Supplementary Material for "Slicing-free Inverse Regression in High-dimensional Sufficient Dimension Reduction"

Qing Mai, Xiaofeng Shao, Runmin Wang and Xin Zhang

In the supplement, we present some additional simulation results in Section S1, additional real data analysis results in Section S2, some discussion of computational complexity in Section S3, proofs of Proposition 2 & Lemma 1 in Section S4, proofs of Theorem 1 in Section S5, proofs of Theorem 2 & 3 in Section S6.

S1 Additional Simulation Results

In this section, we shall present all simulation results for models described in Section 6, except the ones which have already been presented in the main paper. Specifically, Table S1 and S2 contain the results for single index models ($\mathcal{M}_1$ and $\mathcal{M}_2$), Table S3 and S4 contain results for multiple index models ($\mathcal{M}_3$ - $\mathcal{M}_5$). Results for PFC model (univariate response) $\mathcal{M}_6$ are summarized in Table S5, and we gather the rest of results for models with multivariate response variables ($\mathcal{M}_7$-$\mathcal{M}_9$) in Table S6.

The patterns are similar to those presented in the paper. Overall the newly proposed method outperforms the competitors in most scenarios, especially when the dimensionality is significantly larger than the sample size (i.e., high-dimensional setting). It is also observed that SIR-based methods are rather sensitive to the choice of number of slices, whereas our method is slicing-free and is thus easier to use in practice.

			MDDM		Oracle-SIR(3)		Oracle-SIR(10)		Rifle-SIR(3)		Rifle-SIR(10)		LassoSIR(3)		LassoSIR(10)	
			Error	SE	Error	SE	Error	SE	Error	SE	Error	SE	Error	SE	Error	SE
		p = 200	10.3	0.1	12.6	0.1	10.5	0.1	16.4	0.6	30.0	1.2	28.5	0.2	30.6	0.3
		p = 500	10.4	0.1	12.7	0.1	10.6	0.1	22.1	0.9	47.1	1.4	34.6	0.3	47.7	0.6
	n = 200	p = 800	10.1	0.1	12.5	0.1	10.3	0.1	25.2	1.0	53.7	1.4	37.9	0.4	59.9	0.7
		p = 1200	10.0	0.1	12.4	0.1	10.3	0.1	27.5	1.0	54.3	1.4	42.4	0.5	71.4	0.7
		p = 2000	10.1	0.1	12.6	0.1	10.5	0.1	29.7	1.1	63.4	1.4	48.4	0.6	81.7	0.6
		p = 200	6.3	0.1	7.6	0.1	6.3	0.1	8.9	0.4	12.1	0.7	15.0	0.1	12.8	0.1
		p = 500	6.4	0.1	7.9	0.1	6.5	0.1	11.8	0.6	21.6	1.1	16.3	0.1	14.6	0.1
$\mathcal{M}_1$	n = 500	p = 800	6.2	0.1	7.7	0.1	6.4	0.1	13.2	0.7	26.1	1.2	16.6	0.1	16.2	0.2
		p = 1200	6.4	0.1	7.6	0.1	6.4	0.1	13.4	0.7	30.3	1.3	17.3	0.2	17.9	0.2
		p = 2000	6.3	0.1	7.7	0.1	6.3	0.1	14.9	0.8	32.9	1.3	18.3	0.2	21.8	0.3
		p = 200	5.0	0.1	6.0	0.1	5.0	0.1	7.4	0.4	8.7	0.6	11.1	0.1	9.3	0.1
		p = 500	5.2	0.1	6.1	0.1	5.1	0.1	8.3	0.4	12.9	0.8	11.9	0.1	10.1	0.1
	n =800	p = 800	5.1	0.1	6.1	0.1	5.1	0.1	9.5	0.6	20.1	1.1	12.4	0.1	11.1	0.1
		p = 1200	5.1	0.1	6.1	0.1	5.1	0.1	9.5	0.6	20.1	1.1	12.4	0.1	11.1	0.1
		p = 2000	4.9	0.1	6.1	0.1	5.0	0.1	10.6	0.6	23.1	1.2	12.8	0.1	12.2	0.1
		p = 200	10.4	0.1	13.1	0.1	10.7	0.1	17.5	0.6	29.7	1.2	30.3	0.2	31.6	0.3
		p = 500	10.6	0.1	13.3	0.1	10.8	0.1	23.8	0.9	48.9	1.4	36.7	0.3	49.9	0.6
	n = 200	p = 800	10.3	0.1	13.1	0.1	10.6	0.1	26.1	1.0	54.7	1.4	40.1	0.4	61.5	0.7
		p = 1200	54.7	0.8	12.9	0.1	10.4	0.1	74.4	0.8	95.1	0.4	45.0	0.5	71.4	0.7
		p = 2000	55.3	0.8	13.1	0.1	10.6	0.1	76.5	0.7	96.8	0.3	51.2	0.6	82.7	0.6
		p = 200	6.4	0.1	8.0	0.1	6.5	0.1	9.7	0.4	12.0	0.7	15.8	0.1	13.4	0.1
		p = 500	6.7	0.1	8.2	0.1	6.6	0.1	12.4	0.6	21.2	1.1	17.1	0.1	15.1	0.1
$\mathcal{M}_2$	n = 500	p = 800	6.5	0.1	8.0	0.1	6.5	0.1	13.6	0.7	25.1	1.2	17.6	0.2	16.7	0.2
		p = 1200	17.1	0.7	8.0	0.1	6.5	0.1	37.4	1.1	74.6	1.0	18.2	0.2	18.3	0.2
		p = 2000	16.9	0.8	8.0	0.1	6.5	0.1	38.3	1.2	77.4	0.9	19.0	0.2	22.2	0.3
		p = 200	5.1	0.1	6.3	0.1	5.1	0.1	6.9	0.3	8.3	0.5	11.6	0.1	9.6	0.1
		p = 500	5.3	0.1	6.4	0.1	5.2	0.1	7.8	0.4	13.1	0.8	12.5	0.1	10.5	0.1
	n =800	p = 800	5.0	0.1	6.3	0.1	5.1	0.1	8.4	0.4	16.9	1.0	12.6	0.1	10.8	0.1
		p = 1200	10.8	0.6	6.4	0.1	5.2	0.1	23.9	1.0	53.9	1.3	13.1	0.1	11.4	0.1
		p = 2000	11.3	0.6	6.3	0.1	5.1	0.1	26.6	1.1	57.6	1.2	13.6	0.1	12.4	0.1

Table S1: $d(V, \hat{V})$ and corresponding standard errors (in 10^{-2}) for single index models (identity variance)

			MDDM		Oracle-SIR(3)		Oracle-SIR(10)		Rifle-SIR(3)		Rifle-SIR(10)		LassoSIR(3)		LassoSIR(10)	
			Error	SE	Error	SE	Error	SE	Error	SE	Error	SE	Error	SE	Error	SE
		p = 200	17.6	0.2	20.8	0.2	17.8	0.2	26.2	0.5	25.3	0.7	32.1	0.3	28.4	0.3
		p = 500	18.3	0.3	21.1	0.2	18.0	0.2	32.9	0.7	35.0	1.0	34.5	0.3	33.1	0.3
	n = 200	p = 800	18.7	0.3	21.0	0.2	17.6	0.2	34.7	0.8	39.8	1.1	35.3	0.3	35.5	0.3
		p = 1200	26.1	0.6	21.3	0.2	18.0	0.2	47.0	1.0	81.2	0.7	36.7	0.3	41.2	0.4
		p = 2000	26.3	0.7	21.0	0.2	17.7	0.2	47.8	1.0	81.4	0.7	39.1	0.4	49.6	0.6
		p = 200	10.8	0.1	13.2	0.1	11.0	0.1	13.7	0.2	12.7	0.4	18.6	0.2	15.3	0.1
		p = 500	11.0	0.1	13.3	0.1	11.2	0.1	14.3	0.3	15.7	0.6	19.5	0.2	16.5	0.2
$\mathcal{M}_1$	n = 500	p = 800	10.9	0.1	13.5	0.1	11.1	0.1	15.1	0.4	18.4	0.8	20.1	0.2	17.1	0.2
		p = 1200	14.2	0.5	13.4	0.1	11.1	0.1	23.1	0.8	49.9	1.1	20.3	0.2	17.8	0.2
		p = 2000	13.6	0.5	13.5	0.1	11.2	0.1	26.9	0.9	53.8	1.2	20.7	0.2	18.6	0.2
		p = 200	8.8	0.1	10.7	0.1	8.9	0.1	10.6	0.1	9.7	0.3	14.5	0.1	11.9	0.1
		p = 500	8.7	0.1	10.5	0.1	8.9	0.1	11.1	0.3	9.9	0.3	14.8	0.1	12.5	0.1
	n =800	p = 800	8.6	0.1	10.5	0.1	8.8	0.1	11.7	0.4	12.4	0.6	15.0	0.1	12.5	0.1
		p = 1200	10.2	0.4	10.4	0.1	8.8	0.1	18.3	0.8	37.9	1.1	15.1	0.1	12.9	0.1
		p = 2000	11.1	0.5	10.6	0.1	8.8	0.1	20.5	0.8	38.2	1.1	15.8	0.1	13.4	0.1
		p = 200	14.0	0.2	20.8	0.2	14.8	0.2	25.4	0.5	21.4	0.7	31.8	0.3	23.3	0.2
		p = 500	14.3	0.2	21.1	0.2	15.0	0.2	31.4	0.7	29.0	1.0	34.1	0.3	27.4	0.3
	n = 200	p = 800	14.2	0.2	20.7	0.2	14.8	0.2	33.1	0.7	33.6	1.1	34.6	0.3	30.5	0.3
		p = 1200	23.9	0.7	21.1	0.2	14.9	0.2	45.1	1.0	75.9	0.9	36.9	0.3	34.9	0.4
		p = 2000	24.9	0.8	20.6	0.2	14.6	0.2	47.9	1.0	77.2	0.9	38.6	0.4	42.5	0.5
		p = 200	8.7	0.1	13.2	0.1	9.0	0.1	13.7	0.3	10.4	0.4	18.6	0.2	12.6	0.1
		p = 500	8.9	0.1	13.3	0.1	9.3	0.1	14.3	0.3	13.4	0.6	19.4	0.2	13.4	0.1
$\mathcal{M}_2$	n = 500	p = 800	8.8	0.1	13.3	0.1	9.1	0.1	14.0	0.3	14.9	0.7	19.9	0.2	13.8	0.1
		p = 1200	12.8	0.5	13.3	0.1	9.2	0.1	24.3	0.8	45.9	1.2	20.1	0.2	14.6	0.1
		p = 2000	12.5	0.5	13.5	0.1	9.3	0.1	27.0	0.9	49.7	1.2	20.6	0.2	15.1	0.1
		p = 200	7.1	0.1	10.6	0.1	7.4	0.1	10.6	0.1	7.9	0.2	14.4	0.1	9.8	0.1
		p = 500	7.1	0.1	10.6	0.1	7.3	0.1	10.9	0.2	9.3	0.4	15.0	0.1	10.2	0.1
	n =800	p = 800	7.1	0.1	10.6	0.1	7.3	0.1	11.6	0.3	10.8	0.6	15.1	0.1	10.2	0.1
		p = 1200	8.8	0.4	10.4	0.1	7.2	0.1	17.5	0.7	32.8	1.1	15.0	0.1	10.5	0.1
		p = 2000	9.7	0.5	10.5	0.1	7.3	0.1	19.7	0.8	34.3	1.2	15.9	0.1	10.8	0.1

Table S2: $d(V, \hat{V})$ and corresponding standard errors (in 10^{-2}) for single index models (AR-type variance)

