for $x \ge \log(2)$,

$$\frac{\mathrm{d}}{\mathrm{d}x} \Phi^{\dagger}(x) = \frac{\exp(-x)}{\varphi(\Phi^{-1}(1-\exp(-x)))} \le \frac{\exp(-x)}{\varphi((2x)^{1/2})} = (2\pi)^{1/2},$$

where the inequality holds since $\Phi^{-1}(1 - e^{-x}) > 0$ for $x > \log(2)$ and $\varphi(t)$ is decreasing in t for $t > 0$. □

Lemma D.13. *For $x > \log(2)$, let $g(x)$ be the function of x such that*

$$\Phi^{\dagger}(x) = \{2x - \log(2x) - \log(2\pi) + g(x)\}^{1/2}.$$

Then $g(x) \to 0$ as $x \to +\infty$.

Proof. The inequality (B.1) implies that for $x > 0$, $x \le \{-2\log(1 - \Phi(x))\}^{1/2}$. Then for $x > 1$,

$$1 - \Phi(x) \ge \left(1 - \frac{1}{x^2}\right) \frac{1}{x} \varphi(x) \ge \left(1 - \frac{1}{x^2}\right) \{-2\log(1-\Phi(x))\}^{-1/2} \varphi(x),$$

where the first inequality follows from (B.2). The above inequality implies that for $x > 1$,

$$x^2 \ge -2\log(1-\Phi(x)) - \log(-2\log(1-\Phi(x))) - \log(2\pi) + 2\log\left(1 - \frac{1}{x^2}\right).$$

Thus, as $x \to +\infty$,

$$x \ge \{-2\log(1-\Phi(x)) - \log(-2\log(1-\Phi(x))) - \log(2\pi) + o(1)\}^{1/2}. \tag{D.8}$$

Now we prove the other direction of the inequality. The inequality (D.8) implies that as $x \to +\infty$, $x \geq (1+o(1))\left\{-2\log(1-\Phi(x))\right\}^{1/2}$. Combine this inequality and the inequality (B.2), we have as $x \to +\infty$,

$$1-\Phi(x) \leq \frac{1}{x}\varphi(x) \leq (1+o(1))\big(-2\log(1-\Phi(x))\big)^{-1/2}(2\pi)^{-1/2}\exp(-x^2/2).$$

That is, as $x \to +\infty$, $x \leq (-2\log(1-\Phi(x)) - \log(-2\log(1-\Phi(x))) - \log(2\pi) + o(1))^{1/2}$. Thus, as $x \to +\infty$, $x = (-2\log(1-\Phi(x)) - \log(-2\log(1-\Phi(x))) - \log(2\pi) + o(1))^{1/2}$, which is equivalent to the conclusion. □

Lemma D.14. *Let $\xi_{(1)} \leq \xi_{(2)} \leq \cdots \leq \xi_{(N)}$ be the order statistics of N independent standard normal random variables. Suppose k is an integer and $(N+1)/2 \leq k \leq N$. Then for any $t > 0$,*

$$\Pr\Big\{\xi_{(k)} - \Phi^{\dagger}\Big(\sum_{i=1}^{k}\frac{1}{N-i+1}\Big) > (2\pi)^{1/2}\Big(\Big(\frac{2kt}{(N-k+1/2)(N+1/2)}\Big)^{1/2} + \frac{t}{N-k+1}\Big)\Big\}$$

$$\leq \exp(-t);$$

and for

$$0 < t \leq \frac{(N-k+1/2)(N+1/2)}{2k}\left(\log\left(\frac{N+1}{2(N-k+1)}\right)\right)^2,$$

we have

$$\Pr\Big\{\xi_{(k)} - \Phi^{\dagger}\Big(\sum_{i=1}^{k}\frac{1}{N-i+1}\Big) < -\Big(\frac{4\pi kt}{(N-k+1/2)(N+1/2)}\Big)^{1/2}\Big\} \leq \exp(-t).$$

Proof. For $(N+1)/2 \leq k \leq N$, we have

$$\sum_{i=1}^{k}\frac{1}{N-i+1} > \int_{N-k+1}^{N+1}\frac{1}{x}\,\mathrm{d}x = \log\left(\frac{N+1}{N-k+1}\right) \geq \log(2).$$

Then from Taylor's theorem and Lemma D.12, for any $t > 0$,

$$\begin{aligned}
&\Phi^{\dagger}\left(\sum_{i=1}^{k}\frac{1}{N-i+1} + \Big(\frac{2kt}{(N-k+1/2)(N+1/2)}\Big)^{1/2} + \frac{t}{N-k+1}\right)\\
\leq&\Phi^{\dagger}\left(\sum_{i=1}^{k}\frac{1}{N-i+1}\right) + (2\pi)^{1/2}\left(\Big(\frac{2kt}{(N-k+1/2)(N+1/2)}\Big)^{1/2} + \frac{t}{N-k+1}\right).
\end{aligned}$$

The above inequality, together with Lemma D.10, leads to the first conclusion.

Similarly, for $k \geq (N+1)/2$ and

$$0 < t \leq \frac{(N-k+1/2)(N+1/2)}{2k}\left(\log\left(\frac{N+1}{2(N-k+1)}\right)\right)^2,$$

we have

$$\log\Big(\frac{N+1}{N-k+1}\Big) - \Big(\frac{2kt}{(N-k+1/2)(N+1/2)}\Big)^{1/2} \geq \log(2),$$

which leads to

$$\begin{aligned}
&\Phi^{\dagger}\Big(\sum_{i=1}^{k}\frac{1}{N-i+1} - \Big(\frac{2kt}{(N-k+1/2)(N+1/2)}\Big)^{1/2}\Big)\\
\geq&\Phi^{\dagger}\Big(\sum_{i=1}^{k}\frac{1}{N-i+1}\Big) - \Big(\frac{4\pi kt}{(N-k+1/2)(N+1/2)}\Big)^{1/2}.
\end{aligned}$$

The above inequality, together with Lemma D.10, leads to the second conclusion.

□

Lemma D.15. *Let $\xi_{(1)} \le \xi_{(2)} \le \cdots \le \xi_{(N)}$ be the order statistics of N independent standard normal random variables. Suppose $m/(N+1) \le 1/(2e)$. Then for*

$$0 < t \le \frac{1}{4}\log\left(\frac{N+1}{2}\right)\log\left(\frac{N+1}{2m}\right),$$

we have

$$\Pr\Big\{\Big|\sum_{i=N-m+1}^{N}\xi_{(i)} - \sum_{i=N-m+1}^{N}\Phi^{\dagger}\Big(\sum_{\ell=1}^{i}\frac{1}{N-\ell+1}\Big)\Big| > (2\pi)^{1/2}(4(mt)^{1/2}+\log(3m)t)\Big\}$$
$$\le 2m\exp(-t),$$

and

$$\Pr\Bigg\{\Big|\sum_{i=N-m+1}^{N}\xi_{(i)}^2 - \sum_{i=N-m+1}^{N}\Big\{\Phi^{\dagger}\Big(\sum_{\ell=1}^{i}\frac{1}{N-\ell+1}\Big)\Big\}^2\Big| > 8\pi\left(2\log(3m)t+t^2\right)$$
$$+ 4(2\pi)^{1/2}\Big\{\sum_{i=N-m+1}^{N}\Big\{\Phi^{\dagger}\Big(\sum_{\ell=1}^{i}\frac{1}{N-\ell+1}\Big)\Big\}^2\Big\}^{1/2}((2\log(3m)t)^{1/2}+t)\Bigg\} \le 2m\exp(-t).$$

Proof. To apply Lemma D.14, we need to give a lower bound of

$$\frac{(N-i+1/2)(N+1/2)}{2i}\left(\log\left(\frac{N+1}{2(N-i+1)}\right)\right)^2, \quad i = N-m+1,\ldots,N.$$

For $N-m+1 \le i \le N$, we have

$$\begin{aligned}
&\frac{(N-i+1/2)(N+1/2)}{2i}\left(\log\left(\frac{N+1}{2(N-i+1)}\right)\right)^2 \\
\ge&\frac{N-i+1}{4}\log\left(\frac{N+1}{2(N-i+1)}\right)\log\left(\frac{N+1}{2m}\right) \\
=&\left\{\frac{2(N-i+1)}{N+1}\log\left(\frac{N+1}{2(N-i+1)}\right)\right\}\frac{N+1}{8}\log\left(\frac{N+1}{2m}\right).
\end{aligned}$$

To lower bound the terms within curly braces, we note that the function $x\log(1/x)$ is increasing for $x \in (0, e^{-1}]$ and for $N-m+1 \le i \le N$, we have

$$\frac{2}{N+1} \le \frac{2(N-i+1)}{N+1} \le \frac{2m}{N+1} \le e^{-1}.$$

Hence for $N-m+1 \le i \le N$,

$$\begin{aligned}
&\frac{(N-i+1/2)(N+1/2)}{2i}\left(\log\left(\frac{N+1}{2(N-i+1)}\right)\right)^2 \\
\ge&\left\{\frac{2}{N+1}\log\left(\frac{N+1}{2}\right)\right\}\frac{N+1}{8}\log\left(\frac{N+1}{2m}\right) \\
=&\frac{1}{4}\log\left(\frac{N+1}{2}\right)\log\left(\frac{N+1}{2m}\right).
\end{aligned}$$

Then Lemma D.14 implies that for $N-m+1 \le i \le N$ and

$$0 < t \le \frac{1}{4}\log\left(\frac{N+1}{2}\right)\log\left(\frac{N+1}{2m}\right),$$

we have

$$\begin{aligned}
&\Pr\left\{\left|\xi_{(i)} - \Phi^{\dagger}\Big(\sum_{\ell=1}^{i}\frac{1}{N-\ell+1}\Big)\right| > (2\pi)^{1/2}\Big(2\Big(\frac{t}{N-i+1}\Big)^{1/2}+\frac{t}{N-i+1}\Big)\right\} \\
\le&\Pr\left\{\left|\xi_{(i)} - \Phi^{\dagger}\Big(\sum_{\ell=1}^{i}\frac{1}{N-\ell+1}\Big)\right| > (2\pi)^{1/2}\Big(\Big(\frac{2it}{(N-i+1/2)(N+1/2)}\Big)^{1/2}+\frac{t}{N-i+1}\Big)\right\} \\
\le&2\exp(-t). \qquad \text{(D.9)}
\end{aligned}$$

By union bound, for

$$0 < t \le \frac{1}{4}\log\left(\frac{N+1}{2}\right)\log\left(\frac{N+1}{2m}\right),$$

we have

$$\Pr\Bigg\{\Big|\sum_{i=N-m+1}^{N}\xi_{(i)} - \sum_{i=N-m+1}^{N}\Phi^{\dagger}\Big(\sum_{\ell=1}^{i}\frac{1}{N-\ell+1}\Big)\Big| \\ > (2\pi)^{1/2}\sum_{i=N-m+1}^{N}\left(2\Big(\frac{t}{N-i+1}\Big)^{1/2} + \frac{t}{N-i+1}\right)\Bigg\} \le 2m\exp(-t).$$

Note that

$$\sum_{i=N-m+1}^{N}\frac{t}{N-i+1} = \sum_{i=1}^{m}\frac{t}{i} \le \int_{1/2}^{m+1/2}\frac{t}{x}\,\mathrm{d}x \le \log(3m)t,$$
$$\sum_{i=N-m+1}^{N}\Big(\frac{t}{N-i+1}\Big)^{1/2} = \sum_{i=1}^{m}\Big(\frac{t}{i}\Big)^{1/2} \le \int_{0}^{m}\Big(\frac{t}{x}\Big)^{1/2}\mathrm{d}x \le 2(mt)^{1/2}.$$

Hence the first claim holds.

Now we prove the second claim. Note that for real numbers a, b, c, if $|a-b| \le |c|$, then $|a^2-b^2| \le (a-b)^2 + 2|b||a-b| \le c^2 + 2|b||c|$. Conversely, if $|a^2-b^2| > c^2+2|b||c|$, then $|a-b| > |c|$. As a consequence of this simple fact and the inequality (D.9), for $N-m+1 \le i \le N$ and

$$0 < t \le \frac{1}{4}\log\left(\frac{N+1}{2}\right)\log\left(\frac{N+1}{2m}\right),$$

we have

$$\Pr\Bigg\{\Big|\xi_{(i)}^2 - \Big\{\Phi^{\dagger}\Big(\sum_{\ell=1}^{i}\frac{1}{N-\ell+1}\Big)\Big\}^2\Big| > 2\pi\Big(2\Big(\frac{t}{N-i+1}\Big)^{1/2} + \frac{t}{N-i+1}\Big)^2 \\ + 2(2\pi)^{1/2}\Big|\Phi^{\dagger}\Big(\sum_{\ell=1}^{i}\frac{1}{N-\ell+1}\Big)\Big|\Big(2\Big(\frac{t}{N-i+1}\Big)^{1/2} + \frac{t}{N-i+1}\Big)\Bigg\}$$
$$\le \Pr\left\{\Big|\xi_{(i)} - \Phi^{\dagger}\Big(\sum_{\ell=1}^{i}\frac{1}{N-\ell+1}\Big)\Big| > (2\pi)^{1/2}\Big(2\Big(\frac{t}{N-i+1}\Big)^{1/2} + \frac{t}{N-i+1}\Big)\right\}$$
$$\le 2\exp(-t).$$

Then by union bound and Cauchy-Schwarz inequality, for

$$0 < t \le \frac{1}{4}\log\left(\frac{N+1}{2}\right)\log\left(\frac{N+1}{2m}\right),$$

we have

$$\Pr\Bigg\{\Big|\sum_{i=N-m+1}^{N}\xi_{(i)}^2 - \sum_{i=N-m+1}^{N}\Big\{\Phi^{\dagger}\left(\sum_{\ell=1}^{i}\frac{1}{N-\ell+1}\right)\Big\}^2\Big| \\ > 2\pi\sum_{i=N-m+1}^{N}\Big(2\Big(\frac{t}{N-i+1}\Big)^{1/2} + \frac{t}{N-i+1}\Big)^2 \\ + 2\Big\{2\pi\Big\{\sum_{i=N-m+1}^{N}\Big\{\Phi^{\dagger}\Big(\sum_{\ell=1}^{i}\frac{1}{N-\ell+1}\Big)\Big\}^2\Big\}\Big\{\sum_{i=N-m+1}^{N}\Big(2\Big(\frac{t}{N-i+1}\Big)^{1/2} + \frac{t}{N-i+1}\Big)^2\Big\}\Big\}^{1/2}\Bigg\}$$
$$\le 2m\exp(-t).$$

But

$$\begin{aligned}\sum_{i=N-m+1}^{N}\Big(2\Big(\frac{t}{N-i+1}\Big)^{1/2}+\frac{t}{N-i+1}\Big)^2 \leq& 8t\sum_{i=1}^{m}\frac{1}{i}+2t^2\sum_{i=1}^{m}\frac{1}{i^2}\\ \leq& 8t\int_{1/2}^{m+1/2}\frac{1}{x}\,\mathrm{d}x+2t^2\int_{1/2}^{m+1/2}\frac{1}{x^2}\,\mathrm{d}x\\ \leq& 8\log(3m)t+4t^2.\end{aligned}$$

The above two inequalities lead to the second conclusion. □

E Proofs of Theorems 1 and 2

Lemma E.16. *Suppose Assumption 1 holds, the sketching matrix* $\mathbf{O}$ *is an* $N\times n$ *matrix with full column rank. Assume that* $\mathbf{O}$ *is independent of* $\varepsilon_1,\ldots,\varepsilon_N$ *and with probability* 1, $\mathbf{O}^\top\mathbf{X}$ *has full column rank. Then*

$$\mathrm{E}\left\{\|\hat{\boldsymbol{\beta}}_{\mathbf{O}}-\boldsymbol{\beta}\|^2\mid\mathbf{Z}\right\}\geq\sigma_\varepsilon^2\,\mathrm{tr}\left\{\left(\mathbf{X}^\top\mathrm{E}(\mathbf{O}(\mathbf{O}^\top\mathbf{O})^{-1}\mathbf{O}^\top\mid\mathbf{Z})\mathbf{X}\right)^{-1}\right\},$$

Proof. The solution to the sketched least square problem (2) is $\hat{\boldsymbol{\beta}}_{\mathbf{O}}=(\mathbf{X}^\top\mathbf{O}\mathbf{O}^\top\mathbf{X})^{-1}\mathbf{X}^\top\mathbf{O}\mathbf{O}^\top\mathbf{y}$ which is an unbiased estimator of $\boldsymbol{\beta}$. It can be seen that

$$\begin{aligned}\mathrm{E}\left\{\|\hat{\boldsymbol{\beta}}_{\mathbf{O}}-\boldsymbol{\beta}\|^2\mid\mathbf{Z},\mathbf{O}\right\}=&\sigma_\varepsilon^2\,\mathrm{tr}\left\{\left(\mathbf{X}^\top\mathbf{O}\mathbf{O}^\top\mathbf{X}\right)^{-1}\mathbf{X}^\top\mathbf{O}\mathbf{O}^\top\mathbf{O}\mathbf{O}^\top\mathbf{X}\left(\mathbf{X}^\top\mathbf{O}\mathbf{O}^\top\mathbf{X}\right)^{-1}\right\}\\ =&\sigma_\varepsilon^2\,\mathrm{tr}(\mathbf{B}_1^\top\mathbf{B}_1),\end{aligned}$$

where $\mathbf{B}_1=(\mathbf{O}^\top\mathbf{O})^{1/2}\mathbf{O}^\top\mathbf{X}(\mathbf{X}^\top\mathbf{O}\mathbf{O}^\top\mathbf{X})^{-1}$. Define $\mathbf{B}_2=(\mathbf{O}^\top\mathbf{O})^{-1/2}\mathbf{O}^\top\mathbf{X}$. Note that the matrix $\mathbf{B}_2(\mathbf{B}_2^\top\mathbf{B}_2)^{-1}\mathbf{B}_2^\top$ is a projection matrix. Hence

$$\begin{aligned}\mathrm{E}\left\{\|\hat{\boldsymbol{\beta}}_{\mathbf{O}}-\boldsymbol{\beta}\|^2\mid\mathbf{Z},\mathbf{O}\right\}\geq&\sigma_\varepsilon^2\,\mathrm{tr}\left(\mathbf{B}_1^\top\mathbf{B}_2(\mathbf{B}_2^\top\mathbf{B}_2)^{-1}\mathbf{B}_2^\top\mathbf{B}_1\right)\\ =&\sigma_\varepsilon^2\,\mathrm{tr}\left\{\left(\mathbf{X}^\top\mathbf{O}(\mathbf{O}^\top\mathbf{O})^{-1}\mathbf{O}^\top\mathbf{X}\right)^{-1}\right\}.\end{aligned}$$

It is known that the function $\mathbf{B}\mapsto\mathrm{tr}(\mathbf{B}^{-1})$ is a convex function for positive definite $\mathbf{B}$; see, e.g., Bhatia [2007a], Section 1.5. Thus, Jensen's inequality implies that

$$\mathrm{E}\left[\mathrm{tr}\left\{\left(\mathbf{X}^\top\mathbf{O}(\mathbf{O}^\top\mathbf{O})^{-1}\mathbf{O}^\top\mathbf{X}\right)^{-1}\right\}\mid\mathbf{Z}\right]\geq\mathrm{tr}\left\{\left(\mathbf{X}^\top\mathrm{E}(\mathbf{O}(\mathbf{O}^\top\mathbf{O})^{-1}\mathbf{O}^\top\mid\mathbf{Z})\mathbf{X}\right)^{-1}\right\},$$

which completes the proof.