			MDDM		Oracle-SIR(3)		Oracle-SIR(10)		Rifle-SIR(3)		Rifle-SIR(10)		LassoSIR(3)		LassoSIR(10)	
			Error	SE	Error	SE	Error	SE	Error	SE	Error	SE	Error	SE	Error	SE
		p = 200	17.5	0.2	40.6	0.2	27.7	0.2	71.3	0.0	71.2	0.0	69.6	0.2	67.0	0.3
		p = 500	18.1	0.2	40.7	0.2	28.2	0.2	71.3	0.0	71.2	0.0	74.8	0.2	80.1	0.3
	n = 200	p = 800	17.7	0.2	40.8	0.2	27.7	0.2	71.3	0.0	71.2	0.0	76.5	0.2	85.0	0.2
		p = 1200	17.9	0.2	40.7	0.2	28.0	0.2	31.7	0.4	18.6	0.2	78.4	0.2	88.8	0.2
		p = 2000	18.1	0.2	40.8	0.2	28.0	0.2	32.1	0.4	18.9	0.2	80.6	0.2	92.5	0.2
		p = 200	10.6	0.1	27.8	0.2	17.0	0.1	70.9	0.0	70.9	0.0	48.6	0.3	30.1	0.2
		p = 500	10.8	0.1	27.5	0.2	17.2	0.1	70.9	0.0	70.9	0.0	53.7	0.3	39.1	0.3
$\mathcal{M}_3$	n = 500	p = 800	10.8	0.1	27.4	0.2	17.3	0.1	70.9	0.0	70.9	0.0	57.1	0.2	46.0	0.4
		p = 1200	10.7	0.1	27.4	0.2	17.1	0.1	18.9	0.2	11.0	0.1	59.1	0.2	53.0	0.4
		p = 2000	10.7	0.1	27.6	0.2	17.1	0.1	19.0	0.2	11.2	0.1	62.1	0.2	60.7	0.4
		p = 200	8.2	0.1	22.0	0.1	13.5	0.1	70.8	0.0	70.8	0.0	36.0	0.2	20.8	0.1
		p = 500	8.3	0.1	21.9	0.1	13.5	0.1	70.8	0.0	70.8	0.0	40.4	0.2	24.0	0.2
	n =800	p = 800	8.2	0.1	21.9	0.1	13.3	0.1	70.8	0.0	70.8	0.0	42.2	0.3	26.2	0.2
		p = 1200	8.3	0.1	22.0	0.1	13.4	0.1	14.9	0.1	8.7	0.1	44.7	0.2	29.8	0.3
		p = 2000	8.3	0.1	22.2	0.1	13.5	0.1	14.8	0.1	8.6	0.1	47.9	0.3	36.4	0.4
		p = 200	23.1	0.2	46.2	0.3	36.3	0.3	72.0	0.0	71.6	0.0	78.1	0.2	78.2	0.3
		p = 500	22.8	0.2	45.8	0.3	35.9	0.3	72.1	0.0	71.6	0.0	83.0	0.2	87.8	0.2
	n = 200	p = 800	23.0	0.2	45.8	0.3	36.4	0.3	71.9	0.0	71.6	0.0	85.2	0.2	91.5	0.2
		p = 1200	23.2	0.3	45.8	0.3	36.3	0.3	38.1	0.4	25.2	0.3	87.3	0.2	93.8	0.2
		p = 2000	23.1	0.3	45.7	0.3	36.2	0.3	54.0	0.5	34.1	0.5	89.2	0.2	95.9	0.1
		p = 200	13.4	0.1	31.0	0.2	21.7	0.1	71.2	0.0	71.0	0.0	55.3	0.3	43.2	0.3
		p = 500	13.5	0.1	31.0	0.2	21.6	0.1	71.2	0.0	71.0	0.0	61.3	0.3	55.3	0.4
$\mathcal{M}_4$	n = 500	p = 800	13.4	0.1	31.0	0.2	21.8	0.1	71.2	0.0	71.0	0.0	64.7	0.2	62.8	0.4
		p = 1200	13.4	0.1	30.8	0.2	21.6	0.1	21.5	0.2	14.3	0.1	67.2	0.2	67.8	0.3
		p = 2000	13.5	0.1	31.0	0.2	21.8	0.1	22.0	0.2	14.4	0.1	69.7	0.2	73.3	0.3
		p = 200	10.5	0.1	25.2	0.2	17.2	0.1	71.0	0.0	70.9	0.0	42.7	0.2	28.7	0.2
		p = 500	10.4	0.1	24.9	0.2	17.1	0.1	71.0	0.0	70.9	0.0	47.3	0.3	34.2	0.3
	n =800	p = 800	10.5	0.1	25.1	0.2	17.1	0.1	71.0	0.0	70.9	0.0	50.2	0.3	39.9	0.4
		p = 1200	10.3	0.1	24.9	0.2	17.0	0.1	16.9	0.2	11.0	0.1	52.4	0.3	45.8	0.4
		p = 2000	10.6	0.1	25.2	0.2	17.2	0.1	17.3	0.2	11.3	0.1	56.5	0.3	53.8	0.4
		p = 200	30.6	0.6	29.1	0.1	22.0	0.1	71.6	0.0	71.2	0.0	58.8	0.3	56.6	0.3
		p = 500	30.4	0.6	28.9	0.2	22.2	0.1	71.5	0.0	71.2	0.0	67.4	0.3	73.5	0.4
	n = 200	p = 800	30.8	0.6	28.8	0.2	22.1	0.1	71.6	0.0	71.2	0.0	71.2	0.3	81.3	0.3
		p= 1200	31.0	0.6	29.1	0.1	22.3	0.1	20.0	0.2	14.6	0.1	74.1	0.3	86.4	0.3
		p = 2000	31.3	0.6	28.7	0.2	22.1	0.1	19.3	0.2	14.4	0.1	77.8	0.3	90.3	0.2
		p = 200	12.4	0.2	18.4	0.1	13.6	0.1	71.1	0.0	70.9	0.0	31.7	0.2	24.5	0.2
		p = 500	11.8	0.2	18.4	0.1	13.6	0.1	71.0	0.0	70.9	0.0	35.9	0.2	29.2	0.2
$\mathcal{M}_5$	n = 500	p = 800	11.9	0.2	18.4	0.1	13.6	0.1	71.0	0.0	70.9	0.0	37.9	0.3	32.9	0.3
		p = 1200	11.6	0.2	18.4	0.1	13.7	0.1	12.1	0.1	8.6	0.1	39.7	0.3	37.3	0.3
		p = 2000	12.0	0.2	18.5	0.1	13.6	0.1	12.1	0.1	8.6	0.1	42.1	0.3	46.3	0.4
		p = 200	7.9	0.1	14.7	0.1	10.7	0.1	70.9	0.0	70.8	0.0	23.8	0.1	17.4	0.1
		p = 500	7.8	0.1	14.5	0.1	10.6	0.1	70.9	0.0	70.8	0.0	25.7	0.2	19.2	0.1
	n = 800	p = 800	7.8	0.1	14.6	0.1	10.7	0.1	70.9	0.0	70.8	0.0	27.1	0.2	20.7	0.2
		p = 1200	7.7	0.1	14.6	0.1	10.6	0.1	9.3	0.1	6.5	0.1	28.3	0.2	21.6	0.2
		p = 2000	7.9	0.1	14.4	0.1	10.7	0.1	9.4	0.1	6.6	0.1	29.9	0.2	25.0	0.2

Table S3: $d(V, \hat{V})$ and corresponding standard errors (in 10^{-2}) for multiple index models (identity variance)

			MDDM		Oracle-SIR(3)		Oracle-SIR(10)		Rifle-SIR(3)		Rifle-SIR(10)		LassoSIR(3)		LassoSIR(10)	
			Error	SE	Error	SE	Error	SE	Error	SE	Error	SE	Error	SE	Error	SE
		p = 200	41.4	0.4	58.8	0.2	50.2	0.2	72.6	0.0	72.2	0.0	67.0	0.2	62.8	0.2
		p = 500	42.1	0.4	59.0	0.2	50.4	0.2	72.9	0.1	72.4	0.0	70.3	0.2	71.1	0.2
	n = 200	p = 800	43.0	0.4	59.1	0.2	50.4	0.2	72.7	0.0	72.3	0.0	72.1	0.2	75.1	0.2
		p = 1200	42.7	0.4	59.4	0.2	50.8	0.2	53.0	0.4	39.9	0.4	73.0	0.2	78.6	0.2
		p = 2000	43.5	0.4	59.2	0.2	50.8	0.2	53.5	0.4	40.2	0.4	74.8	0.2	82.1	0.2
		p = 200	25.6	0.3	44.6	0.2	34.3	0.2	71.4	0.0	71.3	0.0	51.5	0.2	42.4	0.2
		p = 500	25.8	0.3	44.4	0.2	34.5	0.2	71.5	0.0	71.3	0.0	53.8	0.2	45.7	0.2
$\mathcal{M}_3$	n = 500	p = 800	25.2	0.3	44.6	0.2	34.1	0.2	71.5	0.0	71.3	0.0	54.8	0.2	47.1	0.3
		p = 1200	25.9	0.3	44.4	0.2	34.4	0.2	33.8	0.4	23.8	0.2	55.8	0.2	50.1	0.3
		p = 2000	25.7	0.3	44.3	0.2	34.4	0.2	35.0	0.4	23.8	0.2	56.9	0.2	54.0	0.3
		p = 200	20.1	0.2	37.0	0.2	27.8	0.2	71.2	0.0	71.0	0.0	43.8	0.2	34.9	0.2
		p = 500	19.8	0.2	37.8	0.2	27.6	0.2	71.2	0.0	71.1	0.0	46.1	0.2	36.6	0.2
	n =800	p = 800	20.0	0.2	37.3	0.2	27.9	0.2	71.2	0.0	71.0	0.0	47.0	0.2	37.9	0.2
		p = 1200	19.8	0.2	37.1	0.2	27.8	0.2	26.7	0.3	18.5	0.2	47.6	0.2	38.6	0.2
		p = 2000	19.9	0.2	37.4	0.2	27.7	0.2	27.5	0.3	18.7	0.2	48.8	0.2	40.7	0.2
		p = 200	58.1	0.4	75.9	0.2	70.1	0.3	80.4	0.2	78.2	0.2	85.7	0.2	84.6	0.2
		p = 500	59.3	0.5	75.8	0.2	70.2	0.3	95.2	0.1	94.2	0.1	88.8	0.2	90.2	0.2
	n = 200	p = 800	59.1	0.5	75.1	0.2	69.9	0.3	81.0	0.2	78.7	0.2	89.7	0.2	92.1	0.2
		p = 1200	59.5	0.5	75.3	0.2	69.9	0.3	73.9	0.3	62.3	0.5	90.5	0.2	93.9	0.2
		p = 2000	60.4	0.5	75.4	0.2	69.9	0.3	74.1	0.4	63.0	0.5	91.9	0.2	95.2	0.1
		p = 200	38.6	0.4	61.4	0.2	50.9	0.3	74.5	0.1	73.5	0.1	70.5	0.2	63.0	0.3
		p = 500	38.7	0.4	61.6	0.2	51.1	0.3	74.7	0.1	73.4	0.1	74.0	0.2	69.0	0.3
$\mathcal{M}_4$	n = 500	p = 800	39.3	0.4	61.3	0.2	50.8	0.2	77.5	0.2	74.9	0.1	75.6	0.2	72.5	0.3
		p = 1200	39.5	0.4	61.3	0.2	51.0	0.2	54.9	0.4	40.2	0.4	77.1	0.2	76.0	0.3
		p = 2000	39.7	0.4	61.5	0.2	51.3	0.3	56.1	0.4	40.6	0.4	78.7	0.2	79.7	0.3
		p = 200	30.7	0.3	53.1	0.2	42.2	0.2	73.1	0.1	72.4	0.0	61.3	0.2	51.4	0.3
		p = 500	30.5	0.3	52.9	0.2	42.0	0.2	73.1	0.1	72.4	0.0	64.1	0.2	54.4	0.3
	n =800	p = 800	30.6	0.3	53.7	0.2	42.3	0.2	73.4	0.1	72.5	0.0	65.9	0.2	58.2	0.3
		p = 1200	30.8	0.3	53.5	0.2	42.2	0.2	45.3	0.4	31.5	0.3	67.2	0.2	60.8	0.3
		p = 2000	30.7	0.3	53.6	0.2	42.2	0.2	45.2	0.4	31.1	0.3	68.6	0.2	64.6	0.3
		p = 200	45.5	0.6	46.5	0.2	35.7	0.2	73.9	0.1	72.4	0.0	62.6	0.2	51.8	0.2
		p = 500	46.9	0.6	46.9	0.2	35.8	0.2	73.9	0.1	72.4	0.0	65.6	0.2	57.4	0.3
	n = 200	p = 800	46.2	0.6	46.4	0.2	35.5	0.2	73.8	0.1	72.4	0.0	66.5	0.2	61.4	0.3
		p = 1200	46.9	0.6	46.3	0.2	35.8	0.2	33.0	0.3	23.9	0.2	67.2	0.3	65.7	0.3
		p = 2000	46.1	0.6	46.5	0.2	35.7	0.2	33.6	0.3	23.8	0.2	68.8	0.3	72.0	0.3
		p = 200	24.7	0.4	31.4	0.2	22.6	0.1	72.0	0.0	71.3	0.0	44.8	0.2	32.6	0.1
		p = 500	24.7	0.4	31.1	0.2	22.4	0.1	71.9	0.0	71.3	0.0	46.8	0.2	35.4	0.2
$\mathcal{M}_5$	n = 500	p = 800	25.4	0.4	31.3	0.2	22.9	0.1	72.0	0.0	71.3	0.0	48.0	0.2	36.7	0.2
		p = 1200	25.2	0.4	31.5	0.2	22.7	0.1	20.8	0.2	14.3	0.1	48.2	0.2	37.6	0.2
		p = 2000	25.5	0.4	31.7	0.2	22.8	0.1	20.9	0.2	14.4	0.1	49.1	0.2	39.5	0.2
		p = 200	18.7	0.3	25.6	0.1	18.0	0.1	71.5	0.0	71.1	0.0	36.7	0.1	25.6	0.1
		p = 500	18.1	0.3	25.3	0.1	17.7	0.1	71.5	0.0	71.1	0.0	38.4	0.2	27.1	0.1
	n =800	p = 800	18.6	0.3	25.4	0.1	18.0	0.1	71.5	0.0	71.1	0.0	39.3	0.2	28.4	0.1
		p = 1200	18.7	0.3	25.3	0.1	17.9	0.1	16.3	0.1	11.0	0.1	40.2	0.2	29.2	0.1
		p = 2000	19.1	0.3	25.3	0.1	18.0	0.1	16.5	0.1	11.2	0.1	41.0	0.2	30.3	0.1

Table S4: $d(V, \hat{V})$ and corresponding standard errors (in 10^{-2}) for multiple index models (AR-type variance)

			MDDM		Oracle-SIR(3)		Oracle-SIR(10)		Rifle-SIR(3)		Rifle-SIR(10)		LassoSIR(3)		LassoSIR(10)	
			Error	SE	Error	SE	Error	SE	Error	SE	Error	SE	Error	SE	Error	SE
$\mathcal{M}_6$	n = 200	p = 200	34.3	0.5	49.1	0.5	33.5	0.4	50.0	0.7	30.6	0.5	70.7	0.0	70.7	0.0
		p = 500	34.2	0.5	48.7	0.5	32.8	0.4	49.5	0.7	30.1	0.5	70.7	0.0	70.7	0.0
		p = 800	34.6	0.6	48.9	0.5	33.4	0.5	50.1	0.7	30.8	0.6	70.7	0.0	70.7	0.0
		p = 1200	44.5	0.7	49.3	0.5	33.6	0.5	67.0	0.7	38.9	0.7	70.7	0.0	70.7	0.0
		p = 2000	44.8	0.8	48.1	0.5	33.2	0.5	66.7	0.8	39.4	0.7	70.7	0.0	70.7	0.0
	n = 500	p = 200	22.0	0.4	35.5	0.4	22.6	0.3	33.0	0.5	19.0	0.3	70.7	0.0	70.7	0.0
		p = 500	21.8	0.4	34.7	0.4	22.3	0.3	32.4	0.5	18.8	0.4	70.7	0.0	70.7	0.0
		p = 800	21.7	0.3	34.6	0.4	22.5	0.3	32.3	0.5	18.9	0.3	70.7	0.0	70.7	0.0
		p = 1200	27.1	0.5	35.5	0.4	22.5	0.3	43.4	0.7	23.4	0.4	70.7	0.0	70.7	0.0
		p = 2000	26.9	0.5	34.8	0.4	22.4	0.3	42.3	0.7	23.6	0.5	70.7	0.0	70.7	0.0
	n =800	p = 200	17.2	0.3	29.2	0.3	18.6	0.3	25.8	0.4	14.9	0.2	70.7	0.0	70.7	0.0
		p = 500	16.7	0.3	28.5	0.3	18.1	0.2	25.3	0.4	14.3	0.3	70.7	0.0	70.7	0.0
		p = 800	16.5	0.2	28.4	0.3	17.9	0.2	25.0	0.4	14.1	0.2	70.7	0.0	70.7	0.0
		p =1200	20.8	0.4	29.0	0.4	18.2	0.3	31.7	0.5	18.5	0.4	70.7	0.0	70.7	0.0
		p = 2000	21.2	0.4	28.7	0.4	18.3	0.3	32.4	0.6	18.8	0.3	70.7	0.0	70.7	0.0

Table S5: $d(V, \hat{V})$ and corresponding standard errors (in 10^{-2}) for isotropic PFC models ($\mathcal{M}_6$)

		$n = 100$				$n = 200$				$n = 400$			
		$p = 800$		$p = 1200$		$p = 800$		$p = 1200$		$p = 800$		$p = 1200$	
		Error	SE	Error	SE	Error	SE	Error	SE	Error	SE	Error	SE
$\mathcal{M}_7$	MDDM	45.0	0.5	45.9	0.5	27.5	0.3	28.7	0.4	18.8	0.3	19.6	0.3
	PR-Oracle-SIR(3)	26.3	0.2	26.5	0.2	18.3	0.1	18.4	0.2	12.6	0.1	12.6	0.1
	PR-Oracle-SIR(10)	33.0	0.3	33.0	0.3	20.1	0.2	20.2	0.2	13.0	0.1	13.1	0.1
	PR-SIR(3)	96.6	0.0	97.7	0.0	93.2	0.0	95.3	0.0	87.6	0.0	91.1	0.0
	PR-SIR(10)	97.4	0.0	98.2	0.0	95.0	0.0	96.6	0.0	90.0	0.1	93.7	0.0
$\mathcal{M}_8$	MDDM	93.5	0.6	94.9	0.5	73.6	1.1	77.1	1.1	36.7	1.1	37.8	1.2
	PR-Oracle-SIR(3)	80.1	0.6	80.1	0.6	64.0	0.7	64.7	0.7	42.5	0.5	41.1	0.5
	PR-Oracle-SIR(10)	79.1	0.6	78.9	0.6	58.4	0.6	58.7	0.6	34.5	0.4	34.2	0.4
	PR-SIR(3)	99.9	0.0	100.0	0.0	99.9	0.0	100.0	0.0	99.9	0.0	99.9	0.0
	PR-SIR(10)	99.9	0.0	100.0	0.0	99.9	0.0	100.0	0.0	99.9	0.0	100.0	0.0
$\mathcal{M}_9$	MDDM	17.3	0.4	17.5	0.4	10.0	0.1	10.1	0.1	7.1	0.1	7.1	0.1
	PR-Oracle-SIR(3)	15.5	0.2	15.2	0.2	10.6	0.1	10.6	0.1	7.5	0.1	7.5	0.1
	PR-Oracle-SIR(10)	14.2	0.2	13.9	0.2	9.6	0.1	9.6	0.1	6.8	0.1	6.8	0.1
	PR-SIR(3)	92.2	0.1	95.0	0.1	83.0	0.1	88.3	0.1	71.0	0.1	77.9	0.1
	PR-SIR(10)	90.0	0.1	93.4	0.1	80.1	0.1	85.8	0.1	67.4	0.1	74.7	0.1

Table S6: Averaged subspace estimation errors and the corresponding standard errors (after multiplied by 100) for multivariate response models.

S2 More on Real Data Analysis

S2.1 Choice of Tuning Parameter on Real Data

To apply our proposal on real data, we need to determine the tuning parameter, s, i.e, the desired level of sparsity. In penalized problems such as sparse PCA and sparse SIR, tuning parameters are often chosen with cross-validation. We could also employ cross-validation to choose s. However, as with almost any procedure, cross-validation would considerably slow down the computation. Moreover, as we observe in our theoretical and simulation studies, our method is not very sensitive to s; the result is reasonably stable as long as s is larger than d. Hence, we resort to a faster tuning method on our real data as follows. We start with a sequence of reasonable sparsity levels $\mathcal{S}$, which is set to be $\{1, \ldots, 45\}$. Then for each element in $\mathcal{S}$, we calculate $\hat{\boldsymbol{\beta}}$ and the sample distance covariance (Székely et al. 2007) between $\mathbf{Y}_i$ and $\hat{\boldsymbol{\beta}}^T\mathbf{X}_i$ for all $i = 1, 2, ..., n$. Here distance covariance is used as a model-free measure of the dependence between $\mathbf{Y}_i$ and $\hat{\boldsymbol{\beta}}^T\mathbf{X}_i$. Intuitively, the distance covariance increases as the pre-specified sparsity increases. Therefore, we plot the sample distance covariance against the sparsity levels in Figure S1, and pick the sparsity corresponding to a large enough distance covariance, while any larger sparsity levels will not lead to a significantly larger distance covariance (i.e. the "elbow method"). Based on Figure S1, a sparsity level between 20 and 25 seems to be reasonable, and we pick $s = 25$ as the pre-specified sparsity.

S2.2 Additional Real Data Analysis Results

To further demonstrate our methods on the real data, we also construct the scatterplot from the leading two directions $\boldsymbol{\beta}_1^{\mathrm{T}}\mathbf{X}$ and $\boldsymbol{\beta}_2^{\mathrm{T}}\mathbf{X}$ in Figure S2. For the first direction, the experimental units from leukemia cancer (LE) and colorectal cancer (CO) are on the different side of the plot. For the second direction, CO and LE have similar values whereas the units obtained from central nervous system cancer (CN), breast cancer (BR), lung cancer (LC) and ovarian cancer (OV) are on the opposite side of the figure. This pattern coincides with the first two canonical correlation analysis directions in a previous study of the same data set (Figure 8, Cruz-Cano & Lee 2014). Such a finding is very encouraging as our slicing-free approach automatically detect the most significant associations between the response and the predictor, while directly applied to the multivariate

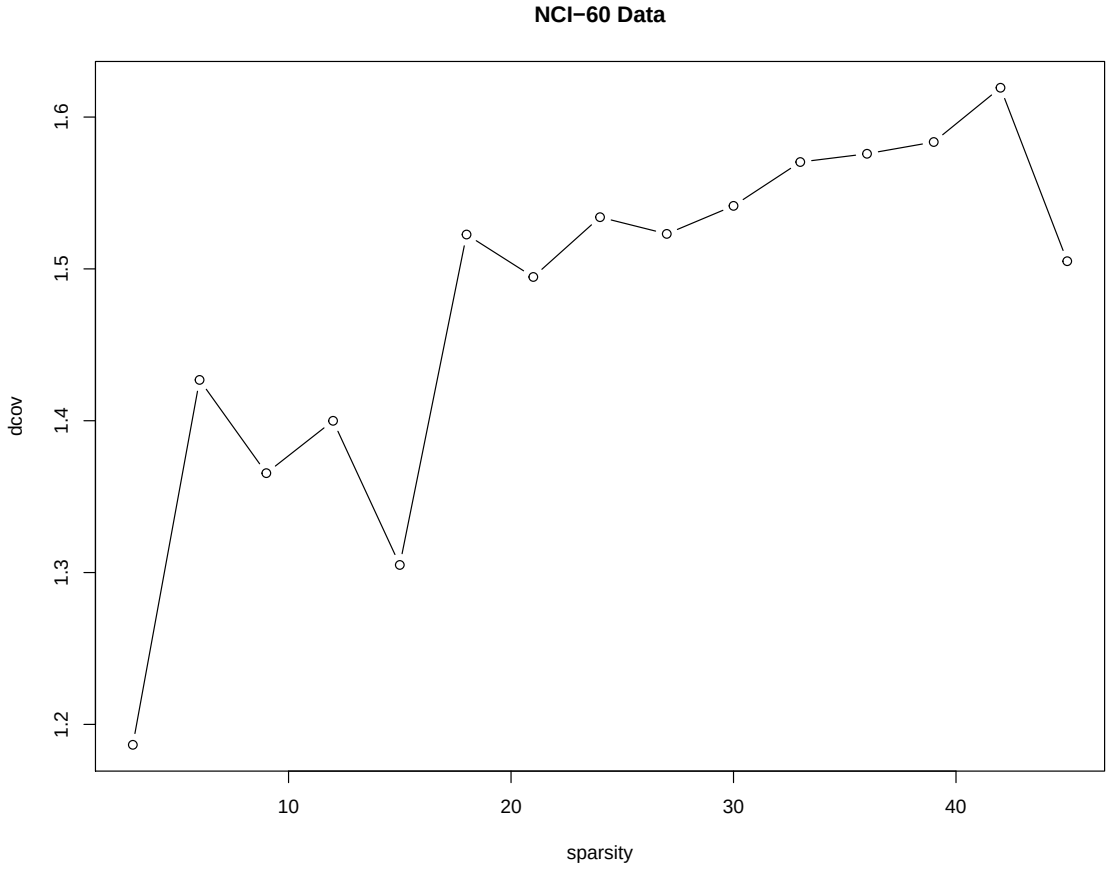

Figure S1: Distance covariance for different pre-specified sparsities.

response.

S3 Computation complexity

We briefly analyze the computational complexity of our proposed methods. For both algorithms, we need to compute the sample MDDM. The current computational complexity of sample MDDM is $O(n^2p)$. If we adapt the fast computing algorithm of Huo & Székely (2016) developed for distance correlation to MDDM, we might be able to reduce the complexity to $O(pn \log n)$. In Algorithm 2, we further need to compute the sample covariance at the complexity level of $O(np^2)$.

To apply the two algorithms, we assume that the maximum number of iterations is T in finding the K directions (Step 3(a) in both algorithms). After obtaining MDDM, Algorithm 1 has a computational complexity of $O(KT(ps+p)+(K-1)(s^2+ps))$, where $(ps+p)$ is the computation

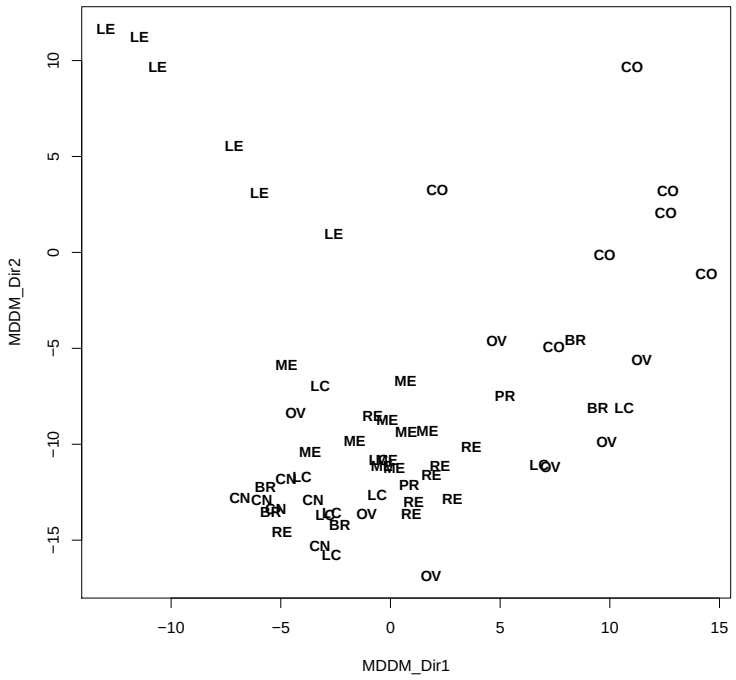

Figure S2: The scatter plot between the first leading directions.

complexity of each iteration (Yuan & Zhang 2013) and $(s^2 + ps)$ is the computation complexity of deflating $\widehat{\mathrm{M}}_k$ in Step 3(b). Note that, in Step 3(b), we repeatedly exploit the sparsity of $\widehat{\boldsymbol{\beta}}_k$ to reduce the computation complexity. For example, to compute $\widehat{\boldsymbol{\beta}}_k\widehat{\boldsymbol{\beta}}_k^{\mathrm{T}}$, we only need to compute the s^2 nonzero elements, and the same applies to $\widehat{\boldsymbol{\beta}}_k^{\mathrm{T}}\widehat{\mathrm{M}}\widehat{\boldsymbol{\beta}}_k$, which has a computational complexity of $O(ps)$. For Algorithm 2, the computational complexity is $O(KT(ps + p) + (K - 1)(p^2 + ps))$. Note that this computational complexity only differs from that of Algorithm 1 in the term p^2. This term is the cost to deflate $\widehat{\mathrm{M}}$ in Step 3(b). Since $\widehat{\boldsymbol{\Sigma}}_{\mathbf{X}}\widehat{\boldsymbol{\beta}}_k\widehat{\boldsymbol{\beta}}_k^{\mathrm{T}}\widehat{\boldsymbol{\Sigma}}_{\mathbf{X}}$ is not guaranteed to be sparse as $\widehat{\boldsymbol{\beta}}_k\widehat{\boldsymbol{\beta}}_k^{\mathrm{T}}$, the deflation is more computationally expensive. Otherwise, each iteration in Step 3(a) of Algorithm 2 has the same complexity as Step 3(a) of Algorithm 1 (Tan et al. 2018).