□

Proof of Theorem 1. From Lemma E.16 and the fact that

$$\mathbf{X}^\top\mathrm{E}(\mathbf{O}(\mathbf{O}^\top\mathbf{O})^{-1}\mathbf{O}^\top\mid\mathbf{Z})\mathbf{X}\leq\left\|\mathrm{E}(\mathbf{O}(\mathbf{O}^\top\mathbf{O})^{-1}\mathbf{O}^\top\mid\mathbf{Z})\right\|\mathbf{X}^\top\mathbf{X},$$

we have

$$\mathrm{E}\left\{\|\hat{\boldsymbol{\beta}}_{\mathbf{O}}-\boldsymbol{\beta}\|^2\mid\mathbf{Z}\right\}\geq\frac{\mathrm{tr}\left\{(\mathbf{X}^\top\mathbf{X})^{-1}\right\}\sigma_\varepsilon^2}{\|\mathrm{E}(\mathbf{O}(\mathbf{O}^\top\mathbf{O})^{-1}\mathbf{O}^\top\mid\mathbf{Z})\|}.$$

Note that

$$\begin{aligned}(\mathbf{X}^\top\mathbf{X})^{-1}=&\begin{pmatrix}N & N\bar{Z}^\top\\ N\bar{Z} & N\bar{Z}\bar{Z}^\top+\sum_{i=1}^N(Z_i-\bar{Z})(Z_i-\bar{Z})^\top\end{pmatrix}^{-1}\\ =&\begin{pmatrix}\frac{1}{N}+\bar{Z}^\top\left(\sum_{i=1}^N(Z_i-\bar{Z})(Z_i-\bar{Z})^\top\right)^{-1}\bar{Z} & -\bar{Z}^\top\left(\sum_{i=1}^N(Z_i-\bar{Z})(Z_i-\bar{Z})^\top\right)^{-1}\\ -\left(\sum_{i=1}^N(Z_i-\bar{Z})(Z_i-\bar{Z})^\top\right)^{-1}\bar{Z} & \left(\sum_{i=1}^N(Z_i-\bar{Z})(Z_i-\bar{Z})^\top\right)^{-1}\end{pmatrix}.\end{aligned}$$

Hence

$$\operatorname{tr}\left\{(\mathbf{X}^\top\mathbf{X})^{-1}\right\} = \frac{1}{N} + \bar{Z}^\top\left(\sum_{i=1}^N (Z_i-\bar{Z})(Z_i-\bar{Z})^\top\right)^{-1}\bar{Z} + \operatorname{tr}\left\{\left(\sum_{i=1}^N (Z_i-\bar{Z})(Z_i-\bar{Z})^\top\right)^{-1}\right\}.$$

The matrix $\sum_{i=1}^N (Z_i-\bar{Z})(Z_i-\bar{Z})^\top$ has Wishart distribution with parameter $\mathbf{I}_p$ and $N-1$ degrees of freedom. Lemma B.5 implies that $\|(N-1)^{-1}\sum_{i=1}^N (Z_i-\bar{Z})(Z_i-\bar{Z})^\top - \mathbf{I}_p\| = o_P(1)$. Then Lemma B.4 implies that

$$\left\|(N-1)\left(\sum_{i=1}^N (Z_i-\bar{Z})(Z_i-\bar{Z})^\top\right)^{-1} - \mathbf{I}_p\right\| = o_P(1).$$

The above equation, combined with Lemma B.2, yields

$$\operatorname{tr}\left\{(\mathbf{X}^\top\mathbf{X})^{-1}\right\} = \frac{1}{N} + (1+o_P(1))\frac{1}{N}\bar{Z}^\top\bar{Z} + (1+o_P(1))\frac{p}{N}.$$

Hence the conclusion follows.

□

Proof of Theorem 2. From Lemma E.16, we only need to lower bound the trace of $\left(\mathbf{X}^\top \operatorname{E}(\mathbf{O}(\mathbf{O}^\top\mathbf{O})^{-1}\mathbf{O}^\top \mid \mathbf{Z})\mathbf{X}\right)^{-1}$. From Cauchy-Schwarz inequality,

$$\operatorname{tr}\left\{\left(\mathbf{X}^\top \operatorname{E}(\mathbf{O}(\mathbf{O}^\top\mathbf{O})^{-1}\mathbf{O}^\top \mid \mathbf{Z})\mathbf{X}\right)^{-1}\right\} \geq \frac{p^2}{\operatorname{tr}\left(\mathbf{X}^\top \operatorname{E}(\mathbf{O}(\mathbf{O}^\top\mathbf{O})^{-1}\mathbf{O}^\top \mid \mathbf{Z})\mathbf{X}\right)}.$$

We have

$$\operatorname{tr}(\mathbf{X}^\top \operatorname{E}(\mathbf{O}(\mathbf{O}^\top\mathbf{O})^{-1}\mathbf{O}^\top \mid \mathbf{Z})\mathbf{X}) = \sum_{i=1}^N d_i\|X_i\|^2 = n + \sum_{i=1}^N d_i\|Z_i\|^2.$$

Let $\xi_{(1)} \leq \cdots \leq \xi_{(N)}$ be the order statistics of $\|Z_1\|^2, \ldots, \|Z_n\|^2$. Note that $d_i \geq 0$ and $\sum_{i=1}^N d_i = \operatorname{tr}\operatorname{E}(\mathbf{O}(\mathbf{O}^\top\mathbf{O})^{-1}\mathbf{O}^\top \mid \mathbf{Z}) = n$. Thus,

$$\sum_{i=1}^N d_i\|Z_i\|^2 \leq \sup_{\substack{\sum_{i=1}^N d_i = n,\\ d_i\geq 0, i=1,\ldots,N}} \sum_{k=1}^N d_k\|Z_k\|^2 \leq \sup_{\substack{\sum_{i=1}^N d_i = n,\\ d_i\geq 0, i=1,\ldots,N}} \sum_{k=1}^N d_k\xi_{(k)} \leq d_{\max}\sum_{k=1}^{\lfloor n/d_{\max}\rfloor+1} \xi_{(N-k+1)}.$$

Note that $\|Z_k\|^2$ has $\chi^2(p)$ distribution. Then Lemma D.10 implies that for any $1 \leq k \leq N$ and $t > 0$,

$$\Pr\left\{\xi_{(N-k+1)} > F^\dagger_{\chi^2(p)}\Big(\sum_{i=k}^N \frac{1}{i} + 2\Big(\frac{t}{k}\Big)^{1/2} + \frac{t}{k}\Big)\right\} \leq \exp(-t).$$

where $F^\dagger_{\chi^2(p)}(t) = F^{-1}_{\chi^2(p)}(1-e^{-t})$ and $F_{\chi^2(p)}(x)$ is the distribution function of a $\chi^2(p)$ random variable. From Lemma 1 of Laurent and Massart [2000], for any $t > 0$,

$$F^\dagger_{\chi^2(p)}(t) \leq p + 2\sqrt{pt} + 2t \leq 2p + 3t.$$

It follows that

$$\Pr\left\{\xi_{(N-k+1)} > 2p + 3\Big(\sum_{i=k}^N \frac{1}{i} + 2\Big(\frac{t}{k}\Big)^{1/2} + \frac{t}{k}\Big)\right\} \leq \exp(-t).$$

We replace t by $t + \log(\lfloor n/d_{\max}\rfloor + 1)$ in the above inequality. Note that $\lfloor n/d_{\max}\rfloor + 1 \leq 2n/d_{\max}$. Then the union bound implies that for any $t > 0$,

$$\Pr\left\{\sum_{k=1}^{\lfloor n/d_{\max}\rfloor+1} \xi_{(N-k+1)} > \frac{4np}{d_{\max}} + 3\sum_{k=1}^{\lfloor n/d_{\max}\rfloor+1}\Big(\sum_{i=k}^N \frac{1}{i} + 2\Big(\frac{t+\log(2n/d_{\max})}{k}\Big)^{1/2} + \frac{t+\log(2n/d_{\max})}{k}\Big)\right\} \leq \exp(-t).$$

We have

$$\begin{aligned}\sum_{k=1}^{\lfloor n/d_{\max}\rfloor+1}\sum_{i=k}^{N}\frac{1}{i} =& \sum_{k=1}^{\lfloor n/d_{\max}\rfloor+1}\sum_{i=k}^{\lfloor n/d_{\max}\rfloor+1}\frac{1}{i}+\sum_{k=1}^{\lfloor n/d_{\max}\rfloor+1}\sum_{i=\lfloor n/d_{\max}\rfloor+2}^{N}\frac{1}{i}\\ =&(\lfloor n/d_{\max}\rfloor+1)\left(1+\sum_{i=\lfloor n/d_{\max}\rfloor+2}^{N}\frac{1}{i}\right)\\ \le&(\lfloor n/d_{\max}\rfloor+1)\left(1+\int_{\lfloor n/d_{\max}\rfloor+1}^{N}x^{-1}\,\mathrm{d}x\right)\\ \le&2n/d_{\max}\left(1+\log\left(\frac{Nd_{\max}}{n}\right)\right).\end{aligned}$$

Also,

$$\sum_{k=1}^{\lfloor n/d_{\max}\rfloor+1}\frac{1}{k^{1/2}}\le\int_{0}^{\lfloor n/d_{\max}\rfloor+1}x^{-1/2}\,\mathrm{d}x=2\sqrt{\lfloor n/d_{\max}\rfloor+1}\le 4\sqrt{n/d_{\max}}.$$

And

$$\sum_{k=1}^{\lfloor n/d_{\max}\rfloor+1}\frac{1}{k}\le\int_{1/2}^{\lfloor n/d_{\max}\rfloor+3/2}x^{-1}\,\mathrm{d}x=\log(2\lfloor n/d_{\max}\rfloor+3)\le\log(5n/d_{\max}).$$

Combining the above bounds yields

$$\begin{aligned}\sum_{k=1}^{\lfloor n/d_{\max}\rfloor+1}\xi_{(N-k+1)}\le&\frac{4np}{d_{\max}}+\frac{6n}{d_{\max}}\left(1+\log\left(\frac{Nd_{\max}}{n}\right)\right)\\ &+O_P\left(\left(\frac{n}{d_{\max}}\log(2n/d_{\max})\right)^{1/2}+\left(\log(2n/d_{\max})\right)^2\right)\\ \le&\frac{6n}{d_{\max}}\left(p+\log\left(\frac{Nd_{\max}}{n}\right)\right)+O_P\left(\frac{n}{d_{\max}}\right).\end{aligned}$$

Thus,

$$\mathrm{tr}(\mathbf{X}^\top\,\mathrm{E}(\mathbf{O}(\mathbf{O}^\top\mathbf{O})^{-1}\mathbf{O}^\top\mid\mathbf{Z})\mathbf{X})\le 6n\left(p+\log\left(\frac{Nd_{\max}}{n}\right)\right)+O_P\left(n\right).$$

The conclusion follows. □

F Proof of Theorem 3

In this section, we provide a proof of Theorem 3 of the main text. We approach this goal by proving some more general results and showing that they can imply Theorem 3.

Define sets $\mathcal{A}_j=\bigcap_{\ell=1}^{j}\{Z:\gamma_{1,\ell}<z_\ell<\gamma_{2,\ell}\}$, $j=1,\ldots,p$. Let G_j denote the distribution function of z_j given that $Z\in\mathcal{A}_{j-1}$. That is,

$$G_j(x)=\frac{\int_{\{z_j\le x\}\cap\mathcal{A}_{j-1}}f(Z)\,\mathrm{d}Z}{\int_{\mathcal{A}_{j-1}}f(Z)\,\mathrm{d}Z},\quad x\in\mathbb{R}.$$

Here we emphasis that G_j relies on the random variables $\{\gamma_{1,\ell},\gamma_{2,\ell}\}_{\ell=1}^{j-1}$. Hence G_j itself may be random. Let $\gamma_{3,j}$ and $\gamma_{4,j}$ denote the $(r+1)$th smallest element and the $(r+1)$th largest element among $\{z_{i,j}:i\in\{1,\ldots,N\}\backslash\big(\bigcup_{\ell=1}^{j-1}(\mathscr{L}_{r,\ell}\cup\mathscr{R}_{r,\ell})\big)\}$. Define sets

$$\mathcal{A}_j^L=\{Z:z_j\le\gamma_{3,\ell}\}\cap A_{j-1},\quad \mathcal{A}_j^R=\{Z:z_j\ge\gamma_{4,\ell}\}\cap\mathcal{A}_{j-1}.$$

Proposition 1. *Suppose Assumption 1 holds. Then the* 2*-dimensional random vectors* $(G_1(\gamma_{1,1}), G_2(\gamma_{2,1}))^\top, \ldots, (G_p(\gamma_{1,p}), G_p(\gamma_{2,p}))^\top$ *are mutually independent. And the distributions of* $G_j(\gamma_{1,j})$*,* $G_j(\gamma_{2,j})$*,* $G_j(\gamma_{3,j})$*,* $G_j(\gamma_{4,j})$ *and* $G_j(\gamma_{2,j}) - G_j(\gamma_{1,j})$ *are* $\mathrm{Beta}(r, N - r(2j-1) + 1)$*,* $\mathrm{Beta}(N - r(2j-1) + 1, r)$*,* $\mathrm{Beta}(r+1, N - r(2j-1))$*,* $\mathrm{Beta}(N - r(2j-1), r+1)$ *and* $\mathrm{Beta}(N - 2rj + 1, 2r)$*, respectively.*

Proof. The definition of $\gamma_{i,j}$ implies that given $\{\gamma_{i,\ell} : i = 1, 2,\ \ell = 1, \ldots, j-1\}$, the thresholds $\gamma_{1,j}$ and $\gamma_{2,j}$ are the rth and the $(N - r(2j-1)+1)$th order statistics of $N - 2r(j-1)$ i.i.d. random variables with distribution G_j. Then given $\{\gamma_{i,\ell} : i = 1, 2,\ \ell = 1, \ldots, j-1\}$, $G_j(\gamma_{1,j})$ and $G_j(\gamma_{2,j})$ are the rth and the $(N - r(2j-1)+1)$th order statistics of $N - 2r(j-1)$ i.i.d. random variables with uniform distribution on the interval $(0, 1)$, which does not rely on $\{\gamma_{i,\ell} : i = 1, 2,\ \ell = 1, \ldots, j-1\}$. Thus, the 2-dimensional random vectors $(G_1(\gamma_{1,1}), G_2(\gamma_{2,1}))^\top, \ldots, (G_p(\gamma_{1,p}), G_p(\gamma_{2,p}))^\top$ are mutually independent. The distributions of $G_j(\gamma_{1,j})$, $G_j(\gamma_{2,j})$, $G_j(\gamma_{3,j})$, $G_j(\gamma_{4,j})$ and $G_j(\gamma_{2,j}) - G_j(\gamma_{1,j})$ follow from Lemma B.7. □

Proposition 2. *Suppose Assumption 1 holds. Then* $\int_{\mathcal{A}_j} f(Z)\,\mathrm{d}Z$ *has distribution* $\mathrm{Beta}(N - 2rj + 1, 2rj)$.