S4 Proofs for Proposition 2 & Lemma 1

Proof of Proposition 2. From the basic properties of MDDM (c.f. beginning of Section 1), $\mathrm{E}(\mathbf{A}^{\mathrm{T}}\mathbf{X} \mid Y) = \mathrm{E}(\mathbf{A}^{\mathrm{T}}\mathbf{X})$ is equivalent to $\mathrm{MDDM}(\mathbf{A}^{\mathrm{T}}\mathbf{X} \mid Y) = 0$.

Suppose the rank of $\mathrm{MDDM}(\mathbf{X} \mid Y)$ is d, then there exists an orthogonal basis matrix for

$\mathbb{R}^p$, $(\boldsymbol{\beta}, \boldsymbol{\beta}_0)$ with $\boldsymbol{\beta} \in \mathbb{R}^{p\times d}$ and $\boldsymbol{\beta}_0 \in \mathbb{R}^{p\times(p-d)}$, such that $\mathrm{span}(\boldsymbol{\beta}) = \mathrm{span}\{\mathrm{MDDM}(\mathbf{X} \mid Y)\}$. This implies $\boldsymbol{\beta}_0^{\mathrm{T}}\mathrm{MDDM}(\mathbf{X} \mid Y) = 0$ and equivalently $\boldsymbol{\beta}_0^{\mathrm{T}}\{\mathrm{E}(\mathbf{X} \mid Y) - \mathrm{E}(\mathbf{X})\} = 0$. Therefore, $\mathrm{span}(\boldsymbol{\beta}_0) \subseteq \mathcal{S}^{\perp}_{\mathrm{E}(\mathbf{X}|Y)}$, which leads to $\mathcal{S}_{\mathrm{E}(\mathbf{X}|Y)} \equiv \mathrm{span}\{\mathrm{E}(\mathbf{X} \mid Y) - \mathrm{E}(\mathbf{X})\} \subseteq \mathrm{span}(\boldsymbol{\beta})$.

Similarly, for any vector $\mathbf{v} \in \mathcal{S}^{\perp}_{\mathrm{E}(\mathbf{X}|Y)}$ we have $\mathbf{v}^{\mathrm{T}}\{\mathrm{E}(\mathbf{X} \mid Y) - \mathrm{E}(\mathbf{X})\} = 0$ and hence $\mathbf{v}^{\mathrm{T}}\mathrm{MDDM}(\mathbf{X} \mid Y)\mathbf{v} = 0$. This implies that $\mathbf{v} \in \mathrm{span}(\boldsymbol{\beta}_0)$ and hence, $\mathcal{S}^{\perp}_{\mathrm{E}(\mathbf{X}|Y)} \subseteq \mathrm{span}(\boldsymbol{\beta}_0)$ and $\mathrm{span}(\boldsymbol{\beta}) \subseteq \mathcal{S}_{\mathrm{E}(\mathbf{X}|Y)}$.

□

For the proof of Lemma 1, we need the following elementary lemma. We include its proof for completeness.

Lemma S1. *Let $\boldsymbol{\alpha}_k$ be the normalized kth eigenvector of $\boldsymbol{\Sigma}_{\mathbf{X}}^{-1/2}\mathbf{M}\boldsymbol{\Sigma}_{\mathbf{X}}^{-1/2}$. Then we must have $\boldsymbol{\beta}_k = \boldsymbol{\Sigma}_{\mathbf{X}}^{-1/2}\boldsymbol{\alpha}_k$.*

Proof of Lemma S1. Let $\boldsymbol{\alpha} = \boldsymbol{\Sigma}_{\mathbf{X}}^{1/2}\boldsymbol{\beta}$ in (4.8), we have that $\boldsymbol{\beta}_k = \boldsymbol{\Sigma}_{\mathbf{X}}^{-1/2}\boldsymbol{\alpha}_k$, where

$$\boldsymbol{\alpha}_k = \arg\max_{\boldsymbol{\alpha}} \boldsymbol{\alpha}^{\mathrm{T}}\boldsymbol{\Sigma}_{\mathbf{X}}^{-1/2}\mathbf{M}\boldsymbol{\Sigma}_{\mathbf{X}}^{-1/2}\boldsymbol{\alpha} \text{ s.t } \boldsymbol{\alpha}^{\mathrm{T}}\boldsymbol{\alpha} = 1, \boldsymbol{\alpha}^{\mathrm{T}}\boldsymbol{\alpha}_l = 0, l < k \tag{S1}$$

It is easy to see that $\boldsymbol{\alpha}_k$ is the kth eigenvector of $\boldsymbol{\Sigma}_{\mathbf{X}}^{-1/2}\mathbf{M}\boldsymbol{\Sigma}_{\mathbf{X}}^{-1/2}$ and the conclusion follows. □

Proof of Lemma 1. By Lemma S1, we have $\boldsymbol{\Sigma}_{\mathbf{X}}^{-1/2}\mathbf{M}\boldsymbol{\Sigma}_{\mathbf{X}}^{-1/2} = \sum_{j=1}^{p}\lambda_j\boldsymbol{\alpha}_j\boldsymbol{\alpha}_j^{\mathrm{T}}$. It follows that $\mathbf{M} = \boldsymbol{\Sigma}_{\mathbf{X}}\{\sum_{j=1}^{p}\lambda_j\boldsymbol{\beta}_j\boldsymbol{\beta}_j^{\mathrm{T}}\}\boldsymbol{\Sigma}_{\mathbf{X}}$. Hence, $\mathbf{M}_k = \boldsymbol{\Sigma}_{\mathbf{X}}\{\sum_{j=k}^{p}\lambda_j\boldsymbol{\beta}_j\boldsymbol{\beta}_j^{\mathrm{T}}\}\boldsymbol{\Sigma}_{\mathbf{X}}$ and $\boldsymbol{\beta}_k$ is its leading generalized eigenvector subject to $\boldsymbol{\beta}^{\mathrm{T}}\boldsymbol{\Sigma}_{\mathbf{X}}\boldsymbol{\beta} = 1$. □

S5 Proof for Theorem 1

Let $(\mathbf{V}_{k\cdot}, \mathbf{U}_{k\cdot})_{k=1}^{n}$ be iid sample from the joint distribution of $(\mathbf{V}, \mathbf{U})$, where $\mathbf{V}_{k\cdot} = (V_{k1}, \cdots, V_{kp})^{\mathrm{T}}$ is the kth sample. Write $\mathrm{MDDM}(\mathbf{V}|\mathbf{U}) = -\mathrm{E}\{(\mathbf{V} - \mathrm{E}(\mathbf{V}))(\mathbf{V}' - \mathrm{E}(\mathbf{V}'))^{\mathrm{T}}|\mathbf{U} - \mathbf{U}'|_q\} = -\mathbf{R} - \mathbf{S} + 2\mathbf{T}$, where

$$\begin{aligned}
\mathbf{R} &= \mathrm{E}(\mathbf{V}\mathbf{V}'^{\mathrm{T}}|\mathbf{U} - \mathbf{U}'|_q) \\
\mathbf{S} &= \mathrm{E}(\mathbf{V})\mathrm{E}(\mathbf{V}')^{\mathrm{T}}\mathrm{E}(|\mathbf{U} - \mathbf{U}'|_q) \\
\mathbf{T} &= \mathrm{E}[\mathbf{V}\mathbf{V}'^{\mathrm{T}}|\mathbf{U}' - \mathbf{U}''|_q] = \mathrm{E}[\mathrm{E}(\mathbf{V})\mathbf{V}'^{\mathrm{T}}|\mathbf{U} - \mathbf{U}'|_q]
\end{aligned}$$

with $(\mathbf{V}', \mathbf{U}')$ and $(\mathbf{V}'', \mathbf{U}'')$ being iid copies of $(\mathbf{V}, \mathbf{U})$. Note that $\mathbf{V}' = (V_1', \cdots, V_p')^T$. At the sample level, we have

$$\mathrm{MDDM}_n(\mathbf{V}|\mathbf{U}) = -\frac{1}{n^2}\sum_{k,l=1}^{n}(\mathbf{V}_{k\cdot} - \bar{\mathbf{V}}_n)(\mathbf{V}_{l\cdot} - \bar{\mathbf{V}}_n)^T|\mathbf{U}_{k\cdot} - \mathbf{U}_{l\cdot}|_q = -\mathbf{R}_n - \mathbf{S}_n + 2\mathbf{T}_n,$$

where $\bar{\mathbf{V}}_n = n^{-1}\sum_{k=1}^{n}\mathbf{V}_{k\cdot}$, and

$$\begin{aligned}
\mathbf{R}_n &= n^{-2}\sum_{k,l=1}^{n}\mathbf{V}_{k\cdot}\mathbf{V}_{l\cdot}^T|\mathbf{U}_{k\cdot} - \mathbf{U}_{l\cdot}|_q \\
\mathbf{S}_n &= n^{-2}\sum_{k,l=1}^{n}\mathbf{V}_{k\cdot}\mathbf{V}_{l\cdot}^T n^{-2}\sum_{k,l=1}^{n}|\mathbf{U}_{k\cdot} - \mathbf{U}_{l\cdot}|_q \\
\mathbf{T}_n &= n^{-3}\sum_{k,l,h=1}^{n}\mathbf{V}_{k\cdot}\mathbf{V}_{h\cdot}^T|\mathbf{U}_{h\cdot} - \mathbf{U}_{l\cdot}|_q.
\end{aligned}$$

Proposition S1. *Suppose Condition (C1) holds. There exists a positive integer $n_0 = n_0(\sigma_0, C_0, q) < \infty$, $\gamma = \gamma(\sigma_0, C_0, q) \in (0, 1/2)$ and a finite positive constant $D_0 = D_0(\sigma_0, C_0, q) < \infty$ such that when $n \geq n_0$ and $16 > \epsilon > D_0 n^{-\gamma}$, we have*

$$P(\|\mathbf{R}_n - \mathbf{R}\|_{max} > 4\epsilon) \leq 10p^2\exp\left(-\frac{\epsilon^2 n}{4\log^3 n}\right).$$

Proof of Proposition S1: Throughout the proof, C is a generic positive constant that vary from line to line. We shall find a bound for $P(\|\mathbf{R}_n - \mathbf{R}\|_{max} > 4\epsilon)$ first. For $i, j = 1, \cdots, p$, let $R_{ij} = \mathrm{E}[V_i V_j'|\mathbf{U} - \mathbf{U}'|_q]$ and $R_{n,ij} = n^{-2}\sum_{k,l=1}^{n} V_{ki}V_{lj}|\mathbf{U}_{k\cdot} - \mathbf{U}_{l\cdot}|_q$. Note that

$$P(\|\mathbf{R}_n - \mathbf{R}\|_{max} > 4\epsilon) \leq p^2 \max_{i,j=1,\cdots,p} P(|R_{n,ij} - R_{ij}| > 4\epsilon).$$

We shall focus on the case I, $(i, j) = (1, 2)$, since other cases can be treated in the same fashion and the bound is uniformly over all pair of (i, j)s.

Case I, $(i, j) = (1, 2)$, write $\widetilde{R}_{n,12} = \{n(n-1)\}^{-1}\sum_{k\neq l}^{n} V_{k1}V_{l2}|\mathbf{U}_{k\cdot} - \mathbf{U}_{l\cdot}|_q$. Let $\mathbf{W} = (\mathbf{U}^{\mathrm{T}}, V_1, V_2)^{\mathrm{T}}$ and $\mathbf{W}' = ((\mathbf{U}')^{\mathrm{T}}, V_1', V_2')^{\mathrm{T}}$ and $\mathbf{W}_k = (\mathbf{U}_{k\cdot}^{\mathrm{T}}, V_{k1}, V_{k2})^{\mathrm{T}}$. Define the kernel h_1 as

$$h_1(\mathbf{W}; \mathbf{W}') = \frac{V_1 V_2'|\mathbf{U} - \mathbf{U}'|_q + V_1' V_2|\mathbf{U} - \mathbf{U}'|_q}{2}$$

Then h_1 is symmetric, $\widetilde{R}_{n,12} = \{n(n-1)\}^{-1}\sum_{k\neq l}^{n} h_1(\mathbf{W}_k; \mathbf{W}_l)$ is a U-statistic of order two and $R_{n,12} = \frac{n-1}{n}\widetilde{R}_{n,12}$.

Under Condition (C1), there exists a positive constant $C_1 = C_1(\sigma_0, C_0) < \infty$ such that $|R_{12}| = |\mathrm{E}(V_1 V_2' |\mathbf{U} - \mathbf{U}'|_q)| \le \mathrm{E}^{1/2}(V_1^2)\mathrm{E}^{1/2}(V_2'^2)\mathrm{E}^{1/2}(|\mathbf{U} - \mathbf{U}'|_q^2) < C_1$. When ϵ satisfies $\epsilon \ge C_1/(2n)$, then $|R_{12}|/n \le 2\epsilon$ and

$$\begin{aligned} P(|R_{n,12} - R_{12}| \ge 4\epsilon) &= P\left(\left|\frac{n-1}{n}(\widetilde{R}_{n,12} - R_{12}) - \frac{1}{n}R_{12}\right| \ge 4\epsilon\right) \\ &\le P(|\widetilde{R}_{n,12} - R_{12}| + |R_{12}/n| \ge 4\epsilon) \le P(|\widetilde{R}_{n,12} - R_{12}| \ge 2\epsilon). \end{aligned}$$

Next we decompose

$$\begin{aligned} \widetilde{R}_{n,12} &= \{n(n-1)\}^{-1} \sum_{k \ne l}^{n} h_1(\mathbf{W}_k, \mathbf{W}_l)\mathbf{1}(|h_1(\mathbf{W}_k, \mathbf{W}_l)| \le M) \\ &\quad + \{n(n-1)\}^{-1} \sum_{k \ne l}^{n} h_1(\mathbf{W}_k, \mathbf{W}_l)\mathbf{1}(|h_1(\mathbf{W}_k, \mathbf{W}_l)| > M) \\ &= \widetilde{R}_{n,12,1} + \widetilde{R}_{n,12,2}, \end{aligned}$$

where the choice of M will be addressed at the end of proof. We also decompose its population counterpart $R_{12} = \mathrm{E}[h_1\mathbf{1}(|h_1| \le M)] + \mathrm{E}[h_1\mathbf{1}(|h_1| > M)] = R_{12,1} + R_{12,2}$.

By Lemma C on page 200 of Serfling (1980), we derive that for $m = \lfloor n/2 \rfloor$, and $t > 0$,

$$\mathrm{E}[\exp(t\widetilde{R}_{n,12,1})] \le \mathrm{E}^m[\exp(th_1\mathbf{1}(|h_1| \le M)/m)]$$

which entails that

$$\begin{aligned} P(\widetilde{R}_{n,12,1} - R_{12,1} \ge \epsilon) &\le \exp(-t(\epsilon + R_{12,1}))\mathrm{E}[\exp(t\widetilde{R}_{n,12,1})] \\ &\le \exp(-t\epsilon)\mathrm{E}^m\{\exp(t(h_1\mathbf{1}(|h_1| \le M) - R_{12,1})/m)\} \\ &\le \exp(-t\epsilon)\exp(t^2M^2/(2m)), \end{aligned}$$

where we have applied Markov's inequality and Lemma A(ii) [cf. Page 200 of Serfling (1980)] in the first and third inequality above, respectively. Applying the same argument with $h_1\mathbf{1}(|h_1| \le M)$ replaced by $-h_1\mathbf{1}(|-h_1| \le M)$, we can obtain

$$P(\widetilde{R}_{n,12,1} - R_{12,1} \le -\epsilon) \le \exp(-t\epsilon)\exp(t^2M^2/(2m)).$$

Choosing $t = \epsilon m/M^2$, we obtain that

$$P(|\widetilde{R}_{n,12,1} - R_{12,1}| \ge \epsilon) \le 2\exp(-\epsilon^2 m/(2M^2)) \tag{S2}$$

Next we turn to $\widetilde{R}_{n,12,2}$. First of all, by Cauchy-Schwartz inequality, $|R_{12,2}| \leq \mathrm{E}^{1/2}(h_1^2)P^{1/2}(|h_1| > M)$. Applying the inequality $|ab| \leq (a^2+b^2)/2$ and $(a+b)^2 \leq 2(a^2+b^2)$ for any $a, b \in R$, we derive

$$\begin{aligned} h_1(\mathbf{W}_k, \mathbf{W}_l) &\leq \frac{V_{k1}V_{l2}|\mathbf{U}_{k\cdot} - \mathbf{U}_{l\cdot}|_q + V_{k2}V_{l1}|\mathbf{U}_{k\cdot} - \mathbf{U}_{l\cdot}|_q}{2} \\ &\leq \frac{1}{4}\{(V_{k1}V_{l2} + V_{k2}V_{l1})^2 + |\mathbf{U}_{k\cdot} - \mathbf{U}_{l\cdot}|_q^2\} \\ &\leq \frac{1}{2}\{V_{k1}^2V_{l2}^2 + V_{k2}^2V_{l1}^2 + |\mathbf{U}_{k\cdot}|_q^2 + |\mathbf{U}_{l\cdot}|_q^2\} \\ &\leq \frac{1}{2}\{V_{k1}^4/2 + V_{l2}^4/2 + V_{k2}^4/2 + V_{l1}^4/2 + |\mathbf{U}_{k\cdot}|_q^2 + |\mathbf{U}_{l\cdot}|_q^2\} \end{aligned}$$

Then it is easy to show that under the uniform sub-Gaussian moment assumption in Condition (C1) and the upper bound on $h_1(\mathbf{W}_k, \mathbf{W}_l)$ above, we have that $\mathrm{E}^{1/2}(h_1^2) \leq C_2$ for some $C_2 = C_2(\sigma_0, C_0) < \infty$. Moreover, since $q^{1/2}\|\mathbf{U}\|_{\max} \geq \|\mathbf{U}\|_q$, we can derive that

$$\begin{aligned} & P(|h_1| > M) \\ \leq\ & P[\max\{|V_1|, |V_1'|, |V_2|, |V_2'|, 2q^{1/2}\|\mathbf{U}\|_{\max}, 2q^{1/2}\|\mathbf{U}'\|_{\max}\} \geq (\frac{M}{2})^{1/3}] \\ \leq\ & 2P\{|V_1| \geq (\frac{M}{2})^{1/3}\} + 2P\{|V_2| \geq (\frac{M}{2})^{1/3}\} + 2P\{2q^{1/2}\|\mathbf{U}\|_{\max} \geq (\frac{M}{2})^{1/3}\} \\ \leq\ & 2P\{|V_1| \geq (\frac{M}{2})^{1/3}\} + 2P\{|V_2| \geq (\frac{M}{2})^{1/3}\} + 2\sum_{j=1}^{q} P\{2q^{1/2}|U_j| \geq (\frac{M}{2})^{1/3}\} \quad \text{(S3)} \end{aligned}$$

Because V_1 is sub-Gaussian as assumed in Condition (C1), by Proposition 2.5.2 in Vershynin (2018), we have $P\{|V_1| \geq (\frac{M}{2})^{1/3}\} \leq 2\exp\{-C(\frac{M}{2})^{2/3}\}$ for some positive constant C. We apply similar arguments to all the remaining terms in (S3) and derive that

$$P(|h_1| > M) \leq (8+4q)\exp\{-2Cq^{-1}(\frac{M}{2})^{2/3}\}.$$

Thus $|R_{12,2}| \leq (8+4q)^{1/2}C_2\exp\{-Cq^{-1}(\frac{M}{2})^{2/3}\}$. If we choose $\epsilon > 0$ such that $(8+4q)^{1/2}C_2\exp\{-Cq^{-1}(\frac{M}{2})^{2/3}\} \leq \epsilon/2$, then $|R_{12,2}| \leq \epsilon/2$, which leads to $P(|\widetilde{R}_{n,12,2} - R_{12,2}| \geq$

$\epsilon) \leq P(|\widetilde{R}_{n,12,2}| \geq \epsilon/2)$. To bound $P(|\widetilde{R}_{n,12,2}| \geq \epsilon/2)$, we write

$$\begin{aligned}
|\widetilde{R}_{n,12,2}| &= |\{n(n-1)\}^{-1} \sum_{k\neq l}^{n} h_1(\mathbf{W}_k, \mathbf{W}_l)\mathbf{1}(|h_1(\mathbf{W}_k, \mathbf{W}_l)| > M)| \\
&\leq \{n(n-1)\}^{-1} \sum_{k\neq l}^{n} |h_1(\mathbf{W}_k, \mathbf{W}_l)|\mathbf{1}(\|\mathbf{W}_k\|_{\max} > q^{-1/6}(\frac{M}{2})^{1/3}) \\
&\quad + \{n(n-1)\}^{-1} \sum_{k\neq l}^{n} |h_1(\mathbf{W}_k, \mathbf{W}_l)|\mathbf{1}(\|\mathbf{W}_l\|_{\max} > q^{-1/6}(\frac{M}{2})^{1/3}) \\
&\equiv L_1 + L_2.
\end{aligned}$$

Without loss of generality, we only consider L_1. Define $F_k = \mathbf{1}(\|\mathbf{W}_k\|_{\max} > q^{-1/6}(\frac{M}{2})^{1/3})$. Note that

$$\begin{aligned}
L_1 &= \{n(n-1)\}^{-1} \sum_{k\neq l} |V_{k1}V_{l2}|\|\mathbf{U}_k - \mathbf{U}_l\|F_k \\
&\leq \{n(n-1)\}^{-1} \sum_{k\neq l} |V_{k1}V_{l2}|\|\mathbf{U}_k\|F_k + \{n(n-1)\}^{-1} \sum_{k\neq l} |V_{k1}V_{l2}|\|\mathbf{U}_l\|F_k \\
&\leq \{n(n-1)\}^{-1}(\sum_{k=1}^{n} |V_{k1}|\|\mathbf{U}_k\|F_k) \cdot (\sum_{l=1}^{n} |V_{l2}|) + \{n(n-1)\}^{-1}(\sum_{k=1}^{n} |V_{k1}|F_k) \cdot \sum_{l=1}^{n} |V_{l2}|\|\mathbf{U}_l\| \\
&\equiv L_{11} + L_{12}
\end{aligned}$$

For L_{11}, note that, for any $\lambda > 0$, $\mathrm{E}\exp\{\lambda|V_{l2}|^2\} = \mathrm{E}\exp\{\lambda V_{l2}^2\}$. Since V_{l2} is sub-Gaussian by Condition (C1), we have that $|V_{l2}|$ is also sub-Gaussian by Proposition 2.5.2 in Vershynin (2018). Hence, it follows from Bernstein's inequality [Theorem 2.8.1 in Vershynin (2018)] that for $\epsilon \in (0,1)$, $n \geq 2$,

$$P(|\frac{1}{n-1}\sum_{l=1}^{n}\{|V_{l2}| - \mathrm{E}|V_{l2}|\}| \geq \epsilon) \leq 2\exp(-Cn\epsilon^2) \tag{S4}$$

Regarding $\frac{1}{n}\sum_{k=1}^{n} |V_{k1}|\|\mathbf{U}_k\|F_k$, we note that $|V_{k1}|\|\mathbf{U}_k\|F_k \leq |V_{k1}|\|\mathbf{U}_k\| \cdot F_k \leq |V_{k1}| \cdot \sum_{j=1}^{q} |U_{kj}| \cdot F_k$. Since $|V_{k1}|$ and $|U_{kj}|$ are sub-Gaussian, we have that $|V_{k1}|\|\mathbf{U}_k\|F_k$ is sub-exponential (Lemma 2.7.7 in Vershynin (2018)). Again by Bernstein's inequality, we have that for $\epsilon \in (0,1)$,

$$P(|\frac{1}{n}\sum_{k=1}^{n}\{|V_{k1}|\|\mathbf{U}_k\|F_k - \mathrm{E}(|V_{k1}|\|\mathbf{U}_k\|F_k)\}| \geq \epsilon) \leq 2\exp(-Cn\epsilon^2). \tag{S5}$$

Moreover, it is easy to see that there exists a $C_3 = C_3(\sigma_0, C_0)$, such that $\mathrm{E}^{1/2}(|V_{k1}|\|\mathbf{U}_k\|)^2 \leq C_3$, so $\mathrm{E}(|V_{k1}|\|\mathbf{U}_k\|F_k) \leq \mathrm{E}^{1/2}(|V_{k1}|\|\mathbf{U}_k\|)^2\mathrm{E}^{1/2}F_k \leq C_3[2(2+q)]^{1/2}\exp\{-Cq^{-1/3}(\frac{M}{2})^{2/3}\}$

where we have used the fact that $\mathrm{E}(F_k) = P(\|\mathbf{W}_k\|_{\max} > q^{-1/6}(\frac{M}{2})^{1/3}) \leq 2(2+q)\exp\{-2Cq^{-1/3}(\frac{M}{2})^{2/3}\}$ by a union bound argument and uniform Sub-Gaussinity assumption in Condition (C1). Hence, $P(|L_{11}| \geq \epsilon/8) \leq \xi_1(\epsilon) + \xi_2(\epsilon)$, where $\xi_1(\epsilon) = P(\frac{1}{n}\sum_{k=1}^{n}|V_{k1}|\|\mathbf{U}_k\|F_k > (8E|V_2|+8)^{-1}\epsilon)$ and $\xi_2(\epsilon) = P((n-1)^{-1}\sum_{l=1}^{n}|V_{l2}| > E|V_2|+1)$. Choosing ϵ such that $\epsilon > (E|V_2|+1)C_3 16[2(2+q)]^{1/2}\exp\{-Cq^{-1/3}(\frac{M}{2})^{2/3}\}$ and $\epsilon < 16\mathrm{E}|V_2| + 16$, then it follows from (S5) that

$$\xi_1(\epsilon) \leq P(|\frac{1}{n}\sum_{k=1}^{n}\{|V_{k1}|\|\mathbf{U}_k\|F_k - \mathrm{E}(|V_{k1}|\|\mathbf{U}_k\|F_k)\}| \geq \epsilon(16E|V_2|+16)^{-1}) \leq 2\exp(-Cn\epsilon^2).$$