Proof. It can be seen that $\int_{\mathcal{A}_j} f(Z)\,\mathrm{d}Z = \prod_{\ell=1}^{j} \{G_\ell(\gamma_{2,\ell}) - G_\ell(\gamma_{1,\ell})\}$. Let $\eta_1, \ldots, \eta_N$ be i.i.d. standard exponential random variables. From Proposition 1 and Rényi's representation of the order statistics (see, e.g., Arnold et al. [2008], Theorem 4.6.1), for $\ell = 1, \ldots, p$, $G_\ell(\gamma_{2,\ell}) - G_\ell(\gamma_{1,\ell})$ has the same distribution as

$$\exp\left\{ -\sum_{s=1}^{2r} \frac{\eta_{2r(\ell-1)+s}}{N - 2r(\ell-1) - s + 1} \right\}.$$

Thus, $\prod_{\ell=1}^{j} \{G_\ell(\gamma_{2,\ell}) - G_\ell(\gamma_{1,\ell})\}$ has the same distribution as

$$\exp\left\{ -\sum_{\ell=1}^{j} \sum_{s=1}^{2r} \frac{\eta_{2r(\ell-1)+s}}{N - 2r(\ell-1) - s + 1} \right\} = \exp\left\{ -\sum_{s=1}^{2rj} \frac{\eta_s}{N - s + 1} \right\}.$$

Then the conclusion follows from another application of Rényi's representation. □

Proposition 3. *Suppose Assumption 1 holds,* $r = N/(2p)$ *is an integer and* $\log(p)/r$ *is bounded. Then*

$$\begin{aligned}
\max_{j \in \{1, \ldots, p-1\}} \left| \frac{N}{N - 2rj} \int_{\mathcal{A}_j} f(Z)\,\mathrm{d}Z - 1 \right| &= O_P\left(\left(\frac{\log(p)}{r} \right)^{1/2} \right), \\
\max_{i \in \{1,3\}, j \in \{1, \ldots, p\}} \left| \frac{N - 2r(j-1)}{r} G_j(\gamma_{i,j}) - 1 \right| &= O_P\left(\left(\frac{\log(p)}{r} \right)^{1/2} \right), \\
\max_{i \in \{2,4\}, j \in \{1, \ldots, p\}} \left| \frac{N - 2r(j-1)}{r} (1 - G_j(\gamma_{i,j})) - 1 \right| &= O_P\left(\left(\frac{\log(p)}{r} \right)^{1/2} \right).
\end{aligned}$$

Proof. From Proposition 2, $\int_{\mathcal{A}_j} f(Z)\,\mathrm{d}Z$ has distribution $\mathrm{Beta}(N - 2rj + 1, 2rj)$. From Lemma D.11 and some algebra, for any $t > 0$,

$$\begin{aligned}
\Pr\left\{ \int_{\mathcal{A}_j} f(Z)\,\mathrm{d}Z < \frac{N - 2rj + 1/2}{N + 1/2} \exp\left\{ -\left(\frac{t}{r} \right)^{1/2} - \frac{t}{r} \right\} \right\} &\leq \exp(-t), \\
\Pr\left\{ \int_{\mathcal{A}_j} f(Z)\,\mathrm{d}Z > \frac{N - 2rj + 1}{N + 1} \exp\left\{ \left(\frac{t}{r} \right)^{1/2} \right\} \right\} &\leq \exp(-t).
\end{aligned}$$

The first conclusion follows from the above inequalities and the union bound.

From Proposition 1, $G_j(\gamma_{i,j})$, $i = 1, \ldots, 4$, $j = 1, \ldots, p$, have beta distributions. Then the second and third conclusions can be similarly derived from Lemma D.11 and the union bound. □

Let

$$\mathcal{B}_j^L = \left\{ Z : G_j(z_j) < \frac{1}{2(p-j+1)} \right\} \cap \mathcal{A}_{j-1}, \quad \mathcal{B}_j^R = \left\{ Z : 1 - G_j(z_j) < \frac{1}{2(p-j+1)} \right\} \cap \mathcal{A}_{j-1}.$$

Define matrices

$$\mathbf{D}_L = \left(\frac{\int_{\mathcal{B}_1^L} (Z-\mu) f(Z)\,\mathrm{d}Z}{\int_{\mathcal{B}_1^L} f(Z)\,\mathrm{d}Z}, \ldots, \frac{\int_{\mathcal{B}_p^L} (Z-\mu) f(Z)\,\mathrm{d}Z}{\int_{\mathcal{B}_p^L} f(Z)\,\mathrm{d}Z} \right),$$
$$\mathbf{D}_R = \left(\frac{\int_{\mathcal{B}_1^R} (Z-\mu) f(Z)\,\mathrm{d}Z}{\int_{\mathcal{B}_1^R} f(Z)\,\mathrm{d}Z}, \ldots, \frac{\int_{\mathcal{B}_p^R} (Z-\mu) f(Z)\,\mathrm{d}Z}{\int_{\mathcal{B}_p^R} f(Z)\,\mathrm{d}Z} \right).$$

Let $\mathbf{A} = \mathbf{D}_L \mathbf{D}_L^\top + \mathbf{D}_R \mathbf{D}_R^\top$. Let

$$\rho_1 = \max_{j \in \{1,\ldots,p\}} \int_{\mathbb{R}^p} \left| \mathbf{1}_{\mathcal{A}_j^L} - \mathbf{1}_{\mathcal{B}_j^L} \right| f(Z)\,\mathrm{d}Z.$$

Proposition 4. *Suppose Assumption 1 holds, $r = N/(2p)$ is an integer and $\log(p)/r \to 0$. Then*

$$\max_{j \in \{1,\ldots,p\}} \int_{\mathbb{R}^p} \left| \mathbf{1}_{\mathcal{A}_j^L} - \mathbf{1}_{\mathcal{B}_j^L} \right| f(Z)\,\mathrm{d}Z = O_P\left(\left(\frac{\log(p)}{Np} \right)^{1/2} \right).$$

Proof. Recall that by definition, $\mathcal{A}_j^L = \{Z : z_j \le \gamma_{3,\ell}\} \cap A_{j-1}$ and $\mathcal{B}_j^L = \left\{Z : G_j(z_j) < \frac{1}{2(p-j+1)}\right\} \cap \mathcal{A}_{j-1}$. We have

$$\begin{aligned}
\int_{\mathbb{R}^p} \left| \mathbf{1}_{\mathcal{A}_j^L} - \mathbf{1}_{\mathcal{B}_j^L} \right| f(Z)\,\mathrm{d}Z &= \int_{\mathbb{R}^p} \left| \mathbf{1}_{\{z_j < \gamma_{3,j}\}} - \mathbf{1}_{\{G_j(z_\ell) < \frac{1}{2(p-j+1)}\}} \right| \mathbf{1}_{\mathcal{A}_{j-1}} f(Z)\,\mathrm{d}Z \\
&= \left| G_j(\gamma_{3,j}) - \frac{1}{2(p-j+1)} \right| \int_{\mathcal{A}_{j-1}} f(Z)\,\mathrm{d}Z.
\end{aligned}$$

Proposition 3 implies that uniformly for $j = 1, \ldots, p$,

$$\begin{aligned}
G_j(\gamma_{3,j}) - \frac{1}{2(p-j+1)} &= O_P\left(\frac{r}{N - 2r(j-1)} \left(\frac{\log(p)}{r} \right)^{1/2} \right), \\
\int_{\mathcal{A}_{j-1}} f(Z)\,\mathrm{d}Z &= \frac{N-2r(j-1)}{N}(1 + o_P(1)).
\end{aligned}$$

Combining the above equalities yields the conclusion. □

Proposition 5. *Suppose Assumption 1 holds, $r = N/(2p)$ is an integer and $\log(p)/r \to 0$. Then uniformly for $j = 1, \ldots, p$,*

$$\int_{\mathcal{A}_j^L} f(Z)\,\mathrm{d}Z = \frac{r}{N}\left(1 + o_P\left(1\right)\right), \quad \int_{\mathcal{B}_j^L} f(Z)\,\mathrm{d}Z = \frac{r}{N}(1 + o_P(1)).$$

Proof. It can be seen that $\int_{\mathcal{A}_j^L} f(Z)\,\mathrm{d}Z = G_j(\gamma_{3,j}) \int_{\mathcal{A}_{j-1}} f(Z)\,\mathrm{d}Z$. From Proposition 3, uniformly for $j = 1, \ldots, p-1$,

$$\int_{\mathcal{A}_j} f(Z)\,\mathrm{d}Z = \frac{N-2rj}{N}\left(1 + O_P\left(\left(\frac{\log(p)}{r}\right)^{1/2}\right)\right).$$

Also, uniformly for $j = 1, \ldots, p$,

$$G_j(\gamma_{3,j}) = \frac{r}{N - 2r(j-1)}\left(1 + O_P\left(\left(\frac{\log(p)}{r}\right)^{1/2}\right)\right).$$

Then the first conclusion follows.

It follows from Proposition 4 that uniformly for $j = 1, \ldots, p$,

$$\left| \int_{\mathcal{B}_j^L} f(Z)\,\mathrm{d}Z - \int_{\mathcal{A}_j^L} f(Z)\,\mathrm{d}Z \right| = O_P\left(\left(\frac{\log(p)}{Np} \right)^{1/2} \right) = o_P\left(\frac{r}{N} \right),$$

where the last equality follows from the assumption $\log(p)/r \to 0$. This, combined with the first conclusion, leads to the second conclusion.

□

Proposition 6. *Suppose Assumption 1 holds, $r = N/(2p)$ is an integer and $p^3/N \to 0$. Then*

$$\left(\sum_{j=1}^{p} \left\| \frac{1}{r} \sum_{i \in \mathscr{L}_{r,j}} Z_i - \frac{\int_{\mathcal{A}_j^L} Z f(Z)\,\mathrm{d}Z}{\int_{\mathcal{A}_j^L} f(Z)\,\mathrm{d}Z} \right\|^2 \right)^{1/2} = O_P\left(\frac{p^{3/2}}{N^{1/2}} \left(\sup_{\{\mathcal{A}:\Pr(Z\in\mathcal{A})\le \frac{1}{p}\}} (\mathrm{E}(\|Z-\mu\|^2 \mathbf{1}_{\mathcal{A}}))^{1/2} \right) \right).$$

Proof. Without loss of generality, we can assume $\mu = \mathbf{0}_p$. From Theorem C.1, we have

$$\mathrm{E}\left\{ \frac{1}{r} \sum_{i \in \mathscr{L}_{r,j}} Z_i \mid \gamma_{1,1}, \ldots, \gamma_{1,j-1}, \gamma_{2,1}, \ldots, \gamma_{2,j-1}, \gamma_{3,j} \right\} = \frac{\int_{\mathcal{A}_j^L} Z f(Z)\,\mathrm{d}Z}{\int_{\mathcal{A}_j^L} f(Z)\,\mathrm{d}Z},$$

$$\mathrm{Var}\left\{ \frac{1}{r} \sum_{i \in \mathscr{L}_{r,j}} Z_i \mid \gamma_{1,1}, \ldots, \gamma_{1,j-1}, \gamma_{2,1}, \ldots, \gamma_{2,j-1}, \gamma_{3,j} \right\} \le \frac{1}{r} \frac{\int_{\mathcal{A}_j^L} Z Z^\top f(Z)\,\mathrm{d}Z}{\int_{\mathcal{A}_j^L} f(Z)\,\mathrm{d}Z}.$$

It follows that

$$\begin{aligned}
&\mathrm{E}\left(\frac{1}{\int_{\mathcal{A}_j^L} \|Z\|^2 f(Z)\,\mathrm{d}Z} \left\| \frac{1}{r} \sum_{i \in \mathscr{L}_{r,j}} Z_i - \frac{\int_{\mathcal{A}_j^L} Z f(Z)\,\mathrm{d}Z}{\int_{\mathcal{A}_j^L} f(Z)\,\mathrm{d}Z} \right\|^2 \right) \\
=& \mathrm{E}\left\{ \frac{\mathrm{tr}\,\mathrm{Var}\left(\frac{1}{r} \sum_{i \in \mathscr{L}_{r,j}} Z_i \mid \gamma_{1,1}, \ldots, \gamma_{1,j-1}, \gamma_{2,1}, \ldots, \gamma_{2,j-1}, \gamma_{3,j} \right)}{\int_{\mathcal{A}_j^L} \|Z\|^2 f(Z)\,\mathrm{d}Z} \right\} \\
\le& \frac{1}{r} \mathrm{E}\left(\frac{1}{\int_{\mathcal{A}_j^L} f(Z)\,\mathrm{d}Z} \right).
\end{aligned}$$

Note that $\int_{\mathcal{A}_j^L} f(Z)\,\mathrm{d}Z = G_j(\gamma_{3,j}) \int_{\mathcal{A}_{j-1}} f(Z)\,\mathrm{d}Z$. Then from Propositions 1 and 2, we have

$$\mathrm{E}\left(\frac{1}{\int_{\mathcal{A}_j^L} f(Z)\,\mathrm{d}Z} \right) = \frac{N - 2r(j-1)}{r} \frac{N}{N - 2r(j-1)} = \frac{N}{r}.$$

Thus,

$$\mathrm{E}\left\{ \sum_{j=1}^{p} \frac{1}{\int_{\mathcal{A}_j^L} \|Z\|^2 f(Z)\,\mathrm{d}Z} \left\| \frac{1}{r} \sum_{i \in \mathscr{L}_{r,j}} Z_i - \frac{\int_{\mathcal{A}_j^L} Z f(Z)\,\mathrm{d}Z}{\int_{\mathcal{A}_j^L} f(Z)\,\mathrm{d}Z} \right\|^2 \right\} \le \frac{Np}{r^2} = O\left(\frac{p^3}{N} \right).$$

The above inequality, combined with Proposition 5, leads to

$$\begin{aligned}
&\sum_{j=1}^{p} \left\| \frac{1}{r} \sum_{i \in \mathscr{L}_{r,j}} Z_i - \frac{\int_{\mathcal{A}_j^L} Z f(Z)\,\mathrm{d}Z}{\int_{\mathcal{A}_j^L} f(Z)\,\mathrm{d}Z} \right\|^2 \\
\le& \left(\max_{j \in \{1,\ldots,p\}} \int_{\mathcal{A}_j^L} \|Z\|^2 f(Z)\,\mathrm{d}Z \right) \left\{ \sum_{j=1}^{p} \frac{1}{\int_{\mathcal{A}_j^L} \|Z\|^2 f(Z)\,\mathrm{d}Z} \left\| \frac{1}{r} \sum_{i \in \mathscr{L}_{r,j}} Z_i - \frac{\int_{\mathcal{A}_j^L} Z f(Z)\,\mathrm{d}Z}{\int_{\mathcal{A}_j^L} f(Z)\,\mathrm{d}Z} \right\|^2 \right\} \\
=& O_P\left(\frac{p^3}{N} \left(\sup_{\{\mathcal{A}:\Pr(Z\in\mathcal{A})\le \frac{1}{p}\}} \mathrm{E}(\|Z\|^2 \mathbf{1}_{\mathcal{A}}) \right) \right).
\end{aligned}$$

This completes the proof. □

Proposition 7. *Suppose Assumption 1 holds, $r = N/(2p)$ is an integer and $p^3/N \to 0$. Then*

$$\left(\sum_{j=1}^{p}\left\|\frac{\int_{\mathcal{A}_j^L} Zf(Z)\,\mathrm{d}Z}{\int_{\mathcal{A}_j^L} f(Z)\,\mathrm{d}Z}-\frac{\int_{\mathcal{B}_j^L} Zf(Z)\,\mathrm{d}Z}{\int_{\mathcal{B}_j^L} f(Z)\,\mathrm{d}Z}\right\|^2\right)^{1/2}$$
$$=O_P\left(p^{3/2}\sup_{\{\mathcal{A}:\Pr(Z\in\mathcal{A})\le\rho_1\}}\mathrm{E}(\|Z-\mu\|\,\mathbf{1}_{\mathcal{A}})+\frac{p^2\left(\log(p)\right)^{1/2}}{N^{1/2}}\sup_{\{\mathcal{A}:\Pr(Z\in\mathcal{A})\le\frac{1}{p}\}}\mathrm{E}(\|Z-\mu\|\,\mathbf{1}_{\mathcal{A}})\right).$$

Proof. Without loss of generality, we can assume $\mu = \mathbf{0}_p$. We have

$$\begin{aligned}
&\left\|\frac{\int_{\mathcal{A}_j^L} Zf(Z)\,\mathrm{d}Z}{\int_{\mathcal{A}_j^L} f(Z)\,\mathrm{d}Z}-\frac{\int_{\mathcal{B}_j^L} Zf(Z)\,\mathrm{d}Z}{\int_{\mathcal{B}_j^L} f(Z)\,\mathrm{d}Z}\right\|\\
\le&\frac{\left\|\int_{\mathcal{A}_j^L} Zf(Z)\,\mathrm{d}Z-\int_{\mathcal{B}_j^L} Zf(Z)\,\mathrm{d}Z\right\|}{\int_{\mathcal{A}_j^L} f(Z)\,\mathrm{d}Z}+\left|\frac{1}{\int_{\mathcal{A}_j^L} f(Z)\,\mathrm{d}Z}-\frac{1}{\int_{\mathcal{B}_j^L} f(Z)\,\mathrm{d}Z}\right|\left\|\int_{\mathcal{B}_j^L} Zf(Z)\,\mathrm{d}Z\right\|\\
\le&\frac{\int_{\mathbb{R}^p}\left|\mathbf{1}_{\mathcal{A}_j^L}-\mathbf{1}_{\mathcal{B}_j^L}\right|\|Z\|\,f(Z)\,\mathrm{d}Z}{\int_{\mathcal{A}_j^L} f(Z)\,\mathrm{d}Z}+\frac{\int_{\mathbb{R}^p}\left|\mathbf{1}_{\mathcal{A}_j^L}-\mathbf{1}_{\mathcal{B}_j^L}\right|f(Z)\,\mathrm{d}Z}{\int_{\mathcal{A}_j^L} f(Z)\,\mathrm{d}Z\int_{\mathcal{B}_j^L} f(Z)\,\mathrm{d}Z}\int_{\mathcal{B}_j^L}\|Z\|\,f(Z)\,\mathrm{d}Z.
\end{aligned}$$

Note that uniformly for $j = 1, \ldots, p$,

$$\int_{\mathbb{R}^p}\left|\mathbf{1}_{\mathcal{A}_j^L}-\mathbf{1}_{\mathcal{B}_j^L}\right|\|Z\|\,f(Z)\,\mathrm{d}Z\le\sup_{\{\mathcal{A}:\Pr(Z\in\mathcal{A})\le\rho_1\}}\mathrm{E}(\|Z\|\,\mathbf{1}_{\mathcal{A}}).$$

It follows from the above fact and Propositions 4 and 5 that

$$\begin{aligned}
&\max_{j\in\{1,\ldots,p\}}\left\|\frac{\int_{\mathcal{A}_j^L} Zf(Z)\,\mathrm{d}Z}{\int_{\mathcal{A}_j^L} f(Z)\,\mathrm{d}Z}-\frac{\int_{\mathcal{B}_j^L} Zf(Z)\,\mathrm{d}Z}{\int_{\mathcal{B}_j^L} f(Z)\,\mathrm{d}Z}\right\|\\
=&O_P\left(p\sup_{\{\mathcal{A}:\Pr(Z\in\mathcal{A})\le\rho_1\}}\mathrm{E}(\|Z\|\,\mathbf{1}_{\mathcal{A}})+\left(\frac{p^3\log(p)}{N}\right)^{1/2}\sup_{\{\mathcal{A}:\Pr(Z\in\mathcal{A})\le\frac{1}{p}\}}\mathrm{E}(\|Z\|\,\mathbf{1}_{\mathcal{A}})\right).
\end{aligned}$$

Then the conclusion follows from the above equality and the fact that

$$\sum_{j=1}^{p}\left\|\frac{\int_{\mathcal{A}_j^L} Zf(Z)\,\mathrm{d}Z}{\int_{\mathcal{A}_j^L} f(Z)\,\mathrm{d}Z}-\frac{\int_{\mathcal{B}_j^L} Zf(Z)\,\mathrm{d}Z}{\int_{\mathcal{B}_j^L} f(Z)\,\mathrm{d}Z}\right\|^2\le p\max_{j\in\{1,\ldots,p\}}\left\|\frac{\int_{\mathcal{A}_j^L} Zf(Z)\,\mathrm{d}Z}{\int_{\mathcal{A}_j^L} f(Z)\,\mathrm{d}Z}-\frac{\int_{\mathcal{B}_j^L} Zf(Z)\,\mathrm{d}Z}{\int_{\mathcal{B}_j^L} f(Z)\,\mathrm{d}Z}\right\|^2.$$

□

The following theorem gives the asymptotic behavior of the conditional mean squared error of $\hat{\boldsymbol{\beta}}_{\mathrm{A}}$.