In addition, we can use (S4) to derive that $\xi_2(\epsilon) \leq 2\exp(-Cn)$. Combining these results, we have $P(|L_{11}| \geq \epsilon/8) \leq 2\exp(-Cn\epsilon^2)$, when $\epsilon \in ((E|V_2|+1)C_3 16[2(2+q)]^{1/2}\exp\{-Cq^{-1/3}(\frac{M}{2})^{2/3}\}, 16E|V_2|+16)$. Thus if we choose $M = (\log n)^{3/2}$, we can find a n_0, $\gamma \in (0, 1/2)$, and D_0 such that when $n \geq n_0$ and $16 > \epsilon > D_0 n^{-\gamma}$, we have $P(|L_{11}| > \epsilon/8) \leq 2\exp(-Cn\epsilon^2)$. Similar arguments lead to $P(|L_{12}| > \epsilon/8) \leq 2\exp(-Cn\epsilon^2)$. This implies that $P(|L_1| > \epsilon/4) \leq 4\exp(-Cn\epsilon^2)$ and similarly $P(|L_2| > \epsilon/4) \leq 4\exp(-Cn\epsilon^2)$. Therefore $P(|\widetilde{R}_{n,12,2}| \geq \epsilon/2) \leq 8\exp(-Cn\epsilon^2)$. In view of (S2), the desired statement follows by choosing large enough n_0 such that $4\log^3(n_0) > C$.

Proof of Theorem 1: Notice that for any $\epsilon > 0$,

$$\begin{aligned}
&P(\|\mathrm{MDDM}_n(\mathbf{V}|\mathbf{U}) - \mathrm{MDDM}(\mathbf{V}|\mathbf{U})\|_{max} > 12\epsilon) \\
&\leq P(\|\mathbf{R}_n - \mathbf{R}\|_{max} > 4\epsilon) + P(\|\mathbf{S}_n - \mathbf{S}\|_{max} > 4\epsilon) + P(\|\mathbf{T}_n - \mathbf{T}\|_{max} > 4\epsilon)
\end{aligned}$$

The concentration bound for $\|\mathbf{R}_n - \mathbf{R}\|_{max}$ has been obtained in Proposition S1, and we shall address the concentration of $\|\mathbf{T}_n - \mathbf{T}\|_{max}$ in the proof below. The proof for the concentration of $\|\mathbf{S}_n - \mathbf{S}\|_{max}$ is similar and simpler so is omitted. Note that

$$P(\|\mathbf{T}_n - \mathbf{T}\|_{max} > 4\epsilon) \leq p^2 \max_{i,j=1,\cdots,p} P(|T_{n,ij} - T_{ij}| > 4\epsilon).$$

Following the same argument as used in the beginning of proof of Proposition S1, we shall only focus on the case $(i,j) = (1,2)$ as other cases can be treated in exactly the same manner.

Let $T_{12} = \mathrm{E}[V_1 V_2'|\mathbf{U}' - \mathbf{U}''|_q] = \mathrm{E}[\mathrm{E}(V_1)V_2'|\mathbf{U} - \mathbf{U}'|_q]$ and $T_{n,12} = n^{-3}\sum_{k,l,h=1}^{n} V_{k1}V_{h2}|\mathbf{U}_{h\cdot} - \mathbf{U}_{l\cdot}|_q$. Let

$$\begin{aligned}
\widetilde{T}_{n,12} \;=\; & \frac{1}{n(n-1)(n-2)}\sum_{k<l<h}[V_{k1}V_{h2}|\mathbf{U}_{h\cdot} - \mathbf{U}_{l\cdot}|_q + V_{k1}V_{l2}|\mathbf{U}_{l\cdot} - \mathbf{U}_{h\cdot}| \\
& + V_{l1}V_{k2}|\mathbf{U}_{k\cdot} - \mathbf{U}_{h\cdot}|_q + V_{l1}V_{h2}|\mathbf{U}_{h\cdot} - \mathbf{U}_{k\cdot}|_q + V_{h1}V_{l2}|\mathbf{U}_{l\cdot} - \mathbf{U}_{k\cdot}|_q \\
& + V_{h1}V_{k2}|\mathbf{U}_{k\cdot} - \mathbf{U}_{l\cdot}|_q] = 6\{n(n-1)(n-2)\}^{-1}\sum_{k<l<h} h_3(\mathbf{W}_k, \mathbf{W}_l, \mathbf{W}_h),
\end{aligned}$$

where h_3 is a kernel function for U-statistic of order three. Following the same argument to deal with $\widetilde{R}_{n,12}$ in the proof of Proposition S1, we write $\widetilde{T}_{n,12} = \widetilde{T}_{n,12,1} + \widetilde{T}_{n,12,2}$, where

$$\begin{aligned}\widetilde{T}_{n,12,1} &= 6\{n(n-1)(n-2)\}^{-1} \sum_{k<l<h} h_3 \mathbf{1}(|h_3| \le M), \\ \widetilde{T}_{n,12,2} &= 6\{n(n-1)(n-2)\}^{-1} \sum_{k<l<h} h_3 \mathbf{1}(|h_3| > M).\end{aligned}$$

Correspondingly, we define $T_{12} = T_{12,1} + T_{12,2}$, where $T_{12,1} = \mathrm{E}[h_3 \mathbf{1}(|h_3| \le M)]$ and $T_{12,2} = \mathrm{E}[h_3 \mathbf{1}(|h_3| > M)]$. By using the same argument for $\widetilde{R}_{n,12,1}$, we can show that

$$P(|\widetilde{T}_{n,12,1} - T_{12,1}| \ge \epsilon) \le 2\exp(-\epsilon^2 \lfloor n/3 \rfloor / (2M^2))$$

since $\widetilde{T}_{n,12,1}$ is a third order U-statistic. Also we note that by the same argument used in bounding h_1, we can get

$$|h_3| \le \frac{1}{12}\{V_{k1}^4 + V_{l1}^4 + V_{h1}^4 + V_{k2}^4 + V_{l2}^4 + V_{h2}^4 + 8|\mathbf{U}_{h\cdot}|_q^2 + 8|\mathbf{U}_{k\cdot}|_q^2 + 8|\mathbf{U}_{l\cdot}|_q^2\}.$$

It follows from Cauchy-Schwartz inequality that $|T_{12,2}| \le E^{1/2}(h_3^2) P^{1/2}(|h_3| > M)$. By using exactly the same argument used for (S3), we can show that $P(|h_3| > M) \le C\exp(-C_3 M^{2/3})$ for some $C_3 = C_3(\sigma_0, C_0, q) > 0$. Hence $|T_{12,2}| \le C\exp(-C_3 M^{2/3})$. We can choose $\epsilon > 0$ such that $C\exp(-C_3 M^{2/3}) \le \epsilon/2$. Therefore, for $\epsilon \ge 2C\exp(-C_3 M^{2/3})$, we have $P(|\widetilde{T}_{n,12,2} - T_{12,2}| > \epsilon) \le P(|\widetilde{T}_{n,12,2}| \ge \epsilon/2)$. By setting $M = \log^{3/2}(n)$ and adopting the same argument as used in bounding $P(|\widetilde{R}_{n,12,2}| \ge \epsilon/2)$, we can derive that $P(|\widetilde{T}_{n,12,2}| \ge \epsilon/2) \le 12\exp(-C_4 n\epsilon^2)$ when $n \ge n_1$ and $\epsilon \in (D_1 n^{-\gamma_1}, 16)$ for some $C_4 = C_4(\sigma_0, C_0, q) > 0$, $D_1 = D_1(\sigma_0, C_0, q)$, $n_1 = n_1(\sigma_0, C_0, q)$ and $\gamma_1 = \gamma_1(\sigma_0, C_0, q)$.

Combining the above results, we obtain that for $16 > \epsilon > D_1 n^{-\gamma_1}$ and $n \ge n_1$, we have

$$\begin{aligned}P(|\widetilde{T}_{n,12} - T_{12}| \ge 2\epsilon) &\le 12\exp(-C_4 n\epsilon^2) + 2\exp(-\epsilon^2 \lfloor n/3 \rfloor/(2M^2)) \\ &\le 14\exp(-\epsilon^2 n/(6\log^3(n))),\end{aligned}$$

where we choose n_1 such that $6\log^3(n_1) > C_4^{-1}$.

Further we note that

$$T_{n,12} - T_{12} = \frac{(n-1)(n-2)}{n^2}(\widetilde{T}_{n,12} - T_{12}) - \frac{3n-2}{n^2} T_{12} + \frac{n-1}{n^2}(\widetilde{R}_{n,12} - R_{12}) + \frac{n-1}{n^2} R_{12}.$$

There exists a finite positive constant $C_6 = C_6(\sigma_0, C_0, q)$ such that $|R_{12}| \leq C_6$ and $|T_{12}| \leq C_6$ so if we choose $\epsilon \geq 3C_6/n$, then $|\frac{3n-2}{n^2}T_{12}| \leq \epsilon$ and $|\frac{n-1}{n^2}R_{12}| \leq \epsilon/3$. Then for $n \geq n_1$ and $\epsilon > D_1 n^{-\gamma_1}$,

$$\begin{aligned} P(|T_{n,12} - T_{12}| > 4\epsilon) &\leq P(|\widetilde{T}_{n,12} - T_{12}| > 2\epsilon) + P(|\widetilde{R}_{n,12} - R_{12}| > 2\epsilon/3) \\ &\leq 14\exp(-\epsilon^2 n/(6\log^3(n))) + 10\exp\left\{-\frac{\epsilon^2 n}{36\log^3(n)}\right\} \\ &\leq 24\exp\left\{-\frac{\epsilon^2 n}{36\log^3(n)}\right\} \end{aligned}$$

Thus the conclusion follows from the above inequality and Proposition S1.

□

S6 Proofs for Theorems 2 & 3

S6.1 Two Generic Algorithms and Their Properties

We first describe two generic algorithms and their properties that will help our proof for Theorems 2 & 3. Consider two matrices $\mathbf{A}, \mathbf{B} \in \mathbb{R}^{p\times p}$, their estimates $\widehat{\mathbf{A}}, \widehat{\mathbf{B}} \in \mathbb{R}^{p\times p}$ and vectors $\mathbf{v}, \mathbf{v}_0, \mathbf{v}_t \in \mathbb{R}^p$. We have Algorithm 3 for the penalized eigen-decomposition for $\mathbf{A}$, and Algorithm 4 for the penalized generalized eigen-decomposition for $(\mathbf{A}, \mathbf{B})$. Algorithm 3 is originally proposed in Yuan & Zhang (2013), and in our Algorithm 1 we use it repeatedly for K times to perform penalized eigen-decomposition for MDDM. Algorithm 4 is originally proposed as the RIFLE Algorithm in Tan et al. (2018), and we use it for K times to perform penalized generalized eigen-decomposition for MDDM.

Yuan & Zhang (2013) proved a property for Algorithm 3 that is important for our proof. Assume that $\mathbf{A}$ has a unique leading eigenvector $\mathbf{v}_\mathbf{I}^*$ with $\|\mathbf{v}_\mathbf{I}^*\|_0 \leq d$. Denote $\lambda_1^\mathbf{I}, \ldots, \lambda_p^\mathbf{I}$ as the eigenvalues. We assume that there exists a constant $\Delta_\mathbf{I} = \lambda_1^\mathbf{I} - \max_{j>1}\lambda_j^\mathbf{I}$. Also for any positive integer k', define

$$\rho(\mathbf{E}_\mathbf{A}, k') = \sup_{\|\mathbf{u}\|_2=1, \|\mathbf{u}\|_0 \leq k'} |\mathbf{u}^\mathrm{T}\mathbf{E}_\mathbf{A}\mathbf{u}|,$$

where $\mathbf{E}_\mathbf{A} = \widehat{\mathbf{A}} - \mathbf{A}$, with $\widehat{\mathbf{A}}$ being an estimate of $\mathbf{A}$. We have the following proposition.

Proposition S2 ((Yuan & Zhang 2013) c.f Theorem 4)**.** *In Algorithm 3, let $s = d + 2s'$ with $s' \geq d$. Assume that $\rho(\mathbf{E}_\mathbf{A}, s) \leq \Delta_\mathbf{I}$. Define*

$$\gamma(s) = \frac{\lambda_1^\mathbf{I} - \Delta_\mathbf{I} + \rho(\mathbf{E}_\mathbf{A}, s)}{\lambda_1^\mathbf{I} - \rho(\mathbf{E}_\mathbf{A}, s)} < 1, \quad \delta_\mathbf{I}(s) = \frac{\sqrt{2}\rho(\mathbf{E}_\mathbf{A}, s)}{\sqrt{\rho(\mathbf{E}_\mathbf{A}, s)^2 + (\Delta_\mathbf{I} - 2\rho(\mathbf{E}_\mathbf{A}, s))^2}}.$$

Algorithm 3 A generic penalized eigen-decomposition algorithm.

1. Input: $s, \widehat{\mathbf{A}}$.
2. Initialize $\mathbf{v}_0$.
   (a) Iterate over t until convergence:
   (b) Set $\mathbf{v}_t = \widehat{\mathbf{A}}\mathbf{v}_{t-1}$.
   (c) If $\|\mathbf{v}_t\|_0 \leq s$, set
   $$\mathbf{v}_t = \frac{\mathbf{v}_t}{\|\mathbf{v}_t\|_2};$$
   else
   $$\mathbf{v}_t = \frac{\mathrm{HT}(\mathbf{v}_t, s)}{\|\mathrm{HT}(\mathbf{v}_t, s)\|_2}$$
3. Output $\mathbf{v}_\infty$ at convergence.

If $|\mathbf{v}_0^{\mathrm{T}}\mathbf{v}_{\mathbf{I}}^*| \geq \theta + \delta_{\mathbf{I}}(s)$ *for some* $\|\mathbf{v}_0\|_0 \leq s'$, $\|\mathbf{v}_0\| = 1$, *and* $\theta \in (0,1)$ *such that*

$$\mu = \sqrt{[1 + 2\{(\frac{d}{s'})^{1/2} + \frac{d}{s'}\}]\{1 - 0.5\theta(1+\theta)(1-\gamma(s)^2)\}} < 1,$$

then we either have

$$\sqrt{1 - |\mathbf{v}_0^{\mathrm{T}}\mathbf{v}_{\mathbf{I}}^*|} < \sqrt{10}\delta_{\mathbf{I}}(s)/(1-\mu),$$

or for all $t \geq 0$,

$$\sqrt{1 - |\mathbf{v}_t^{\mathrm{T}}\mathbf{v}_{\mathbf{I}}^*|} \leq \mu^t \sqrt{1 - |\mathbf{v}_0^{\mathrm{T}}\mathbf{v}_{\mathbf{I}}^*|} + \sqrt{10}\delta_{\mathbf{I}}(s)/(1-\mu).$$

Based on the results in Tan et al. (2018), we can also derive the following useful results for Algorithm 4. We assume that the matrix pair $(\mathbf{A}, \mathbf{B})$ has the leading generalized eigenvector $\mathbf{v}^*$ such that $\|\mathbf{v}^*\|_0 \leq d$. The generalized eigenvalues of $(\mathbf{A}, \mathbf{B})$ are referred to as $\lambda_j, j = 1, \ldots, p$ and their estimates are $\widehat{\lambda}_j, j = 1, \ldots, p$. We introduce the following notation:

$$\mathrm{cr}(\mathbf{A}, \mathbf{B}) = \min_{\mathbf{v}:\|\mathbf{v}\|_2=1} \{(\mathbf{v}^{\mathrm{T}}\mathbf{A}\mathbf{v})^2 + (\mathbf{v}^{\mathrm{T}}\mathbf{B}\mathbf{v})^2\}^{1/2} > 0 \tag{S6}$$

$$\mathrm{cr}(k') = \inf_{F:|F|\leq k'} \mathrm{cr}(\mathbf{A}_F, \mathbf{B}_F), \tag{S7}$$

$$\delta(k') = \sqrt{\rho(\mathbf{E}_{\mathbf{A}}, k')^2 + \rho(\mathbf{E}_{\mathbf{B}}, k')^2}, \tag{S8}$$

where $\mathbf{E}_{\mathbf{B}} = \mathbf{B} - \widehat{\mathbf{B}}$, with $\widehat{\mathbf{B}}$ being an estimate of $\mathbf{B}$. Also denote $\kappa(\mathbf{B})$ as the condition number of $\mathbf{B}$ and $\omega_1(F) = \sup_{\|\mathbf{u}\|_0 \subset F, \|\mathbf{u}\|_2=1} \frac{\mathbf{u}^{\mathrm{T}}\mathbf{A}\mathbf{u}}{\mathbf{u}^{\mathrm{T}}\mathbf{B}\mathbf{u}}$ for any index set F. We consider the following assumption:

Algorithm 4 A generic penalized generalized eigen-decomposition algorithm.

1. Input: $s, \widehat{\mathbf{A}}, \widehat{\mathbf{B}}$ and step size $\eta > 0$.
2. Initialize $\mathbf{v}_0$.
3. Iterate over t until convergence:
   (a) Set $\rho^{(t-1)} = \dfrac{\mathbf{v}_t^{\mathrm{T}} \widehat{\mathbf{A}} \mathbf{v}_t}{\mathbf{v}_t^{\mathrm{T}} \widehat{\mathbf{B}} \mathbf{v}_t}$.
   (b) $\mathbf{C} = \mathbf{I} + (\eta/\rho^{(t-1)}) \cdot (\widehat{\mathbf{A}} - \rho^{(t-1)} \widehat{\mathbf{B}})$.
   (c) $\mathbf{v}_t = \mathbf{C}\mathbf{v}_{t-1} / \|\mathbf{C}\mathbf{v}_{t-1}\|_2$.
   (d) $\mathbf{v}_t = \dfrac{\mathrm{HT}(\mathbf{v}_t, s)}{\|\mathrm{HT}(\mathbf{v}_t, s)\|_2}$.
4. Output $\mathbf{v}_\infty$ at convergence.

Assumption S1. *For sufficiently large n, there are constants $b, c > 0$, such that $\dfrac{\delta(s)}{cr(s)} \leq b$ and $\rho(\mathbf{E}_{\mathbf{B}}, s) \leq c\lambda_{\min}(\mathbf{B})$ for any $s = o(n)$.*

We also denote $c_{\text{upper}} = \dfrac{1+c}{1-c}$ for c defined in Assumption S1. We estimate $\mathbf{v}^*$ with the RIFLE algorithm with the step size η. We choose η such that $\eta\lambda_{\max}(\mathbf{B}) < 1/(1+c)$. Further, in the RIFLE algorithm, let $s = 2s' + d$ and choose $s' = Cd$ for sufficiently large C. The initial value $\mathbf{v}_0$ satisfies that $\|\mathbf{v}_0\|_2 = 1$.

Proposition S3 (Based on Theorem 1 and Corollary 1 in Tan et al. (2018))**.** *Under Assumption S1, we have the following conclusions:*

1. *For any F such that $\mathrm{supp}(\mathbf{v}_0) \subset F$, there exists a constant a such that*

$$(1-a)\omega_1(F) \leq \hat{\omega}_1(F) \leq (1+a)\omega_1(F). \tag{S9}$$

2. *Choose η such that*

$$\nu = \sqrt{1 + 2\{(d/s')^{1/2} + d/s'\}} \sqrt{1 - \frac{1+c}{8}\eta\lambda_{\min}(\mathbf{B}) \frac{1-\alpha}{c_{upper}\kappa(\mathbf{B}) + \alpha}} < 1, \tag{S10}$$

where $\alpha = \frac{(1+a)\lambda_2}{(1-a)\lambda_1}$. Input an initial vector $\mathbf{v}_0$ such that $\dfrac{|(\mathbf{v}^*)^{\mathrm{T}}\mathbf{v}_0|}{\|\mathbf{v}^*\|_2} \geq 1 - \theta(\mathbf{A}, \mathbf{B})$, where

$$\theta(\mathbf{A}, \mathbf{B}) = \min\left[\frac{1}{8c_{upper}\kappa(\mathbf{B})}, \frac{1/\alpha - 1}{3c_{upper}\kappa(\mathbf{B})}, \frac{1-\alpha}{30(1+c)c_{upper}^2\eta\lambda_{\max}(\mathbf{B})\kappa^2(\mathbf{B})\{c_{upper}\kappa(\mathbf{B}) + \alpha\}}\right]. \tag{S11}$$

Further denote

$$\xi = \min_{j>1} \frac{\lambda_1 - (1+a)\lambda_j}{\sqrt{1+\lambda_1^2}\sqrt{1+(1-a)^2\lambda_j^2}}. \tag{S12}$$

Assume that $\xi > \delta(s)/cr(s)$ *and we have*

$$\sqrt{1 - \frac{|(\mathbf{v}^*)^{\mathrm{T}}\mathbf{v}_t|}{\|\mathbf{v}^*\|_2}} \le \nu^t \sqrt{\theta(\mathbf{A},\mathbf{B})} + \frac{\sqrt{10}}{1-\nu}\frac{2}{\xi\{cr(s)-\delta(s)\}}\delta(s). \tag{S13}$$

We rewrite Proposition S3 in the following more user-friendly form. We denote

$$\phi = \lambda_1 - \lambda_2, \quad a^* = \min\{1/2, \frac{\Delta}{\lambda_1+\lambda_2}, \frac{\lambda_{\min}(\mathbf{B})}{2}\} \tag{S14}$$

$$\xi^* = \frac{\lambda_1-\lambda_2}{2(1+\lambda_1^2)}, \quad \alpha^* = \frac{(1+a^*)\lambda_2}{(1-a^*)\lambda_1}. \tag{S15}$$

We have the following lemma.

Lemma S2. *Assume that Assumption S1 holds. Choose* η *such that*

$$\nu^* = \sqrt{1+2\{(d/s')^{1/2}+d/s'\}}\sqrt{1 - \frac{1+c}{8}\eta\lambda_{\min}(\mathbf{B})\frac{1-\frac{\lambda_2}{\lambda_1}}{c_{upper}\kappa(\mathbf{B})+3\frac{\lambda_2}{\lambda_1}}} < 1.$$

Also assume that $\delta(s) < \min\{1/2, \lambda_{\min}(\mathbf{B})/2\}$, $\frac{2\delta(s)}{\lambda_{\min}(\mathbf{B})} + \frac{2\delta(s)}{\lambda_{\min}(\mathbf{B})\lambda_1(F)} < a^*$ *and* $\xi^* > 2\delta(s)/cr(s)$. *Input an initial vector* $\mathbf{v}_0$ *such that* $\frac{|(\mathbf{v}^*)^{\mathrm{T}}\mathbf{v}_0|}{\|\mathbf{v}^*\|_2} \ge 1-\theta^*(\mathbf{A},\mathbf{B})$, *where* $0 < \theta^*(\mathbf{A},\mathbf{B}) < 1$,

$$\theta^*(\mathbf{A},\mathbf{B}) = \min\left[\frac{1}{8c_{upper}\kappa(\mathbf{B})}, \frac{1/\alpha^*-1}{3c_{upper}\kappa(\mathbf{B})}, \frac{1-\alpha^*}{30(1+c)c_{upper}^2\eta\lambda_{\max}(\mathbf{B})\kappa^2(\mathbf{B})\{c_{upper}\kappa(\mathbf{B})+\alpha^*\}}\right]. \tag{S16}$$

We have

$$|\sin\Theta(\mathbf{v}^*, \mathbf{v}_\infty)| \le \frac{\sqrt{20}}{1-\nu^*}\frac{2}{\xi^*\{cr(s)-\delta(s)\}}\delta(s). \tag{S17}$$

Proof of Lemma S2. We prove Lemma S2 by showing that all the conditions in Proposition S3 are met. Note that

$$\hat{\omega}_1(F) = \sup_{\|\mathbf{u}\|_2=1,\mathrm{supp}(\mathbf{u})\subset F} \frac{\mathbf{u}^{\mathrm{T}}\widehat{\mathbf{A}}\mathbf{u}}{\mathbf{u}^{\mathrm{T}}\widehat{\mathbf{B}}\mathbf{u}} = \sup_{\|\mathbf{u}\|_2=1,\mathrm{supp}(\mathbf{u})\subset F} \frac{\mathbf{u}^{\mathrm{T}}\mathbf{A}\mathbf{u} + \mathbf{u}^{\mathrm{T}}(\widehat{\mathbf{A}}-\mathbf{A})\mathbf{u}}{\mathbf{u}^{\mathrm{T}}\mathbf{B}\mathbf{u} + \mathbf{u}^{\mathrm{T}}(\widehat{\mathbf{B}}-\mathbf{B})\mathbf{u}}. \tag{S18}$$

It is obvious that

$$\sup_{\|\mathbf{u}\|_2=1,\mathrm{supp}(\mathbf{u})\subset F} \frac{\mathbf{u}^{\mathrm{T}}\mathbf{A}\mathbf{u}-\delta(s)}{\mathbf{u}^{\mathrm{T}}\mathbf{B}\mathbf{u}+\delta(s)} \le \hat{\omega}_1(F) \le \sup_{\|\mathbf{u}\|_2=1,\mathrm{supp}(\mathbf{u})\subset F} \frac{\mathbf{u}^{\mathrm{T}}\mathbf{A}\mathbf{u}+\delta(s)}{\mathbf{u}^{\mathrm{T}}\mathbf{B}\mathbf{u}-\delta(s)} \tag{S19}$$

Because $a^* > \dfrac{2\delta(s)}{\lambda_{\min}(\mathbf{B})} + \dfrac{2\delta(s)}{\lambda_{\min}(\mathbf{B})\lambda_1(F)}$, by Lemma S3, we have that Assumption S1 implies

$$(1-a^*)\omega_1(F) \le \hat{\omega}_1(F) \le (1+a^*)\omega_1(F). \tag{S20}$$

Also, by our definition, $a^* \le 1/2$. It follows that $\frac{\lambda_2}{\lambda_1} \le \frac{(1+a)\lambda_2}{(1-a)\lambda_1} = \alpha \le \frac{3\lambda_2}{\lambda_1}$. Hence, $\nu \le \nu^* < 1$, where ν is defined in (S10). In addition, because $a^* \le \frac{\phi}{\lambda_1+\lambda_2} \le \frac{\phi}{2\lambda_2}$, we have $\xi \ge \xi^*$, where ξ is defined in (S12). Finally, because $\dfrac{1}{8c_{\text{upper}}\kappa(\mathbf{B})} < 1$, we have $\theta^*(\mathbf{A},\mathbf{B}) < 1$. Because $a^* \le \dfrac{\phi}{\lambda_1+\lambda_2}$, we have $1-\gamma^* > 0$ and thus $\theta^*(\mathbf{A},\mathbf{B}) > 0$.