Theorem F.2. *Suppose that Assumption 1 holds, $r = N/(2p)$ is an integer, $N > 2p^2$, $p^3/N \to 0$, $\|\mathbf{A}\| = o_P(N/p^2)$, $\|\mathbf{A}^{-1}\| = O_P(1)$ and there exist constants $C_1, C_2 > 0$ such that $C_1 < \lambda_p(\boldsymbol{\Sigma}) \le \lambda_1(\boldsymbol{\Sigma}) < C_2$. Also, suppose that*

$$\|\mathbf{A}\|^{3/2}\sup_{\{\mathcal{A}:\Pr(Z\in\mathcal{A})\le\frac{1}{p}\}}\left(\mathrm{E}(\|Z-\mu\|^2\,\mathbf{1}_{\mathcal{A}})\right)^{1/2}=o_P\left(\frac{N^{1/2}}{p^{3/2}}\right),\tag{F.10}$$

$$\|\mathbf{A}\|^{3/2}\sup_{\{\mathcal{A}:\Pr(Z\in\mathcal{A})\le\frac{1}{p}\}}\mathrm{E}(\|Z-\mu\|\,\mathbf{1}_{\mathcal{A}})=o_P\left(\frac{N^{1/2}}{p^2(\log(p))^{1/2}}\right),\tag{F.11}$$

and for any $M > 0$,

$$\|\mathbf{A}\|^{3/2}\sup_{\{\mathcal{A}:\Pr(Z\in\mathcal{A})\le M\left(\frac{\log(p)}{Np}\right)^{1/2}\}}\mathrm{E}(\|Z-\mu\|\,\mathbf{1}_{\mathcal{A}})=o_P\left(\frac{1}{p^{3/2}}\right).\tag{F.12}$$

Then as $N \to \infty$,

$$\mathrm{E}\left\{\|\hat{\boldsymbol{\beta}}_{\mathrm{A}}-\boldsymbol{\beta}\|^2\mid\mathbf{Z}\right\}=(1+o_P(1))\left(2p\left(\mathrm{tr}(\mathbf{A}^{-1})+\mu^\top\mathbf{A}^{-1}\mu\right)+1\right)\frac{\sigma_\varepsilon^2}{N}.\tag{F.13}$$

Remark 1. *In Theorem F.2, we allow that both N and p tend to infinity, and the primary condition on the dimension is $p^3/N \to 0$. Since the varying p setting is considered, the tail behavior of Z should be controlled. Hence we impose conditions* (F.10)-(F.12) *for the tail behavior of Z.*

Proof of Theorem F.2. We have

$$\mathrm{Var}(\hat{\boldsymbol{\beta}}_{\mathrm{A}} \mid \mathbf{Z}) = \frac{\sigma_\varepsilon^2}{r} \begin{pmatrix} 2p & 2p\bar{Z}^\top \\ 2p\bar{Z} & \sum_{j=1}^p \bar{Z}_j^L \bar{Z}_j^{L\top} + \sum_{j=1}^p \bar{Z}_j^R \bar{Z}_j^{R\top} \end{pmatrix}^{-1}.$$

It can be seen that

$$\sum_{j=1}^p \bar{Z}_j^L \bar{Z}_j^{L\top} + \sum_{j=1}^p \bar{Z}_j^R \bar{Z}_j^{R\top} = 2p\bar{Z}\bar{Z}^\top + \mathbf{C}_L\mathbf{C}_L^\top + \mathbf{C}_R\mathbf{C}_R^\top - 2p(\bar{Z}-\mu)(\bar{Z}-\mu)^\top,$$

where $\mathbf{C}_L = \left(\bar{Z}_1^L - \mu, \ldots, \bar{Z}_p^L - \mu\right)$, and $\mathbf{C}_R = \left(\bar{Z}_1^R - \mu, \ldots, \bar{Z}_p^R - \mu\right)$. Let $\tilde{\mathbf{A}} = \mathbf{C}_L\mathbf{C}_L^\top + \mathbf{C}_R\mathbf{C}_R^\top - 2p(\bar{Z}-\mu)(\bar{Z}-\mu)^\top$. Then it can be seen that

$$\mathrm{Var}(\hat{\boldsymbol{\beta}}_{\mathrm{A}} \mid \mathbf{Z}) = \frac{\sigma_\varepsilon^2}{r} \begin{pmatrix} \frac{1}{2p} + \bar{Z}^\top\tilde{\mathbf{A}}^{-1}\bar{Z} & -\bar{Z}^\top\tilde{\mathbf{A}}^{-1} \\ -\tilde{\mathbf{A}}^{-1}\bar{Z} & \tilde{\mathbf{A}}^{-1} \end{pmatrix}.$$

Consequently,

$$\mathrm{E}\left\{\|\hat{\boldsymbol{\beta}}_{\mathrm{A}} - \boldsymbol{\beta}\|^2 \mid \mathbf{Z}\right\} = \frac{\sigma_\varepsilon^2}{r}\left(\frac{1}{2p} + \bar{Z}^\top\tilde{\mathbf{A}}^{-1}\bar{Z} + \mathrm{tr}(\tilde{\mathbf{A}}^{-1})\right). \tag{F.14}$$

From Propositions 6, 7 and the conditions (F.10), (F.11), (F.12), we have

$$\|\mathbf{C}_L - \mathbf{D}_L\| = o_P\left(\frac{1}{\|\mathbf{A}\|^{3/2}}\right).$$

It follows from the above equation, Lemma B.3 and the fact $\|\mathbf{D}_L\| = \|\mathbf{D}_L\mathbf{D}_L^\top\|^{1/2} \le \|\mathbf{A}\|^{1/2}$ that

$$\|\mathbf{C}_L\mathbf{C}_L^\top - \mathbf{D}_L\mathbf{D}_L^\top\| \le (2\|\mathbf{A}\|^{1/2} + \|\mathbf{C}_L - \mathbf{D}_L\|)\|\mathbf{C}_L - \mathbf{D}_L\| = o_P\left(\frac{1}{\|\mathbf{A}\|}\right).$$

Similarly, it can be shown that $\|\mathbf{C}_R\mathbf{C}_R^\top - \mathbf{D}_R\mathbf{D}_R^\top\| = o_P\left(1/\left(\|\mathbf{A}\|\right)\right)$. Hence $\|\mathbf{C}_L\mathbf{C}_L^\top + \mathbf{C}_R\mathbf{C}_R^\top - \mathbf{A}\| = o_P(1/\|\mathbf{A}\|)$. On the other hand,

$$\mathrm{E}\,\|2p(\bar{Z}-\mu)(\bar{Z}-\mu)^\top\| = \mathrm{tr}(\boldsymbol{\Sigma})/r \le \frac{2C_2p^2}{N} = o(1/\|\mathbf{A}\|),$$

where the last equality follows from the condition $\|\mathbf{A}\| = o(N/p^2)$. Thus, $\|\tilde{\mathbf{A}} - \mathbf{A}\| = o_P(1/\|\mathbf{A}\|)$. This fact, combined with Lemma B.4, leads to

$$\|\tilde{\mathbf{A}}^{-1} - \mathbf{A}^{-1}\| = o_P\left(\frac{1}{\|\mathbf{A}\|}\right). \tag{F.15}$$

From (F.15) and Lemma B.2, we have

$$\begin{aligned}\left|\mathrm{tr}(\tilde{\mathbf{A}}^{-1}) - \mathrm{tr}(\mathbf{A}^{-1})\right| =& \left|\sum_{j=1}^p (\lambda_j(\tilde{\mathbf{A}}^{-1}) - \lambda_j(\mathbf{A}^{-1}))\right| \\ \le& p\|\tilde{\mathbf{A}}^{-1} - \mathbf{A}^{-1}\| \\ =& o_P\left(\frac{p}{\|\mathbf{A}\|}\right) \\ =& o_P(\mathrm{tr}(\mathbf{A}^{-1})).\end{aligned}$$

That is, $\mathrm{tr}(\tilde{\mathbf{A}}^{-1}) = (1 + o_P(1))\,\mathrm{tr}(\mathbf{A}^{-1})$. On the other hand, we have

$$\begin{aligned}\left|\bar{Z}^\top\tilde{\mathbf{A}}^{-1}\bar{Z} - \mu^\top\mathbf{A}^{-1}\mu\right| \le& \left|\bar{Z}^\top\tilde{\mathbf{A}}^{-1}\bar{Z} - \bar{Z}^\top\mathbf{A}^{-1}\bar{Z}\right| + \left|(\bar{Z}+\mu)^\top\mathbf{A}^{-1}(\bar{Z}-\mu)\right| \\ \le& \|\tilde{\mathbf{A}}^{-1} - \mathbf{A}^{-1}\|\|\bar{Z}\|^2 + (2\|\mu\| + \|\bar{Z}-\mu\|)\|\mathbf{A}^{-1}\|\|\bar{Z}-\mu\|.\end{aligned}$$

It follows from (F.15) and the fact $\|\bar{Z}-\mu\| = O_P((p/N)^{1/2})$ that

$$\|\tilde{\mathbf{A}}^{-1}-\mathbf{A}^{-1}\|\|\bar{Z}\|^2 = o_P\left(\frac{p/N+\|\mu\|^2}{\|\mathbf{A}\|}\right) = o_P\left(\mathrm{tr}(\mathbf{A}^{-1})+\mu^\top\mathbf{A}^{-1}\mu\right).$$

We have

$$(2\|\mu\|+\|\bar{Z}-\mu\|)\|\mathbf{A}^{-1}\|\|\bar{Z}-\mu\| = O_P(\|\mu\|\sqrt{p/N}+p/N).$$

Note that $p/N = o_P(p/\|\mathbf{A}\|) = o_P(\mathrm{tr}(\mathbf{A}^{-1}))$. And

$$\|\mu\|\sqrt{p/N} = o_P\left(\frac{\|\mu\|}{(p\|\mathbf{A}\|)^{1/2}}\right) = o_P\left(\frac{\|\mu\|}{(p\|\mathbf{A}\|)^{1/2}}\frac{1}{2p}+\frac{\|\mu\|^2}{\|\mathbf{A}\|}\right) = o_P\left(\frac{1}{2p}+\mu^\top\mathbf{A}^{-1}\mu\right).$$

Hence

$$\bar{Z}^\top\tilde{\mathbf{A}}^{-1}\bar{Z} = \mu^\top\mathbf{A}^{-1}\mu + o_P\left(\frac{1}{2p}+\mu^\top\mathbf{A}^{-1}\mu+\mathrm{tr}(\mathbf{A}^{-1})\right).$$

Thus,

$$\frac{1}{2p}+\bar{Z}^\top\tilde{\mathbf{A}}^{-1}\bar{Z}+\mathrm{tr}(\tilde{\mathbf{A}}^{-1}) = (1+o_P(1))\left(\frac{1}{2p}+\mu^\top\mathbf{A}^{-1}\mu+\mathrm{tr}(\mathbf{A}^{-1})\right).$$

Then the conclusion follows from (F.14) and the above equality.

□

The expression (F.13) gives the asymptotic behavior of the conditional mean squared error of $\hat{\boldsymbol{\beta}}_{\mathrm{A}}$. Note that the (j,j)th element of $\mathbf{D}_L$ is

$$\frac{\int_{\mathbb{R}}\mathbf{1}_{\{z_j:G_j(z_j)<1/(2(p-j+1))\}}(z_j-\mu_j)\,\mathrm{d}G_j(z_j)}{\int_{\mathbb{R}}\mathbf{1}_{\{z_j:G_j(z_j)<1/(2(p-j+1))\}}\,\mathrm{d}G_j(z_j)}.$$

It can be expected that at the tail region of G_j where $G_j(z_j) < 1/(2(p-j+1))$, z_j is largely deviated from μ_j. Consequently, the diagonal elements of $\mathbf{A}$ may be large, which may result in the fast convergence of $\mathrm{E}\{\|\hat{\boldsymbol{\beta}}_{\mathrm{A}}-\boldsymbol{\beta}\|^2 \mid \mathbf{Z}\}$. However, the matrix $\mathbf{A}$ in (F.13) is random which makes it hard to rigorously derive the general convergence rate of $\hat{\boldsymbol{\beta}}_{\mathrm{A}}$. In fact, $\mathbf{A}$ may rely on $\gamma_{i,j}$, $i=1,2$, $j=1,\ldots,p-1$. If p is fixed, $\gamma_{i,j}$ may have a fixed limit and the matrix $\mathbf{A}$ in Theorem F.2 may be replaced by certain nonrandom matrix. However, such nonrandom matrix may not have a tractable form in general. To understand the convergence rate of $\hat{\boldsymbol{\beta}}_{\mathrm{A}}$, we consider the special case that $z_1,\ldots,z_p$ are mutually independent and the distribution of $z_j-\mu_j$ is symmetric. Denote by F_j the distribution function of z_j. Let $\mathbf{A}^\dagger$ be the $p\times p$ diagonal matrix whose (j,j)th element is $2\{\mathrm{E}(z_j-\mu_j \mid F_j(z_j)<1/(2(p-j+1)))\}^2$, $j=1,\ldots,p$. We have the following theorem.

Theorem F.3. *Suppose that Assumption 1 holds, $z_1,\ldots,z_p$ are mutually independent and the distribution of $z_j-\mu_j$ is symmetric about 0. Suppose there exists $C>0$ such that*

$$\min_{j\in\{1,\ldots,p\}}\{\mathrm{E}(z_j-\mu_j \mid F_j(z_j)<1/(2(p-j+1)))\}^2 > C. \tag{F.16}$$

Also suppose that

$$\frac{\log(p)}{r^{1/2}}\max_{j\in\{1,\ldots,p\}}\left|F_j^{-1}\left(\frac{1}{4(p-j+1)}\right)-\mu_j\right| = o\left(\frac{1}{\|\mathbf{A}^\dagger\|^{3/2}}\right). \tag{F.17}$$

Then

$$\left\|\mathbf{A}-\mathbf{A}^\dagger\right\| = o_P\left(\frac{1}{\|\mathbf{A}^\dagger\|}\right).$$

Under the conditions of Theorem F.3, one can replace the matrix $\mathbf{A}$ by $\mathbf{A}^\dagger$ in the expression (F.13). Then the performance of $\hat{\boldsymbol{\beta}}_{\mathrm{A}}$ relies on the squared conditional expectations

$$\{\mathrm{E}(z_j-\mu_j \mid F_j(z_j)<1/(2(p-j+1)))\}^2, \quad j=1,\ldots,p.$$

For a fixed j, if the distribution of z_j does not depend on p and has unbounded support, then as $p \to \infty$,

$$\{\mathrm{E}(z_j - \mu_j \mid F_j(z_j) < 1/(2(p-j+1)))\}^2 \to \infty.$$

In this case, it can be expected that $\mathrm{tr}(\mathbf{A}^{\dagger -1}) = o(p)$ and $\hat{\boldsymbol{\beta}}_{\mathrm{A}}$ has a faster convergence rate than the uniform sampling method. However, the exact convergence rate relies on the tail properties of $z_j - \mu_j$, $j = 1, \ldots, p$.

Proof of Theorem F.3. The (i,j)th element of $\mathbf{D}_L$ is

$$\frac{\int_{\mathcal{B}_j^L} (z_i - \mu_i) f(Z)\,\mathrm{d}Z}{\int_{\mathcal{B}_j^L} f(Z)\,\mathrm{d}Z}.$$

The mutual independence of $z_1, \ldots, z_p$ implies that $G_j(x)$ equals to $F_j(x)$, the marginal distribution of z_j, and

$$\frac{\int_{\mathcal{B}_j^L} (z_i - \mu_i) f(Z)\,\mathrm{d}Z}{\int_{\mathcal{B}_j^L} f(Z)\,\mathrm{d}Z} = \begin{cases} \int_{\gamma_{1,i}}^{\gamma_{2,i}} (x - \mu_i)\,\mathrm{d}F_i(x) / (F_i(\gamma_{2,i}) - F_i(\gamma_{1,i})) & \text{for } i < j, \\ \mathrm{E}\{z_j - \mu_j \mid F_j(z_j) < 1/(2(p-j+1))\} & \text{for } i = j, \\ 0 & \text{for } i > j. \end{cases}$$

Thus,

$$\begin{aligned} &\|\mathbf{D}_L - \mathrm{diag}\left(\mathrm{E}(z_1 - \mu_1 \mid F_1(z_1) < 1/(2p)), \ldots, \mathrm{E}(z_p - \mu_p \mid F_p(z_p) < 1/2)\right)\|_F^2 \\ \leq &\sum_{j=1}^{p-1} (p-j) \left(\frac{\int_{\gamma_{1,j}}^{\gamma_{2,j}} (x - \mu_j)\,\mathrm{d}F_j(x)}{F_j(\gamma_{2,j}) - F_j(\gamma_{1,j})} \right)^2. \end{aligned}$$

Since the distribution of $z_j - \mu_j$ is symmetric about 0, we have $\int_{\gamma_{1,j}}^{2\mu_j - \gamma_{1,j}} (x - \mu_j)\,\mathrm{d}F_j(x) = 0$. Thus,

$$\begin{aligned} \left| \int_{\gamma_{1,j}}^{\gamma_{2,j}} (x - \mu_j)\,\mathrm{d}F_j(x) \right| &= \left| \int_{2\mu_j - \gamma_{1,j}}^{\gamma_{2,j}} (x - \mu_j)\,\mathrm{d}F_j(x) \right| \\ &\leq \max(|\gamma_{1,j} - \mu_j|, |\gamma_{2,j} - \mu_j|) |F_j(\gamma_{2,j}) + F_j(\gamma_{1,j}) - 1|. \end{aligned}$$

From Proposition 3, uniformly for $j = 1, \ldots, p-1$,

$$\begin{aligned} F_j(\gamma_{1,j}) =& \frac{1}{2(p-j+1)} + O_P\left(\frac{1}{p-j} \left(\frac{\log(p)}{r} \right)^{1/2} \right), \\ F_j(\gamma_{2,j}) =& 1 - \frac{1}{2(p-j+1)} + O_P\left(\frac{1}{p-j} \left(\frac{\log(p)}{r} \right)^{1/2} \right). \end{aligned}$$

It follows that uniformly for $j = 1, \ldots, p-1$,

$$(p-j) \left(\frac{\int_{\gamma_{1,j}}^{\gamma_{2,j}} (x - \mu_j)\,\mathrm{d}F_j(x)}{F_j(\gamma_{2,j}) - F_j(\gamma_{1,j})} \right)^2 = O_P\left(\max((\gamma_{1,j} - \mu_j)^2, (\gamma_{2,j} - \mu_j)^2) \frac{\log(p)}{r(p-j)} \right).$$

Also note that with probability tending to 1, uniformly for $j = 1, \ldots, p-1$,

$$\max((\gamma_{1,j} - \mu_j)^2, (\gamma_{2,j} - \mu_j)^2) \leq \left(F_j^{-1}\left(\frac{1}{4(p-j+1)} \right) - \mu_j \right)^2.$$

Thus,

$$\begin{aligned} &\|\mathbf{D}_L - \mathrm{diag}(\mathrm{E}(z_1 - \mu_1 \mid F_1(z_1) < 1/(2p)), \ldots, \mathrm{E}(z_p - \mu_p \mid F_p(z_p) < 1/2))\|_F^2 \\ =& O_P\left(\frac{\log(p)}{r} \left(\sum_{j=1}^{p-1} \frac{1}{p-j} \right) \max_{j \in \{1,\ldots,p\}} \left(F_j^{-1}\left(\frac{1}{4(p-j+1)} \right) - \mu_j \right)^2 \right) \\ =& O_P\left(\frac{(\log(p))^2}{r} \max_{j \in \{1,\ldots,p\}} \left(F_j^{-1}\left(\frac{1}{4(p-j+1)} \right) - \mu_j \right)^2 \right). \end{aligned}$$

It follows from the above equation and Lemma B.3 that

$$\begin{aligned}
&\left\|\mathbf{D}_L\mathbf{D}_L^\top - \operatorname{diag}(\{\mathrm{E}(z_1-\mu_1 \mid F_1(z_1)<1/(2p))\}^2,\ldots,\{\mathrm{E}(z_p-\mu_p \mid F_p(z_p)<1/2)\}^2)\right\| \\
=&O_P\Bigg(\max_{j\in\{1,\ldots,p\}}|\mathrm{E}(z_j-\mu_j \mid F_j(z_j)<1/(2(p-j+1)))| \\
&\qquad\cdot\frac{\log(p)}{r^{1/2}}\max_{j\in\{1,\ldots,p\}}\left|F_j^{-1}\left(\frac{1}{4(p-j+1)}\right)-\mu_j\right|\Bigg) \\
=&O_P\left(\frac{1}{\max_{j\in\{1,\ldots,p\}}\{\mathrm{E}(z_j-\mu_j \mid F_j(z_j)<1/(2(p-j+1)))\}^2}\right),
\end{aligned}$$

where the last equality follows from the conditions (F.16) and (F.17). Since $z_j-\mu_j$ is symmetric about 0, we have

$$\{\mathrm{E}(z_j-\mu_j \mid F_j(z_j)>1-1/(2(p-j+1)))\}^2=\{\mathrm{E}(z_j-\mu_j \mid F_j(z_j)<1/(2(p-j+1)))\}^2.$$

Hence one can similarly obtain

$$\begin{aligned}
&\left\|\mathbf{D}_R\mathbf{D}_R^\top - \operatorname{diag}(\{\mathrm{E}(z_1-\mu_1 \mid F_1(z_1)<1/(2p))\}^2,\ldots,\{\mathrm{E}(z_p-\mu_p \mid F_p(z_p)<1/2)\}^2)\right\| \\
=&O_P\left(\frac{1}{\max_{j\in\{1,\ldots,p\}}\{\mathrm{E}(z_j-\mu_j \mid F_j(z_j)<1/(2(p-j+1)))\}^2}\right).
\end{aligned}$$

The conclusion follows.

□

Proof of Theorem 3. First we show that the conditions of Theorem F.2 and Theorem F.3 hold. To verify the conditions of Theorem F.3, we note that

$$\begin{aligned}
&\mathrm{E}(z_j-\mu_j \mid F_j(z_j)<1/(2(p-j+1))) \\
=&\mathrm{E}(z_1 \mid z_1<\Phi^{-1}(1/(2(p-j+1)))) \\
=&2(p-j+1)\int_{-\infty}^{\Phi^{-1}(1/(2(p-j+1)))}(2\pi)^{-1/2}t\exp(-t^2/2)\,\mathrm{d}t \\
=&-(2/\pi)^{1/2}(p-j+1)\exp\{-(\Phi^{-1}(1/(2(p-j+1))))^2/2\}.
\end{aligned}$$

Lemma D.13 implies that if $p-j\to\infty$, then

$$\left(\Phi^{-1}(1/(2(p-j+1)))\right)^2=2\log(2(p-j+1))-\log(2\log(2(p-j+1)))-\log(2\pi)+o(1).$$

It follows from the above two equations that if $p-j\to\infty$, then

$$\mathrm{E}(z_j-\mu_j \mid F_j(z_j)<1/(2(p-j+1)))=-(1+o(1))\left(2\log(2(p-j+1))\right)^{1/2}. \tag{F.18}$$

Note that the right hand side of (F.18) tends to $-\infty$ as $p-j\to\infty$. As a consequence, the condition (F.16) holds. Now we verify the condition (F.17). From (F.18), we have

$$\max_{j\in\{1,\ldots,p\}}|\mathrm{E}(z_j-\mu_j \mid F_j(z_j)<1/(2(p-j+1)))|^3=(1+o(1))(2\log(2p))^{3/2}.$$

And from Lemma D.13, we have

$$\max_{j\in\{1,\ldots,p\}}\left|F_j^{-1}\left(\frac{1}{4(p-j+1)}\right)-\mu_j\right|=-\Phi^{-1}(1/(4p))=(1+o(1))(2\log(4p))^{1/2}.$$

Thus, the condition (F.17) is equivalent to $(\log(p))^3/r^{1/2}\to 0$. Thus, Theorem F.3 implies that $\|\mathbf{A}-\mathbf{A}^\dagger\|=o_P(1/\log(p))$. As a consequence, $\|\mathbf{A}\|=(1+o_P(1))4\log(2p)$ and $\|\mathbf{A}^\dagger\|=(1+o_P(1))4\log(2p)$.

Now we verify the conditions of Theorem F.2. From Lemma B.6, for any $0<\delta<1$, we have

$$\sup_{\{\mathscr{A}:\Pr(Z\in\mathscr{A})<\delta\}}\mathrm{E}(\|Z\|^2\mathbf{1}_{\mathscr{A}})\le\sum_{j=1}^p\sup_{\{\mathscr{A}:\Pr(Z\in\mathscr{A})<\delta\}}\mathrm{E}(z_j^2\mathbf{1}_{\mathscr{A}})=p\,\mathrm{E}(z_1^2\mathbf{1}_{\{|z_1|>\Phi^{-1}(1-\delta/2)\}}).$$

We note that

$$\begin{aligned}\mathrm{E}(z_1^2\mathbf{1}_{\{|z_1|>\Phi^{-1}(1-\delta/2)\}}) &= \int_0^{+\infty} 2t\Pr(|z_1|\mathbf{1}_{\{|z_1|>\Phi^{-1}(1-\delta/2)\}} > t)\,\mathrm{d}t \\ &= \int_0^{\Phi^{-1}(1-\delta/2)} 2\delta t\,\mathrm{d}t + \int_{\Phi^{-1}(1-\delta/2)}^{+\infty} 2t\Pr(|z_1|>t)\,\mathrm{d}t.\end{aligned}$$

From Lemma D.13, there exists an $\epsilon \in (0,1)$ such that for $0<\delta<\epsilon$,

$$2\log(2/\delta) - \log(2\log(2/\delta)) - \log(3\pi) \le \left(\Phi^{-1}(1-\delta/2)\right)^2 \le 2\log(2/\delta).$$

Hence for $0<\delta<\epsilon$,

$$\int_0^{\Phi^{-1}(1-\delta/2)} 2\delta t\,\mathrm{d}t = \delta\left(\Phi^{-1}(1-\delta/2)\right)^2 \le 2\delta\log(2/\delta).$$

On the other hand,

$$\begin{aligned}\int_{\Phi^{-1}(1-\delta/2)}^{+\infty} 2t\Pr(|z_1|>t)\,\mathrm{d}t \le& 4\int_{\Phi^{-1}(1-\delta/2)}^{+\infty} t\exp(-t^2/2)\,\mathrm{d}t \\ =& 4\exp\{-(\Phi^{-1}(1-\delta/2))^2/2\} \\ \le& \frac{1}{2}\delta(6\pi\log(2/\delta))^{1/2}.\end{aligned}$$

Thus, there exists an absolute constant $C>0$ such that for $0<\delta<\epsilon$,

$$\mathrm{E}(z_1^2\mathbf{1}_{\{|z_1|>\Phi^{-1}(1-\delta/2)\}}) \le C\delta\log(2/\delta).$$

Consequently,

$$\sup_{\{\mathscr{A}:\Pr(Z\in\mathscr{A})<\delta\}} \mathrm{E}(\|Z\|^2\mathbf{1}_{\mathscr{A}}) \le Cp\delta\log(2/\delta).$$

On the other hand,

$$\sup_{\{\mathscr{A}:\Pr(Z\in\mathscr{A})<\delta\}} \mathrm{E}(\|Z\|\mathbf{1}_{\mathscr{A}}) \le \sup_{\{\mathscr{A}:\Pr(Z\in\mathscr{A})<\delta\}} \left(\mathrm{E}(\|Z\|^2\mathbf{1}_{\mathscr{A}})\Pr(\mathscr{A})\right)^{1/2} \le (Cp\log(2/\delta))^{1/2}\,\delta.$$

Thus, the conditions (F.10), (F.11) and (F.12) hold provided $p^3(\log(p))^4\log(N)/N \to 0$.

We have shown that the conditions of Theorem F.2 and Theorem F.3 hold. Consequently,

$$\mathrm{E}\left\{\|\hat{\boldsymbol{\beta}}_{\mathrm{A}} - \boldsymbol{\beta}\|^2 \mid \mathbf{Z}\right\} = (1+o_P(1))\,\mathrm{tr}\left(\mathbf{A}^{\dagger-1}\right)\frac{2p\sigma_\varepsilon^2}{N}.$$

From Stolz-Cesàro theorem (See, e.g., Mureşan [2009], Chapter 3, Theorem 1.22), we have

$$\begin{aligned}\lim_{p\to\infty}\frac{\mathrm{tr}\left(\mathbf{A}^{\dagger-1}\right)}{p/\log(2p)} &= \lim_{p\to\infty}\frac{\sum_{j=1}^{p}\frac{1}{2\{\mathrm{E}(z_1|\Phi(z_1)<1/(2j))\}^2}}{p/\log(2p)} \\ &= \lim_{p\to\infty}\frac{\frac{1}{2\{\mathrm{E}(z_1|\Phi(z_1)<1/(2p))\}^2}}{p/\log(2p) - (p-1)/\log(2(p-1))}.\end{aligned}$$

From (F.18), $2\{\mathrm{E}(z_1 \mid \Phi(z_1)<1/(2p))\}^2 = (1+o(1))4\log(2p)$. On the other hand, it can be seen that as $p\to\infty$, $p/\log(2p) - (p-1)/\log(2(p-1)) = 1/\log(2p) + O(1/(\log(2p))^2)$. Thus,

$$\lim_{p\to\infty}\frac{\mathrm{tr}\left(\mathbf{A}^{\dagger-1}\right)}{p/\log(2p)} = \frac{1}{4}.$$

It follows that

$$\mathrm{E}\left\{\|\hat{\boldsymbol{\beta}}_{\mathrm{A}} - \boldsymbol{\beta}\|^2 \mid \mathbf{Z}\right\} = (1+o_P(1))\,\mathrm{tr}\left(\mathbf{A}^{\dagger-1}\right)\frac{2p\sigma_\varepsilon^2}{N} = (1+o_P(1))\frac{p^2\sigma_\varepsilon^2}{2\log(2p)N},$$

which completes the proof.

□

G Proof of Theorem 4

First we outline the proof structure of Theorem 4. A key idea of the proof is to couple Algorithm 1 with Algorithm 2 which is a variant of Algorithm 1. In Algorithm 2, the thresholds are $\{z_{(r),j}, z_{(N-r+1),j}\}_{j=1}^{p}$ which are more tractable. We prove that under certain conditions, with high probability, Algorithms 1 and 2 produce exactly the same results. Thus, the statistical properties of Algorithm 1 inherits from that of Algorithm 2.

Algorithm 2: A variant of Algorithm 1

Input: Observations $\{Z_i, y_i\}_{i=1}^{N}$, covariate dimension p, subdata sample size n
Output: Estimator of $\boldsymbol{\beta}$
$r \leftarrow \lfloor \frac{n}{2p} \rfloor$
for $j \in \{1, ..., p\}$ **do**
$\mathscr{L}'_{r,j} \leftarrow \{i \in \{1, \ldots, N\} : z_{i,j} \le z_{(r),j}\}$
$\mathscr{R}'_{r,j} \leftarrow \{i \in \{1, \ldots, N\} : z_{i,j} \ge z_{(N-r+1),j}\}$
$\mathscr{I}'_j \leftarrow \mathscr{L}'_{r,j} \cup \mathscr{R}'_{r,j}$
$\hat{\boldsymbol{\beta}}_{\mathrm{I}}^{\dagger} \leftarrow (\sum_{j=1}^{p} \sum_{i \in \mathscr{I}'_j} X_i X_i^{\top})^{-1} (\sum_{j=1}^{p} \sum_{i \in \mathscr{I}'_j} X_i y_i)$
return $\hat{\boldsymbol{\beta}}_{\mathrm{I}}^{\dagger}$

The following theorem shows the equivalence of Algorithms 1 and 2 in the setting that Z is normally distributed.

Theorem G.4. *Suppose Assumption 1 holds and $Z \sim \mathcal{N}(\mu, \mathbf{\Sigma})$. Suppose there exists a constant $0 < \rho < 1/\sqrt{2}$ such that $\max_{1 \le i < j \le p} |\rho_{i,j}| \le \rho$, where $\rho_{i,j} = \sigma_{i,j}/(\sigma_{i,i}\sigma_{j,j})^{1/2}$. Suppose there exist $\epsilon_1, \epsilon_2 \in (0, 1)$ such that for sufficiently large N, $4r \le N^{\epsilon_1}$, $p \le N^{\epsilon_2}$ and the condition (6) holds. Then as $N \to \infty$, with probability tending to 1, Algorithms 1 and 2 produce exactly the same results.*

Remark 2. *Although Theorem G.4 assumes $Z_1, \ldots, Z_N$ are normally distributed, it can be directly generalized to a more general class of distributions. Note that the index sets $\mathscr{I}_j$ and $\mathscr{I}'_j$ are invariant if the covariates are transformed by monotone functions. Hence the conclusion of Theorem G.4 also holds if $z_{i,j} = g_j(\tilde{z}_{i,j})$, where g_j is a monotone function and $(\tilde{z}_{i,1}, \ldots, \tilde{z}_{i,p})^{\top} \sim \mathcal{N}(\mu, \mathbf{\Sigma})$, $i = 1, \ldots, N$, $j = 1, \ldots, p$.*

Remark 3. *In some theoretical analyses of Wang et al. [2019], the algorithm actually studied is in fact Algorithm 2 rather than Algorithm 1. In the proof of Theorem 3 in Wang et al. [2019], it was claimed that if r is fixed as N goes to infinity, using Algorithm 2 instead of Algorithm 1 will not affect the final result. However, this claim was not proved. Theorem G.4 fills this theoretical gap.*

Theorem G.4 allows us to transfer the statistical properties of Algorithm 2 to Algorithm 1. Now we deal with the estimator $\hat{\boldsymbol{\beta}}_{\mathrm{I}}^{\dagger}$ in Algorithm 2. Let

$$\mathbf{D}_N = \operatorname{diag}\left(n, \left(n + 4r \log\left(\frac{N}{r}\right)\right) \mathbf{I}_p\right).$$

The following theorem gives the asymptotic behavior of $\operatorname{Var}(\mathbf{D}_N \hat{\boldsymbol{\beta}}_{\mathrm{I}}^{\dagger} \mid \mathbf{Z})$ in Algorithm 2 for varying n and p under the assumption that Z is normally distributed.

Theorem G.5. *Suppose Assumption 1 holds and $Z \sim \mathcal{N}(\mu, \mathbf{\Sigma})$. Suppose there exist constants $C_1, C_2, C_3 > 0$ such that $C_1 < \lambda_p(\mathbf{\Sigma}) < \lambda_1(\mathbf{\Sigma}) < C_2$ and $\|\mu\| < C_3$. Suppose $n/(2p)$ is an integer. Furthermore, suppose as $N \to \infty$, the condition 7 holds. Then as $N \to \infty$, satisfies*

$$\operatorname{Var}(\mathbf{D}_N \hat{\boldsymbol{\beta}}_{\mathrm{I}}^{\dagger} \mid \mathbf{Z}) = \sigma_{\varepsilon}^2 \begin{pmatrix} 1 + \alpha_N \mu^{\top} \mathbf{W}_N^{-1} \mu & -\alpha_N^{1/2} \mu^{\top} \mathbf{W}_N^{-1} \\ -\alpha_N^{1/2} \mathbf{W}_N^{-1} \mu & \mathbf{W}_N^{-1} \end{pmatrix} + \mathbf{E},$$

where $\|\mathbf{E}\| = o_P(1)$.

Theorem 4 follows from Theorems G.4 and G.5, as indicated by the following proof.

Proof of Theorem 4. From Theorems G.4 and G.5, we have

$$\operatorname{Var}(\mathbf{D}_N \hat{\boldsymbol{\beta}}_{\mathrm{I}} \mid \mathbf{Z}) = \sigma_{\varepsilon}^2 \begin{pmatrix} 1 + \alpha_N \mu^{\top} \mathbf{W}_N^{-1} \mu & -\alpha_N^{1/2} \mu^{\top} \mathbf{W}_N^{-1} \\ -\alpha_N^{1/2} \mathbf{W}_N^{-1} \mu & \mathbf{W}_N^{-1} \end{pmatrix} + \mathbf{E},$$

where $\|\mathbf{E}\| = o_P(1)$. Then from Weyl's inequality (Lemma B.2),

$$\begin{aligned}\mathrm{E}((\hat{\boldsymbol{\beta}}_{0,\mathrm{I}} - \beta_0)^2 \mid \mathbf{Z}) =& (1 + \alpha_N \mu^\top \mathbf{W}_N^{-1} \mu) \frac{\sigma_\varepsilon^2}{n}, \\ \mathrm{E}(\|\hat{\boldsymbol{\beta}}_{1,\mathrm{I}} - \boldsymbol{\beta}_1\|^2 \mid \mathbf{Z}) =& \operatorname{tr}(\mathbf{W}_N^{-1}) \frac{\sigma_\varepsilon^2}{n + 4r \log(N/r)} = \alpha_N \operatorname{tr}(\mathbf{W}_N^{-1}) \frac{\sigma_\varepsilon^2}{n},\end{aligned}$$

where $\hat{\boldsymbol{\beta}}_{0,\mathrm{I}}$ and $\hat{\boldsymbol{\beta}}_{1,\mathrm{I}}$ are the first element and the last p elements of $\hat{\boldsymbol{\beta}}_{\mathrm{I}}$, respectively. The conclusion follows. □

Below we give the detailed proofs of Theorems G.4 and G.5.

Lemma G.17. *Suppose $\{(\xi_i, \eta_i)\}_{i=1}^N$ are independent and identically distributed bivariate normal random variables with distribution*

$$\begin{pmatrix} \xi_i \\ \eta_i \end{pmatrix} \sim \mathscr{N}\left(\begin{pmatrix} \mu_\xi \\ \mu_\eta \end{pmatrix}, \begin{pmatrix} \sigma_{\xi,\xi} & \sigma_{\xi,\eta} \\ \sigma_{\xi,\eta} & \sigma_{\eta,\eta} \end{pmatrix} \right).$$

Suppose $\sigma_{\xi,\xi} > 0$, $\sigma_{\eta\eta} > 0$. For $1 \le r \le N$, let $\mathscr{L}_{r,\xi}$ and $\mathscr{R}_{r,\xi}$ be the indices of the first r smallest and largest observations of $\{\xi_i\}_{i=1}^N$, respectively. Similarly, let $\mathscr{L}_{r,\eta}$ and $\mathscr{R}_{r,\eta}$ denote the indices of the first r smallest and largest observations of $\{\eta_i\}_{i=1}^N$, respectively. Suppose $|\rho_{\xi,\eta}| < 1$ where $\rho_{\xi,\eta} = \sigma_{\xi,\eta}/(\sigma_{\xi,\xi}\sigma_{\eta,\eta})^{1/2}$. Suppose there exist $\epsilon_1, \epsilon_2, \delta \in (0,1)$ such that $4r \le N^{\epsilon_1}$, $0 \le t \le 2\epsilon_2 \log(N)$ and

$$(1+\delta)(1+2\epsilon_2)^{1/2}|\rho_{\xi,\eta}| + (\epsilon_1 + 2\epsilon_2)^{1/2}(1-\rho_{\xi,\eta}^2)^{1/2} < (1-\delta)(1-\epsilon_1)^{1/2}. \tag{G.19}$$

Then there is an N^* only depending on δ, ϵ_1 and ϵ_2 such that for $N > N^*$,

$$\Pr\{(\mathscr{L}_{r,\xi} \cup \mathscr{R}_{r,\xi}) \cap (\mathscr{L}_{r,\eta} \cup \mathscr{R}_{r,\eta}) \neq \emptyset\} \le 4\exp(-t).$$

Proof of Lemma G.17. Note that the conditional distribution of η_i given ξ_i is

$$\eta_i | \xi_i \sim \mathscr{N}\Big(\mu_\eta + \frac{\sigma_{\xi,\eta}}{\sigma_{\xi,\xi}}(\xi_i - \mu_\xi), \sigma_{\eta,\eta} - \frac{\sigma_{\xi,\eta}^2}{\sigma_{\xi,\xi}}\Big).$$

Let $\omega_1, \ldots, \omega_N$ be independent standard normal random variables which are independent of $\xi_1, \ldots, \xi_N$. Then

$$\{(\xi_i, \eta_i)\}_{i=1}^N \overset{\mathscr{L}}{=} \left\{ \left(\xi_i,\ \mu_\eta + \frac{\sigma_{\xi,\eta}}{\sigma_{\xi,\xi}}(\xi_i - \mu_\xi) + \left(\sigma_{\eta,\eta} - \frac{\sigma_{\xi,\eta}^2}{\sigma_{\xi,\xi}}\right)^{1/2} \omega_i \right) \right\}_{i=1}^N.$$

Hence without loss of generality, we assume $\eta_i = \mu_\eta + (\sigma_{\xi,\eta}/\sigma_{\xi,\xi})(\xi_i - \mu_\xi) + (\sigma_{\eta,\eta} - \sigma_{\xi,\eta}^2/\sigma_{\xi,\xi})^{1/2}\omega_i$. Then

$$\max_{i \in \mathscr{L}_{r,\xi} \cup \mathscr{R}_{r,\xi}} \frac{|\eta_i - \mu_\eta|}{\sigma_{\eta,\eta}^{1/2}} \le |\rho_{\xi,\eta}| \max_{1 \le i \le N} \frac{|\xi_i - \mu_\xi|}{\sigma_{\xi,\xi}^{1/2}} + (1-\rho_{\xi,\eta}^2)^{1/2} \max_{i \in \mathscr{L}_{r,\xi} \cup \mathscr{R}_{r,\xi}} |\omega_i|. \tag{G.20}$$

For $t > 0$, define set

$$\mathscr{A}_{1,t} = \left\{ \max_{1 \le i \le N} \frac{|\xi_i - \mu_\xi|}{\sigma_{\xi,\xi}^{1/2}} \le (2\log(2N) + 2t)^{1/2} \right\}.$$

Then

$$\Pr(\mathscr{A}_{1,t}^{\complement}) \le N \Pr\left\{ \frac{|\xi_1 - \mu_\xi|}{\sigma_{\xi,\xi}^{1/2}} > (2\log(2N) + 2t)^{1/2} \right\} \le 2N \exp\{-(\log(2N) + t)\} = \exp(-t),$$

where the first inequality follows from the union bound, and the second inequality follows from the inequality (B.1).

For $t > 0$, define

$$\mathscr{A}_{2,t} = \left\{ \max_{i \in \mathscr{L}_{r,\xi} \cup \mathscr{R}_{r,\xi}} |\omega_i| \le (2\log(4r) + 2t)^{1/2} \right\}.$$

Note that the sets $\mathscr{L}_{r,\xi}$ and $\mathscr{R}_{r,\xi}$ only depend on $\{\xi_i\}_{i=1}^N$. Hence given $\{\xi_i\}_{i=1}^N$, the random variables $\{\omega_i : i \in \mathscr{L}_{r,\xi} \cup \mathscr{R}_{r,\xi}\}$ are independent standard normal random variables. Thus, for any $t > 0$,

$$\begin{aligned}\Pr(\mathscr{A}_{2,t}^{\complement}) =& \mathrm{E}\left\{ \Pr\left\{ \max_{i\in\mathscr{L}_{r,\xi}\cup\mathscr{R}_{r,\xi}} |\omega_i| > (2\log(4r)+2t)^{1/2} \mid \xi_1,\dots,\xi_N \right\}\right\}\\ \le& 2r \Pr\left\{|\omega_1| > (2\log(4r)+2t)^{1/2}\right\}\\ \le& \exp(-t).\end{aligned}$$

The inequality (G.20) implies that on $\mathscr{A}_{1,t} \cap \mathscr{A}_{2,t}$,

$$\max_{i\in\mathscr{L}_{r,\xi}\cup\mathscr{R}_{r,\xi}} \frac{|\eta_i - \mu_\eta|}{\sigma_{\eta,\eta}^{1/2}} \le |\rho_{\xi,\eta}|(2\log(2N)+2t)^{1/2} + \left\{(1-\rho_{\xi,\eta}^2)(2\log(4r)+2t)\right\}^{1/2}. \quad \text{(G.21)}$$

For $t > 0$, define sets

$$\begin{aligned}\mathscr{A}_{3,t} =& \left\{ \frac{\eta_{(N-r+1)} - \mu_\eta}{\sigma_{\eta,\eta}^{1/2}} \ge \Phi^\dagger\left(\sum_{i=r}^N \frac{1}{i} - \Big(\frac{2(N-r+1)t}{(r-1/2)(N+1/2)}\Big)^{1/2}\right)\right\},\\ \mathscr{A}_{4,t} =& \left\{ \frac{\eta_{(r)} - \mu_\eta}{\sigma_{\eta,\eta}^{1/2}} \le -\Phi^\dagger\left(\sum_{i=r}^N \frac{1}{i} - \Big(\frac{2(N-r+1)t}{(r-1/2)(N+1/2)}\Big)^{1/2}\right)\right\}.\end{aligned}$$

From Lemma D.10, for any $t > 0$, $\Pr(\mathscr{A}_{3,t}^{\complement}) \le \exp(-t)$. Note that $(\eta_{(r)} - \mu_\eta) \overset{\mathscr{L}}{=} -(\eta_{(N-r+1)} - \mu_\eta)$. Hence for any $t > 0$, $\Pr(\mathscr{A}_{4,t}^{\complement}) \le \exp(-t)$. On $\mathscr{A}_{3,t} \cap \mathscr{A}_{4,t}$, we have

$$\min_{i\in\mathscr{R}_{\eta,r}\cup\mathscr{L}_{\eta,r}} \frac{|\eta_i - \mu_\eta|}{\sigma_{\eta,\eta}^{1/2}} \ge \Phi^\dagger\left(\sum_{i=r}^N \frac{1}{i} - \Big(\frac{2(N-r+1)t}{(r-1/2)(N+1/2)}\Big)^{1/2}\right). \quad \text{(G.22)}$$

Note that if

$$\Phi^\dagger\left(\sum_{i=r}^N \frac{1}{i} - \Big(\frac{2(N-r+1)t}{(r-1/2)(N+1/2)}\Big)^{1/2}\right) > |\rho_{\xi,\eta}|(2\log(2N)+2t)^{1/2} + \{(1-\rho_{\xi,\eta}^2)(2\log(4r)+2t)\}^{1/2}, \quad \text{(G.23)}$$

then the inequalities (G.21) and (G.22) imply that

$$\mathscr{A}_{1,t} \cap \mathscr{A}_{2,t} \cap \mathscr{A}_{3,t} \cap \mathscr{A}_{4,t} \subset \left\{(\mathscr{L}_{r,\xi} \cup \mathscr{R}_{r,\xi}) \cap (\mathscr{L}_{r,\eta} \cup \mathscr{R}_{r,\eta}) = \emptyset\right\},$$

which leads to

$$\Pr\left\{(\mathscr{L}_{r,\xi} \cup \mathscr{R}_{r,\xi}) \cap (\mathscr{R}_{r,\eta} \cup \mathscr{R}_{r,\eta}) \ne \emptyset\right\} \le \sum_{i=1}^4 \Pr(\mathscr{A}_{i,t}^{\complement}) \le 4\exp(-t).$$

Now we prove that under the assumptions of the lemma, (G.23) holds for $N > N^*$, where N^* only depends on δ, ϵ_1, and ϵ_2.

First we deal with the left hand side of (G.23). Note that

$$\sum_{i=r}^N \frac{1}{i} \ge \sum_{i=r}^N \int_i^{i+1} \frac{1}{x}\,\mathrm{d}x = \int_r^{N+1} \frac{1}{x}\,\mathrm{d}x = \log\left(\frac{N+1}{r}\right) \ge (1-\epsilon_1)\log(N),$$

where the last inequality follows from the assumption $4r \le N^{\epsilon_1}$. On the other hand,

$$\frac{2(N-r+1)t}{(r-1/2)(N+1/2)} \le \frac{2t}{(r-1/2)} \le 4t \le 8\epsilon_2\log(N).$$

Hence there exists N_1^* only depending on δ, ϵ_1 and ϵ_2 such that for $N > N_1^*$,

$$\begin{aligned}\Phi^\dagger\left(\sum_{i=r}^N \frac{1}{i} - \Big(\frac{2(N-r+1)t}{(r-1/2)(N+1/2)}\Big)^{1/2}\right) \ge& \Phi^\dagger\left((1-\epsilon_1)\log(N) - (8\epsilon_2\log(N))^{1/2}\right)\\ \ge& \Phi^\dagger\left((1-\delta)(1-\epsilon_1)\log(N)\right).\end{aligned}$$

From Lemma D.13, for large x, $\Phi^\dagger(x) \geq \{(1-\delta)2x\}^{1/2}$. Then there exists $N_2^* \geq N_1^*$ only depending on δ, ϵ_1 and ϵ_2 such that for $N > N_2^*$,

$$\Phi^\dagger\left(\sum_{i=r}^{N}\frac{1}{i} - \Big(\frac{2(N-r+1)t}{(r-1/2)(N+1/2)}\Big)^{1/2}\right) \geq (1-\delta)\left\{(1-\epsilon_1)2\log(N)\right\}^{1/2}. \tag{G.24}$$

Now we deal with the right hand side of (G.23). There exists $N_3^* \geq N_2^*$ only depending on δ, ϵ_1 and ϵ_2 such that for $N > N_3^*$,

$$\begin{aligned}&|\rho_{\xi,\eta}|(2\log(2N)+2t)^{1/2} + \{(1-\rho_{\xi,\eta}^2)(2\log(4r)+2t)\}^{1/2}\\ &\leq \left((1+\delta)(1+2\epsilon_2)^{1/2}|\rho_{\xi,\eta}| + ((\epsilon_1+2\epsilon_2)(1-\rho_{\xi,\eta}^2))^{1/2}\right)(2\log(N))^{1/2}.\end{aligned} \tag{G.25}$$

It can be seen that the inequalities (G.24), (G.25) and the condition (G.19) lead to (G.23). This completes the proof.

□

Proof of Theorem G.4. Note that the left hand side and the right hand side of (6) are continuous functions of ϵ_1 and ϵ_2. Hence there exists a $0 < \delta < 1$ and $\epsilon_2 < \epsilon_2' < 1$ only depending on ϵ_1 and ϵ_2 such that

$$(1+\delta)(1+2\epsilon_2')^{1/2}|\rho| + \{(\epsilon_1+2\epsilon_2')(1-\rho^2)\}^{1/2} < (1-\delta)(1-\epsilon_1)^{1/2}. \tag{G.26}$$

It can be seen that the left hand side of (G.26) is increasing in $|\rho|$ for $0 \leq |\rho| \leq 2^{-1/2}$. Then (G.26) implies that for any $1 \leq i < j \leq p$,

$$(1+\delta)(1+2\epsilon_2')^{1/2}|\rho_{i,j}| + \{(\epsilon_1+2\epsilon_2')(1-\rho_{i,j}^2)\}^{1/2} < (1-\delta)(1-\epsilon_1)^{1/2}.$$

Then Lemma G.17 implies that there exists an N^* only depending on δ, ϵ_1 and ϵ_2', which in turn only depends on ϵ_1 and ϵ_2, such that for any $1 \leq i < j \leq p$,

$$\Pr\{\mathscr{I}_i' \cap \mathscr{I}_j' \neq \emptyset\} \leq 2\exp(-2\epsilon_2'\log(N)).$$

Thus, for $N > N^*$,

$$\begin{aligned}&1 - \Pr\{\text{Algorithm 1 and Algorithm 2 give exactly the same result}\}\\ &\leq \Pr\{\text{There exist } 1 \leq i < j \leq p \text{ such that } \mathscr{I}_i' \cap \mathscr{I}_j' \neq \emptyset\}\\ &\leq \sum_{1\leq i<j\leq p}\Pr\{\mathscr{I}_i' \cap \mathscr{I}_j' \neq \emptyset\}\\ &\leq p^2\exp(-2\epsilon_2'\log(N))\\ &\leq \exp(-2(\epsilon_2'-\epsilon_2)\log(N)),\end{aligned}$$

which converges to 0 as $N \to \infty$. This completes the proof. □

The following proposition contains essential results for the proof of Theorem G.5.

Proposition 8. *Under the assumptions of Theorem G.5, we have*

$$\sum_{j=1}^{p}\sum_{i\in\mathscr{I}_j'} Z_i = n\mu + \mathbf{E}_1,$$

where $\|\mathbf{E}_1\| = o_P\big(\big(n(n+r\log(N/r))\big)^{1/2}\big)$ *, and*

$$\sum_{j=1}^{p}\sum_{i\in\mathscr{I}_j'} Z_i Z_i^\top = n\left(\mu\mu^\top + \mathbf{\Sigma}\right) + 4r\log\left(\frac{N}{r}\right)\mathbf{\Sigma}\,\mathrm{diag}(\mathbf{\Sigma})^{-1}\mathbf{\Sigma} + \mathbf{E}_2,$$

where $\|\mathbf{E}_2\| = o_P\left(n + r\log(N/r)\right)$.

Proof. For $i = 1, \ldots, N$ and $j = 1, \ldots, p$, the conditional distribution of Z_i given $z_{i,j}$ is

$$Z_i \mid z_{i,j} \sim \mathscr{N}\left(\mu + (z_{i,j} - \mu_j)\sigma_{j,j}^{-1}\mathbf{\Sigma}_{:,j}\,,\ \mathbf{\Sigma}_j^*\right),$$

where $\mathbf{\Sigma}_j^* = \mathbf{\Sigma} - \sigma_{j,j}^{-1}\mathbf{\Sigma}_{:,j}\mathbf{\Sigma}_{:,j}^\top$ is a positive semidefinite but singular matrix. For $i = 1, \ldots, N$ and $j = 1, \ldots, p$, we can write

$$Z_i = \mu + (z_{i,j} - \mu_j)\sigma_{j,j}^{-1}\mathbf{\Sigma}_{:,j} + (\mathbf{\Sigma}_j^*)^{1/2}W_{i,j},$$

where $\{W_{i,j}\}_{i=1}^N$ are independent p-dimensional standard normal random vectors and are independent of $\{z_{i,j}\}_{i=1}^N$. Here we emphasize that for $j \neq j'$, the random vectors $\{W_{i,j}\}_{i=1}^N$ may not be independent of $\{W_{i,j'}\}_{i=1}^N$.

We have

$$\sum_{j=1}^p \sum_{i\in\mathscr{I}_j'} Z_i = n\mu + \sum_{j=1}^p \left(\sum_{i\in\mathscr{I}_j'} \frac{z_{i,j}-\mu_j}{\sigma_{j,j}^{1/2}}\right)\sigma_{j,j}^{-1/2}\mathbf{\Sigma}_{:,j} + \sum_{j=1}^p (\mathbf{\Sigma}_j^*)^{1/2}\left(\sum_{i\in\mathscr{I}_j'} W_{i,j}\right).$$

Note that

$$\begin{aligned}\sum_{i\in\mathscr{I}_j'} \frac{z_{i,j}-\mu_j}{\sigma_{j,j}^{1/2}} =& \sum_{i=1}^r \left(\frac{z_{(i),j}-\mu_j}{\sigma_{j,j}^{1/2}} - \Phi^\dagger\left(\sum_{\ell=1}^i \frac{1}{N-\ell+1}\right)\right)\\ &- \sum_{i=N-r+1}^N \left(-\frac{z_{(i),j}-\mu_j}{\sigma_{j,j}^{1/2}} - \Phi^\dagger\left(\sum_{\ell=1}^i \frac{1}{N-\ell+1}\right)\right).\end{aligned}$$

The distribution of $Z_{(i),j} - \mu_j$ is the same as the distribution of $-(Z_{(N-i+1),j} - \mu_j)$. Hence from Lemma D.15 and the union bound, if $r/(N+1) \le 1/(2e)$, then for

$$0 < t \le \frac{1}{4}\log\left(\frac{N+1}{2}\right)\log\left(\frac{N+1}{2r}\right),$$

we have

$$\Pr\left\{\max_{j\in\{1,\ldots,p\}}\left|\sum_{i\in\mathscr{I}_j'}\frac{z_{i,j}-\mu_j}{\sigma_{j,j}^{1/2}}\right| > 2(2\pi)^{1/2}\left(4(rt)^{1/2} + \log(3r)t\right)\right\} \le 2n\exp(-t).$$

We would like to replace t by $t + \log(n)$. Note that for fixed $t > 0$, we have

$$0 < t + \log(n) \le \frac{1}{4}\log\left(\frac{N+1}{2}\right)\log\left(\frac{N+1}{2r}\right)$$

for sufficiently large N. Hence for any $t > 0$, there exists an N_t^* such that for $N > N_t^*$,

$$\Pr\left\{\max_{j\in\{1,\ldots,p\}}\left|\sum_{i\in\mathscr{I}_j'}\frac{z_{i,j}-\mu_j}{\sigma_{j,j}^{1/2}}\right| > 2(2\pi)^{1/2}\Big(4(r(t+\log(n)))^{1/2} + \log(3r)(t+\log(n))\Big)\right\} \le 2\exp(-t).$$

It follows that

$$\max_{j\in\{1,\ldots,p\}}\left|\sum_{i\in\mathscr{I}_j'}\frac{z_{i,j}-\mu_j}{\sigma_{j,j}^{1/2}}\right| = O_P\left(\max\left((r\log(n))^{1/2},\ \log(r)\log(n)\right)\right) = O_P\left(r^{1/2}\log(n)\right). \tag{G.27}$$

Thus,

$$\begin{aligned}\left\|\sum_{j=1}^p\Big(\sum_{i\in\mathscr{I}_j'}\frac{z_{i,j}-\mu_j}{\sigma_{j,j}^{1/2}}\Big)\sigma_{j,j}^{-1/2}\mathbf{\Sigma}_{:,j}\right\| \le& \frac{C_2 p}{C_1^{1/2}}\max_{j\in\{1,\ldots,p\}}\left|\sum_{i\in\mathscr{I}_j'}\frac{z_{i,j}-\mu_j}{\sigma_{j,j}^{1/2}}\right|\\ =& O_P\left(pr^{1/2}\log(n)\right)\\ =& o_P\left((n(n+r\log(N/r)))^{1/2}\right),\end{aligned}$$

where the last equality follows from the fact $r^{1/2} \le n^{1/2}$ and the condition (7). On the other hand,

$$\Big\| \sum_{j=1}^{p} (\mathbf{\Sigma}_j^*)^{1/2} \Big(\sum_{i \in \mathscr{I}_j'} W_{i,j} \Big) \Big\| \le C_2^{1/2} \sum_{j=1}^{p} \Big\| \sum_{i \in \mathscr{I}_j'} W_{i,j} \Big\| \le \Big(C_2 p \sum_{j=1}^{p} \Big\| \sum_{i \in \mathscr{I}_j'} W_{i,j} \Big\|^2 \Big)^{1/2}.$$

Note that $\sum_{i \in \mathscr{I}_j'} W_{i,j} \sim \mathscr{N}(\mathbf{0}_p, 2r\mathbf{I}_p)$, $j = 1, \ldots, p$. Then $\mathrm{E} \sum_{j=1}^{p} \| \sum_{i \in \mathscr{I}_j'} W_{i,j} \|^2 = np$. Hence the condition (7) implies that

$$\Big\| \sum_{j=1}^{p} (\mathbf{\Sigma}_j^*)^{1/2} \Big(\sum_{i \in \mathscr{I}_j'} W_{i,j} \Big) \Big\| = O_P(n^{1/2} p) = o_P\left((n(n + r \log(N/r)))^{1/2} \right). \tag{G.28}$$

Hence the first claim holds.

Now we turn to the second claim. For $i \in \mathscr{I}_j'$, we have

$$\begin{aligned} Z_i Z_i^\top =& \mu\mu^\top + (z_{i,j} - \mu_j)^2 \sigma_{j,j}^{-2} \mathbf{\Sigma}_{:,j} \mathbf{\Sigma}_{:,j}^\top + (\mathbf{\Sigma}_j^*)^{1/2} W_{i,j} W_{i,j}^\top (\mathbf{\Sigma}_j^*)^{1/2} \\ &+ (z_{i,j} - \mu_j) \sigma_{j,j}^{-1} \left(\mathbf{\Sigma}_{:,j} \mu^\top + \mu \mathbf{\Sigma}_{:,j}^\top \right) + (z_{i,j} - \mu_j) \sigma_{j,j}^{-1} \left(\mathbf{\Sigma}_{:,j} W_{i,j}^\top (\mathbf{\Sigma}_j^*)^{1/2} + (\mathbf{\Sigma}_j^*)^{1/2} W_{i,j} \mathbf{\Sigma}_{:,j}^\top \right) \\ &+ \mu W_{i,j}^\top (\mathbf{\Sigma}_j^*)^{1/2} + (\mathbf{\Sigma}_j^*)^{1/2} W_{i,j} \mu^\top. \end{aligned}$$

Thus,

$$\begin{aligned} \sum_{j=1}^{p} \sum_{i \in \mathscr{I}_j'} Z_i Z_i^\top =& n\mu\mu^\top + \sum_{j=1}^{p} \sum_{i \in \mathscr{I}_j'} (z_{i,j} - \mu_j)^2 \sigma_{j,j}^{-2} \mathbf{\Sigma}_{:,j} \mathbf{\Sigma}_{:,j}^\top + 2r \sum_{j=1}^{p} \mathbf{\Sigma}_j^* \\ &+ \mathbf{A}_1 + \mathbf{A}_2 + \mathbf{A}_2^\top + \mathbf{A}_3 + \mathbf{A}_3^\top + \mathbf{A}_4 + \mathbf{A}_4^\top, \end{aligned}$$

where

$$\mathbf{A}_1 = \sum_{j=1}^{p} (\mathbf{\Sigma}_j^*)^{1/2} \sum_{i \in \mathscr{I}_j'} (W_{i,j} W_{i,j}^\top - \mathbf{I}_p) (\mathbf{\Sigma}_j^*)^{1/2}, \qquad \mathbf{A}_2 = \sum_{j=1}^{p} \Big(\sum_{i \in \mathscr{I}_j'} (z_{i,j} - \mu_j) \Big) \sigma_{j,j}^{-1} \mathbf{\Sigma}_{:,j} \mu^\top,$$

$$\mathbf{A}_3 = \sum_{j=1}^{p} \sigma_{j,j}^{-1} \mathbf{\Sigma}_{:,j} \Big(\sum_{i \in \mathscr{I}_j'} (z_{i,j} - \mu_j) W_{i,j} \Big)^\top (\mathbf{\Sigma}_j^*)^{1/2}, \qquad \mathbf{A}_4 = \mu \sum_{j=1}^{p} (\sum_{i \in \mathscr{I}_j'} W_{i,j})^\top (\mathbf{\Sigma}_j^*)^{1/2}.$$

First we investigate the behavior of $\sum_{j=1}^{p} \sum_{i \in \mathscr{I}_j'} (z_{i,j} - \mu_j)^2 \sigma_{j,j}^{-2} \mathbf{\Sigma}_{:,j} \mathbf{\Sigma}_{:,j}^\top$. We have

$$\sum_{i \in \mathscr{I}_j'} \frac{(z_{i,j} - \mu_j)^2}{\sigma_{j,j}} = \sum_{i=1}^{r} \frac{(z_{(i),j} - \mu_j)^2}{\sigma_{j,j}} + \sum_{i=N-r+1}^{N} \frac{(z_{(i),j} - \mu_j)^2}{\sigma_{j,j}}.$$

Note that the distribution of $Z_{(i),j} - \mu_j$ is the same as the distribution of $-(Z_{(N-i+1),j} - \mu_j)$, $i = 1, \ldots, r$. Then from Lemma D.15 and the union bound, if $r/(N+1) \le 1/(2e)$, then for

$$0 < t \le \frac{1}{4} \log\left(\frac{N+1}{2} \right) \log\left(\frac{N+1}{2r} \right),$$

we have

$$\begin{aligned} &\Pr\Big\{ \max_{j \in \{1, \ldots, p\}} \Big| \sum_{i \in \mathscr{I}_j'} \frac{(z_{i,j} - \mu_j)^2}{\sigma_{j,j}} - 2 \sum_{i=N-r+1}^{N} \Big\{ \Phi^\dagger\Big(\sum_{\ell=1}^{i} \frac{1}{N - \ell + 1} \Big) \Big\}^2 \Big| > 16\pi(2\log(3r)t + t^2) \\ &+ 8(2\pi)^{1/2} \Big\{ \sum_{i=N-r+1}^{N} \Big\{ \Phi^\dagger\Big(\sum_{\ell=1}^{i} \frac{1}{N - \ell + 1} \Big) \Big\}^2 \Big\}^{1/2} \Big((2\log(3r)t)^{1/2} + t \Big) \Big\} \le 2n \exp(-t). \end{aligned}$$

Note that we have assumed that $r/N \to 0$. Then for any fixed $t > 0$, there exists an N_t^* such that for $N > N_t^*$,

$$t + \log(n) \le \frac{1}{4} \log\left(\frac{N+1}{2} \right) \log\left(\frac{N+1}{2r} \right).$$

By replacing t by $t+\log(n)$, we have, for $N > N_t^*$,

$$\begin{aligned}
&\Pr\Big\{ \max_{j\in\{1,\ldots,p\}} \Big| \sum_{i\in\mathscr{I}_j'} \frac{(z_{i,j}-\mu_j)^2}{\sigma_{j,j}} - 2\sum_{i=N-r+1}^{N} \Big\{\Phi^\dagger\Big(\sum_{\ell=1}^{i}\frac{1}{N-\ell+1}\Big)\Big\}^2\Big| \\
&> 16\pi\left(2\log(3r)(t+\log(n)) + (t+\log(n))^2\right) \\
&\quad + 8\Big\{2\pi\sum_{i=N-r+1}^{N}\Big\{\Phi^\dagger\Big(\sum_{\ell=1}^{i}\frac{1}{N-\ell+1}\Big)\Big\}^2\Big\}^{1/2}\Big(\left(2\log(3r)(t+\log(n))\right)^{1/2} + t + \log(n)\Big)\Big\} \\
&\leq 2\exp(-t).
\end{aligned}$$

It follows that

$$\begin{aligned}
&\max_{j\in\{1,\ldots,p\}} \Big| \sum_{i\in\mathscr{I}_j'} \frac{(z_{i,j}-\mu_j)^2}{\sigma_{j,j}} - 2\sum_{i=N-r+1}^{N} \Big\{\Phi^\dagger\Big(\sum_{\ell=1}^{i}\frac{1}{N-\ell+1}\Big)\Big\}^2\Big| \\
&= O_P\Big((\log(n))^2 + \log(n)\Big(\sum_{i=N-r+1}^{N}\Big(\Phi^\dagger\Big(\sum_{\ell=1}^{i}\frac{1}{N-\ell+1}\Big)\Big)^2\Big)^{1/2}\Big). \qquad \text{(G.29)}
\end{aligned}$$

To deal with $\sum_{i=N-r+1}^{N}\big\{\Phi^\dagger\big(\sum_{\ell=1}^{i}1/(N-\ell+1)\big)\big\}^2$, we note that uniformly for $i\in\{N-r+1,\ldots,N\}$,

$$\sum_{\ell=1}^{i}\frac{1}{N-\ell+1} \geq \sum_{\ell=1}^{N-r+1}\frac{1}{N-\ell+1} \geq \int_r^{N+1}\frac{1}{x}\,\mathrm{d}x = \log\left(\frac{N+1}{r}\right) \to \infty.$$

From Lemma D.13, for any $\epsilon\in(0,1)$, there is an N_ϵ^* such that for $N > N_\epsilon^*$, uniformly for $i\in\{N-r+1,\ldots,N\}$,

$$(1-\epsilon)2\sum_{\ell=1}^{i}\frac{1}{N-\ell+1} \leq \Big\{\Phi^\dagger\Big(\sum_{\ell=1}^{i}\frac{1}{N-\ell+1}\Big)\Big\}^2 \leq (1+\epsilon)2\sum_{\ell=1}^{i}\frac{1}{N-\ell+1}.$$

Note that

$$\sum_{i=N-r+1}^{N}\sum_{\ell=1}^{i}\frac{1}{N-\ell+1} = \sum_{i=1}^{r}\sum_{\ell=i}^{N}\frac{1}{\ell} = \sum_{i=1}^{r}\sum_{\ell=i}^{r}\frac{1}{\ell} + \sum_{i=1}^{r}\sum_{\ell=r+1}^{N}\frac{1}{\ell} = r\left(1+\sum_{\ell=r+1}^{N}\frac{1}{\ell}\right).$$

Since $r/N\to 0$, we have $\sum_{\ell=r+1}^{N}1/\ell = (1+o(1))\log(N/r)$. Thus,

$$\sum_{i=N-r+1}^{N}\Big\{\Phi^\dagger\Big(\sum_{\ell=1}^{i}\frac{1}{N-\ell+1}\Big)\Big\}^2 = (1+o(1))2r\log\left(\frac{N}{r}\right).$$

Combining (G.29) and the above equality yields

$$\begin{aligned}
&\max_{j\in\{1,\ldots,p\}} \Big| \sum_{i\in\mathscr{I}_j'} \frac{(z_{i,j}-\mu_j)^2}{\sigma_{j,j}} - 4r\log\left(\frac{N}{r}\right)\Big| \\
&= o_P\Big(r\log\Big(\frac{N}{r}\Big)\Big) + O_P\Big((\log(n))^2 + (\log(n))\Big(r\log\Big(\frac{N}{r}\Big)\Big)^{1/2}\Big).
\end{aligned}$$

It can be seen that $\log(n) = o\left(n + r\log\left(N/r\right)\right)^{1/2}$. Hence we have

$$\max_{j\in\{1,\ldots,p\}}\left|\sum_{i\in\mathscr{I}_j'}\frac{(z_{i,j}-\mu_j)^2}{\sigma_{j,j}} - 4r\log\left(\frac{N}{r}\right)\right| = o_P\left(n + r\log\left(\frac{N}{r}\right)\right). \qquad \text{(G.30)}$$

Note that

$$\begin{aligned}
&\sum_{j=1}^{p}\sum_{i\in\mathscr{I}_j'}(z_{i,j}-\mu_j)^2\sigma_{j,j}^{-2}\mathbf{\Sigma}_{:,j}\mathbf{\Sigma}_{:,j}^\top - 4r\log\left(\frac{N}{r}\right)\mathbf{\Sigma}\,\mathrm{diag}(\mathbf{\Sigma})^{-1}\mathbf{\Sigma} \\
&= \sum_{j=1}^{p}\Big(\sum_{i\in\mathscr{I}_j'}\frac{(z_{i,j}-\mu_j)^2}{\sigma_{j,j}} - 4r\log\left(\frac{N}{r}\right)\Big)\sigma_{j,j}^{-1}\mathbf{\Sigma}_{:,j}\mathbf{\Sigma}_{:,j}^\top.
\end{aligned}$$

But (G.30) implies that for any $\epsilon > 0$,

$$\Pr\Big\{ -\epsilon\Big(n + r\log\Big(\frac{N}{r}\Big)\Big)\mathbf{\Sigma}\,\mathrm{diag}(\mathbf{\Sigma})^{-1}\mathbf{\Sigma} \le \sum_{j=1}^{p}\Big(\sum_{i\in\mathscr{I}_j'}\frac{(z_{i,j}-\mu_j)^2}{\sigma_{j,j}} - 4r\log\Big(\frac{N}{r}\Big)\Big)\sigma_{j,j}^{-1}\mathbf{\Sigma}_{:,j}\mathbf{\Sigma}_{:,j}^\top \le \epsilon\Big(n + r\log\Big(\frac{N}{r}\Big)\Big)\mathbf{\Sigma}\,\mathrm{diag}(\mathbf{\Sigma})^{-1}\mathbf{\Sigma}\Big\} \to 1.$$

It follows that

$$\Big\|\sum_{j=1}^{p}\sum_{i\in\mathscr{I}_j'}(z_{i,j}-\mu_j)^2\sigma_{j,j}^{-2}\mathbf{\Sigma}_{:,j}\mathbf{\Sigma}_{:,j}^\top - 4r\log\Big(\frac{N}{r}\Big)\mathbf{\Sigma}\,\mathrm{diag}(\mathbf{\Sigma})^{-1}\mathbf{\Sigma}\Big\| = o_P\Big(n + r\log\Big(\frac{N}{r}\Big)\Big).$$

Thus,

$$\begin{aligned}&\Big\|\sum_{j=1}^{p}\sum_{i\in\mathscr{I}_j'}Z_iZ_i^\top - \Big(n(\mu\mu^\top+\mathbf{\Sigma}) + 4r\log\Big(\frac{N}{r}\Big)\mathbf{\Sigma}\,\mathrm{diag}(\mathbf{\Sigma})^{-1}\mathbf{\Sigma}\Big)\Big\|\\ \le&\sum_{i=1}^{4}2\|A_i\| + o_P\Big(n + r\log\Big(\frac{N}{r}\Big)\Big).\end{aligned}$$

It remains to show that $\|\mathbf{A}_i\| = o_P(n + r\log(N/r))$, $i = 1,\dots,4$.

Note that $\mathscr{I}_j'$ only relies on $\{z_{i,j}\}_{i=1}^N$. Hence $\mathscr{I}_j'$ is independent of $\{W_{i,j}\}_{i=1}^N$. From Lemma B.5, for any $t > 0$,

$$\begin{aligned}&\Pr\Big(\Big\|\sum_{i\in\mathscr{I}_j'}(W_{i,j}W_{i,j}^\top - \mathbf{I}_p)\Big\| > 3\max\Big((2rp)^{1/2} + (4rt)^{1/2}\,,\,\big(p^{1/2}+(2t)^{1/2}\big)^2\Big)\Big)\\ =&\,\mathrm{E}\Big\{\Pr\Big(\Big\|\sum_{i\in\mathscr{I}_j'}(W_{i,j}W_{i,j}^\top - \mathbf{I}_p)\Big\| > 3\max\Big((2rp)^{1/2} + (4rt)^{1/2}\,,\,\big(p^{1/2}+(2t)^{1/2}\big)^2\Big) \mid z_{1,j},\dots,z_{N,j}\Big)\Big\}\\ \le&\,2\exp(-t).\end{aligned}$$

Then from the union bound and the fact $\|\mathbf{\Sigma}_j^*\| \le \|\mathbf{\Sigma}\| \le C_2$, for $t > 0$,

$$\Pr\Big(\|\mathbf{A}_1\| > 3C_2p\max\Big((2rp)^{1/2} + (4rt)^{1/2}\,,\,\big(p^{1/2}+(2t)^{1/2}\big)^2\Big)\Big) \le 2p\exp(-t).$$

By replacing t by $t + \log(p)$, we obtain $\|\mathbf{A}_1\| = O_P(\max(pn^{1/2}, p^2)) = o\left(n + r\log\left(N/r\right)\right)$.

For $\mathbf{A}_2$, we have

$$\begin{aligned}\|\mathbf{A}_2\| =&\left\|\mathbf{\Sigma}\{\mathrm{diag}(\mathbf{\Sigma})\}^{-1/2}\begin{pmatrix}\sum_{i\in\mathscr{I}_1'}\frac{z_{i,1}-\mu_1}{\sigma_{1,1}^{1/2}}\\ \sum_{i\in\mathscr{I}_2'}\frac{z_{i,2}-\mu_2}{\sigma_{2,2}^{1/2}}\\ \vdots\\ \sum_{i\in\mathscr{I}_p'}\frac{z_{i,p}-\mu_p}{\sigma_{p,p}^{1/2}}\end{pmatrix}\mu^\top\right\|\\ \le&\|\mu\|\left\|\mathbf{\Sigma}\{\mathrm{diag}(\mathbf{\Sigma})\}^{-1/2}\right\|\Big(\sum_{j=1}^{p}\Big(\sum_{i\in\mathscr{I}_j'}\frac{z_{i,j}-\mu_j}{\sigma_{j,j}^{1/2}}\Big)^2\Big)^{1/2}\\ \le&\frac{C_2C_3}{C_1^{1/2}}p^{1/2}\max_{j\in\{1,\dots,p\}}\Big|\sum_{i\in\mathscr{I}_j'}\frac{z_{i,j}-\mu_j}{\sigma_{j,j}^{1/2}}\Big|.\end{aligned}$$

Then from (G.27), $\|\mathbf{A}_2\| = O_P\left(n^{1/2}\log(n)\right) = o_P(n + r\log\left(N/r\right))$.

For $\mathbf{A}_3$, we have

$$\begin{aligned}
\|\mathbf{A}_3\| =& \left\| \mathbf{\Sigma}\{\mathrm{diag}(\mathbf{\Sigma})\}^{-1/2} \begin{pmatrix} \sum_{i\in\mathscr{I}_1'} \frac{z_{i,1}-\mu_1}{\sigma_{1,1}^{1/2}} W_{i,1}^\top (\mathbf{\Sigma}_1^*)^{1/2} \\ \sum_{i\in\mathscr{I}_2'} \frac{z_{i,2}-\mu_2}{\sigma_{2,2}^{1/2}} W_{i,2}^\top (\mathbf{\Sigma}_2^*)^{1/2} \\ \vdots \\ \sum_{i\in\mathscr{I}_p'} \frac{z_{i,p}-\mu_p}{\sigma_{p,p}^{1/2}} W_{i,p}^\top (\mathbf{\Sigma}_p^*)^{1/2} \end{pmatrix} \right\| \\
\le& \frac{C_2}{C_1^{1/2}} \left\| \begin{pmatrix} \sum_{i\in\mathscr{I}_1'} \frac{z_{i,1}-\mu_1}{\sigma_{1,1}^{1/2}} W_{i,1}^\top (\mathbf{\Sigma}_1^*)^{1/2} \\ \sum_{i\in\mathscr{I}_2'} \frac{z_{i,2}-\mu_2}{\sigma_{2,2}^{[}1/2]} W_{i,2}^\top (\mathbf{\Sigma}_2^*)^{1/2} \\ \vdots \\ \sum_{i\in\mathscr{I}_p'} \frac{z_{i,p}-\mu_p}{\sigma_{p,p}^{1/2}} W_{i,p}^\top (\mathbf{\Sigma}_p^*)^{1/2} \end{pmatrix} \right\|_F \\
=& \frac{C_2}{C_1^{1/2}} \Big\{ \sum_{j=1}^p \Big\| \sum_{i\in\mathscr{I}_j'} \frac{z_{i,j}-\mu_j}{\sigma_{j,j}^{1/2}} W_{i,j}^\top (\mathbf{\Sigma}_j^*)^{1/2} \Big\|^2 \Big\}^{1/2} \\
\le& \frac{C_2^{3/2}}{C_1^{1/2}} \Big\{ \sum_{j=1}^p \Big\| \sum_{i\in\mathscr{I}_j'} \frac{z_{i,j}-\mu_j}{\sigma_{j,j}^{1/2}} W_{i,j} \Big\|^2 \Big\}^{1/2}.
\end{aligned}$$

Note that for $j = 1, \ldots, p$,

$$\sum_{i\in\mathscr{I}_j'} \frac{z_{i,j}-\mu_j}{\sigma_{j,j}^{1/2}} W_{i,j} \mid z_{1,j}, \ldots, z_{N,j} \sim \mathscr{N}\Big(\mathbf{0}_p, \sum_{i\in\mathscr{I}_j'} \frac{(z_{i,j}-\mu_j)^2}{\sigma_{j,j}} \mathbf{I}_p\Big).$$

From Laurent and Massart [2000], Lemma 1, for $t > 0$,

$$\Pr\Big\{ \Big\| \sum_{i\in\mathscr{I}_j'} \frac{z_{i,j}-\mu_j}{\sigma_{j,j}^{1/2}} W_{i,j} \Big\|^2 \ge \Big(\sum_{i\in\mathscr{I}_j'} \frac{(z_{i,j}-\mu_j)^2}{\sigma_{j,j}} \Big) \Big(p + 2(pt)^{1/2} + 2t \Big) \Big\} \le \exp(-t).$$

By the union bound, for $t > 0$,

$$\Pr\Big\{ \sum_{j=1}^p \Big\| \sum_{i\in\mathscr{I}_j'} \frac{z_{i,j}-\mu_j}{\sigma_{j,j}^{1/2}} W_{i,j} \Big\|^2 \ge \Big(\sum_{j=1}^p \sum_{i\in\mathscr{I}_j'} \frac{(z_{i,j}-\mu_j)^2}{\sigma_{j,j}} \Big) \Big(p + 2(pt)^{1/2} + 2t \Big) \Big\} \le p\exp(-t).$$

By replacing t by $t + \log(p)$, we obtain

$$\sum_{j=1}^p \Big\| \sum_{i\in\mathscr{I}_j'} \frac{z_{i,j}-\mu_j}{\sigma_{j,j}^{1/2}} W_{i,j} \Big\|^2 = O_P\left(p \sum_{j=1}^p \sum_{i\in\mathscr{I}_j'} \frac{(z_{i,j}-\mu_j)^2}{\sigma_{j,j}} \right).$$

It follows from the above equation and (G.30) that

$$\sum_{j=1}^p \Big\| \sum_{i\in\mathscr{I}_j'} \frac{z_{i,j}-\mu_j}{\sigma_{j,j}^{1/2}} W_{i,j} \Big\|^2 = O_P\left(np \log\left(\frac{N}{r}\right)\right) + o_P(np^2).$$

Then

$$\|\mathbf{A}_3\| = O_P\Big(\Big(np\log\Big(\frac{N}{r}\Big)\Big)^{1/2} \Big) + o_P(n^{1/2}p) = o_P\left(n + r\log\left(\frac{N}{r}\right)\right).$$

For $\mathbf{A}_4$, we have

$$\|\mathbf{A}_4\| = \|\mu\| \Big\| \sum_{j=1}^p (\mathbf{\Sigma}_j^*)^{1/2} \Big(\sum_{i\in\mathscr{I}_j'} W_{i,j} \Big) \Big\|.$$

Then from (G.28), $\|\mathbf{A}_4\| = o_P\left(n + r\log\left(N/r\right)\right)$. This completes the proof.

□

Proof of Theorem G.5. From Proposition 8, we have

$$\begin{aligned}
&\mathbf{D}_N^{-1/2}\left(\sum_{j=1}^{p}\sum_{i\in\mathscr{I}_j'}X_iX_i^\top\right)\mathbf{D}_N^{-1/2}\\
&=\begin{pmatrix}1 & \frac{1}{(n(n+4r\log(N/r)))^{1/2}}\sum_{j=1}^{p}\sum_{i\in\mathscr{I}_j'}Z_i^\top\\ \frac{1}{(n(n+4r\log(N/r)))^{1/2}}\sum_{j=1}^{p}\sum_{i\in\mathscr{I}_j'}Z_i & \frac{1}{n+4r\log(N/r)}\sum_{j=1}^{p}\sum_{i\in\mathscr{I}_j'}Z_iZ_i^\top\end{pmatrix}\\
&=\begin{pmatrix}1 & \alpha_N^{1/2}\mu^\top\\ \alpha_N^{1/2}\mu & \alpha_N\mu\mu^\top+\mathbf{W}_N\end{pmatrix}+\mathbf{E}^*,
\end{aligned}$$

where $\|\mathbf{E}^*\|=o_P(1)$. It can be seen that the matrix

$$\begin{pmatrix}1 & \alpha_N^{1/2}\mu^\top\\ \alpha_N^{1/2}\mu & \alpha_N\mu\mu^\top+\mathbf{W}_N\end{pmatrix}$$

is positive definite and its eigenvalues are bounded from 0 and infinity. And

$$\begin{pmatrix}1 & \alpha_N^{1/2}\mu^\top\\ \alpha_N^{1/2}\mu & \alpha_N\mu\mu^\top+\mathbf{W}_N\end{pmatrix}^{-1}=\begin{pmatrix}1+\alpha_N\mu^\top\mathbf{W}_N^{-1}\mu & -\alpha_N^{1/2}\mu^\top\mathbf{W}_N^{-1}\\ -\alpha_N^{1/2}\mathbf{W}_N^{-1}\mu & \mathbf{W}_N^{-1}\end{pmatrix}.$$

Then from Lemma B.4,

$$\left\|\mathbf{D}_N^{1/2}\left(\sum_{j=1}^{p}\sum_{i\in\mathscr{I}_j'}X_iX_i^\top\right)^{-1}\mathbf{D}_N^{1/2}-\begin{pmatrix}1+\alpha_N\mu^\top\mathbf{W}_N^{-1}\mu & -\alpha_N^{1/2}\mu^\top\mathbf{W}_N^{-1}\\ -\alpha_N^{1/2}\mathbf{W}_N^{-1}\mu & \mathbf{W}_N^{-1}\end{pmatrix}\right\|=o_P(1),$$

which completes the proof.

□

References

B. C. Arnold, N. Balakrishnan, and H. N. Nagaraja. *A first course in order statistics*. Society for Industrial and Applied Mathematics (SIAM), Philadelphia, PA, 2008.

R. Bhatia. *Positive definite matrices*. Princeton Series in Applied Mathematics. Princeton University Press, Princeton, NJ, 2007a.

R. Bhatia. *Perturbation Bounds for Matrix Eigenvalues*. 2007b.

S. Boucheron and M. Thomas. Concentration inequalities for order statistics. *Electronic Communications in Probability*, 17:1–12, 2012.

S. Boucheron, G. Lugosi, and P. Massart. *Concentration Inequalities: A Nonasymptotic Theory of Independence*. Oxford University Press, USA, 2013.

R. Durrett. *Probability: Theory and Examples*. Cambridge, 2019.

B. Laurent and P. Massart. Adaptive estimation of a quadratic functional by model selection. *The Annals of Statistics*, 28(5):1302–1338, Oct. 2000.

M. Mureşan. *A concrete approach to classical analysis*. CMS Books in Mathematics/Ouvrages de Mathématiques de la SMC. Springer, New York, 2009.

R. Vershynin. Introduction to the non-asymptotic analysis of random matrices. *arXiv: 1011.3027v7*, 2010.

H. Wang, M. Yang, and J. Stufken. Information-based optimal subdata selection for big data linear regression. *Journal of the American Statistical Association*, 114(525):393–405, 2019.

S. Zhang, B. Guo, A. Dong, J. He, Z. Xu, and S. X. Chen. Cautionary tales on air-quality improvement in Beijing. *Proceedings of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 473(2205):1–14, sep 2017.