By Proposition S3, we have

$$\begin{aligned}
\sqrt{1-\frac{|(\mathbf{v}^*)^{\mathrm{T}}\mathbf{v}_t|}{\|\mathbf{v}^*\|_2}} &\le \nu^t\sqrt{\theta(\mathbf{A},\mathbf{B})} + \frac{\sqrt{10}}{1-\nu}\frac{2}{\xi\{cr(s)-\delta(s)\}}\delta(s) &\text{(S21)}\\
&\le (\nu^*)^t\sqrt{\theta^*(\mathbf{A},\mathbf{B})} + \frac{\sqrt{10}}{1-\nu^*}\frac{2}{\xi^*\{cr(s)-\delta(s)\}}\delta(s). &\text{(S22)}
\end{aligned}$$

Let $t \to \infty$ and we have that $\sqrt{1-\dfrac{|(\mathbf{v}^*)^{\mathrm{T}}\mathbf{v}_\infty|}{\|\mathbf{v}^*\|_2}} \le \dfrac{\sqrt{10}}{1-\nu^*}\dfrac{2}{\xi^*\{cr(s)-\delta(s)\}}\delta(s)$. Finally, we note that $1-\dfrac{|(\mathbf{v}^*)^{\mathrm{T}}\mathbf{v}_\infty|}{\|\mathbf{v}^*\|_2} = 1-|\cos\Theta(\mathbf{v}^*,\mathbf{v}_\infty)|$. Since $\sin^2\Theta(\mathbf{v}^*,\mathbf{v}_\infty) = (1+|\cos\Theta(\mathbf{v}^*,\mathbf{v}_\infty)|)(1-|\cos\Theta(\mathbf{v}^*,\mathbf{v}_\infty)|) \le 2(1-|\cos\Theta(\mathbf{v}^*,\mathbf{v}_\infty)|)$, we have the desired conclusion. □

Lemma S3. *Consider two symmetric matrices $\mathbf{A},\mathbf{B}$, where $\lambda_{\min}(\mathbf{B}) > 0$. For any $F \subset \{1,\dots,p\}$, denote*

$$\lambda_1(F) = \sup_{\|\mathbf{u}\|_2=1,\mathrm{supp}(\mathbf{u})\subset F} \frac{\mathbf{u}^{\mathrm{T}}\mathbf{A}\mathbf{u}}{\mathbf{u}^{\mathrm{T}}\mathbf{B}\mathbf{u}}. \tag{S23}$$

For any $0 < \epsilon < \min\{\frac{1}{2}, \dfrac{\lambda_{\min}(\mathbf{B})}{2}\}$, we have that

$$\begin{aligned}
\sup_{\|\mathbf{u}\|_2=1,\mathrm{supp}(\mathbf{u})\subset F} \frac{\mathbf{u}^{\mathrm{T}}\mathbf{A}\mathbf{u}+\epsilon}{\mathbf{u}^{\mathrm{T}}\mathbf{B}\mathbf{u}-\epsilon} &\le (1+a)\lambda_1(F), &\text{(S24)}\\
\sup_{\|\mathbf{u}\|_2=1,\mathrm{supp}(\mathbf{u})\subset F} \frac{\mathbf{u}^{\mathrm{T}}\mathbf{A}\mathbf{u}-\epsilon}{\mathbf{u}^{\mathrm{T}}\mathbf{B}\mathbf{u}+\epsilon} &\ge (1-a)\lambda_1(F), &\text{(S25)}
\end{aligned}$$

where $a = \dfrac{2\epsilon}{\lambda_{\min}(\mathbf{B})} + \dfrac{2\epsilon}{\lambda_{\min}(\mathbf{B})\omega_1(F)}$.

Proof of Lemma S3. For (S24), note that, for any $\mathbf{u}$, we have $\mathbf{u}^{\mathrm{T}}\mathbf{B}\mathbf{u} - \epsilon \geq \{1 - \frac{\epsilon}{\lambda_{\min}(\mathbf{B})}\}\mathbf{u}^{\mathrm{T}}\mathbf{B}\mathbf{u}$. So

$$\sup_{\|\mathbf{u}\|_2=1,\mathrm{supp}(\mathbf{u})\subset F} \frac{\mathbf{u}^{\mathrm{T}}\mathbf{A}\mathbf{u} + \epsilon}{\mathbf{u}^{\mathrm{T}}\mathbf{B}\mathbf{u} - \epsilon} \tag{S26}$$

$$\leq \frac{1}{1 - \frac{\epsilon}{\lambda_{\min}(\mathbf{B})}} \cdot \sup_{\|\mathbf{u}\|_2=1,\mathrm{supp}(\mathbf{u})\subset F} \frac{\mathbf{u}^{\mathrm{T}}\mathbf{A}\mathbf{u}}{\mathbf{u}^{\mathrm{T}}\mathbf{B}\mathbf{u}} + \frac{1}{1 - \frac{\epsilon}{\lambda_{\min}(\mathbf{B})}} \cdot \sup_{\|\mathbf{u}\|_2=1,\mathrm{supp}(\mathbf{u})\subset F} \frac{1}{\mathbf{u}^{\mathrm{T}}\mathbf{B}\mathbf{u}} \tag{S27}$$

$$\leq \frac{1}{1 - \frac{\epsilon}{\lambda_{\min}(\mathbf{B})}} \cdot \omega_1(F) + \frac{2\epsilon}{\lambda_{\min}(\mathbf{B})\omega_1(F)} \cdot \omega_1(F), \tag{S28}$$

where in the last inequality we use the fact that $\epsilon < \dfrac{\lambda_{\min}(\mathbf{B})}{2}$. Also note that, for any $0 < x < 1/2$, we have $\dfrac{1}{1-x} < 1 + 2x$. It follows that $\dfrac{1}{1 - \frac{\epsilon}{\lambda_{\min}(\mathbf{B})}} \leq 1 + \dfrac{2\epsilon}{\lambda_{\min}(\mathbf{B})}$. Hence,

$$\sup_{\|\mathbf{u}\|_2=1,\mathrm{supp}(\mathbf{u})\subset F} \frac{\mathbf{u}^{\mathrm{T}}\mathbf{A}\mathbf{u} + \epsilon}{\mathbf{u}^{\mathrm{T}}\mathbf{B}\mathbf{u} - \epsilon} \leq \left\{1 + \frac{2\epsilon}{\lambda_{\min}(\mathbf{B})} + \frac{2\epsilon}{\lambda_{\min}(\mathbf{B})\omega_1(F)}\right\} \omega_1(F) = (1 + a)\omega_1(F). \tag{S29}$$

Similarly, for (S25), we note that, for any $\mathbf{u}$, $\mathbf{u}^{\mathrm{T}}\mathbf{B}\mathbf{u} + \epsilon \leq \{1 + \frac{\epsilon}{\lambda_{\min}(\mathbf{B})}\}\mathbf{u}^{\mathrm{T}}\mathbf{B}\mathbf{u}$. So

$$\sup_{\|\mathbf{u}\|_2=1,\mathrm{supp}(\mathbf{u})\subset F} \frac{\mathbf{u}^{\mathrm{T}}\mathbf{A}\mathbf{u} - \epsilon}{\mathbf{u}^{\mathrm{T}}\mathbf{B}\mathbf{u} + \epsilon} \geq \frac{1}{1 + \frac{\epsilon}{\lambda_{\min}(\mathbf{B})}} \cdot \omega_1(F) - \frac{1}{1 + \frac{\epsilon}{\lambda_{\min}(\mathbf{B})}} \cdot \frac{\epsilon}{\lambda_{\max}(\mathbf{B})}, \tag{S30}$$

$$\geq \frac{1}{1 + \frac{\epsilon}{\lambda_{\min}(\mathbf{B})}} \cdot \omega_1(F) - \frac{\epsilon}{\lambda_{\max}(\mathbf{B})\omega_1(F)} \cdot \omega_1(F). \tag{S31}$$

Because $\frac{1}{1+x} > 1 - x$ for any $x > 0$, we have $\dfrac{1}{1 + \frac{\epsilon}{\lambda_{\min}(\mathbf{B})}} > 1 - \dfrac{\epsilon}{\lambda_{\min}(\mathbf{B})}$. Hence,

$$\sup_{\|\mathbf{u}\|_2=1,\mathrm{supp}(\mathbf{u})\subset F} \frac{\mathbf{u}^{\mathrm{T}}\mathbf{A}\mathbf{u} - \epsilon}{\mathbf{u}^{\mathrm{T}}\mathbf{B}\mathbf{u} + \epsilon} \geq (1 - \frac{\epsilon}{\lambda_{\min}(\mathbf{B})} - \frac{\epsilon}{\lambda_{\max}(\mathbf{B})\omega_1(F)})\omega_1(F) \geq (1 - a)\omega_1(F). \tag{S32}$$

The conclusion follows. □

S6.2 Additional technical lemmas

We first derive several lemmas concerning a parameter $\boldsymbol{\beta}_k$ (either in the penalized eigen-decomposition or the penalized generalized eigen-decomposition) and its estimate $\widehat{\boldsymbol{\beta}}_k$. We denote $\eta_k = |\sin\Theta(\boldsymbol{\beta}_k, \hat{\boldsymbol{\beta}}_k)|$, and $\hat{\lambda}_k = \widehat{\boldsymbol{\beta}}_k^{\mathrm{T}}\widehat{\mathbf{M}}\widehat{\boldsymbol{\beta}}_k$.

Lemma S4. *If $\|\boldsymbol{\beta}_k\|_0 \leq s, k = 1, \ldots, K$, we have that*

$$\|\mathrm{vec}(\boldsymbol{\beta}_k\boldsymbol{\beta}_k^{\mathrm{T}} - \hat{\boldsymbol{\beta}}_k\hat{\boldsymbol{\beta}}_k^{\mathrm{T}})\|_1 \leq 2s\eta_k \tag{S33}$$

Proof of Lemma S4. For (S33), set $\boldsymbol{\zeta}_k = \mathrm{vec}(\boldsymbol{\beta}_k\boldsymbol{\beta}_k^{\mathrm{T}} - \hat{\boldsymbol{\beta}}_k\hat{\boldsymbol{\beta}}_k^{\mathrm{T}})$. We have that

$$\begin{aligned} \|\boldsymbol{\zeta}_k\|_2^2 &= \|\boldsymbol{\beta}_k\boldsymbol{\beta}_k^{\mathrm{T}} - \hat{\boldsymbol{\beta}}_k\hat{\boldsymbol{\beta}}_k^{\mathrm{T}}\|_F^2 = \mathrm{Tr}((\boldsymbol{\beta}_k\boldsymbol{\beta}_k^{\mathrm{T}} - \hat{\boldsymbol{\beta}}_k\hat{\boldsymbol{\beta}}_k^{\mathrm{T}})(\boldsymbol{\beta}_k\boldsymbol{\beta}_k^{\mathrm{T}} - \hat{\boldsymbol{\beta}}_k\hat{\boldsymbol{\beta}}_k^{\mathrm{T}})) \quad \text{(S34)} \\ &= 2(1 - (\boldsymbol{\beta}_k^{\mathrm{T}}\hat{\boldsymbol{\beta}}_k)^2) = 2\eta_k^2 \quad \text{(S35)} \end{aligned}$$

Hence, by the Cauchy-Schwarz inequality,

$$\|\boldsymbol{\zeta}_k\|_1 \leq \sqrt{\|\boldsymbol{\zeta}_k\|_0}\|\boldsymbol{\zeta}_k\|_2 \leq \sqrt{2s^2} \cdot \sqrt{2\eta_k^2} = 2s\eta_k \quad \text{(S36)}$$

□

Lemma S5. *If $\|\boldsymbol{\beta}_k\|_0 \leq s$, we have that*

$$\|\mathrm{vec}(\boldsymbol{\beta}_k\boldsymbol{\beta}_k^{\mathrm{T}})\|_1 \leq s. \quad \text{(S37)}$$

Proof of Lemma S5. By the Cauchy-Schwarz inequality, we have that

$$\|\mathrm{vec}(\boldsymbol{\beta}_k\boldsymbol{\beta}_k^{\mathrm{T}})\|_1 \leq \sqrt{\|\mathrm{vec}(\boldsymbol{\beta}_k\boldsymbol{\beta}_k^{\mathrm{T}})\|_0}\|\mathrm{vec}(\boldsymbol{\beta}_k\boldsymbol{\beta}_k^{\mathrm{T}})\|_2 \quad \text{(S38)}$$

Note that $\|\mathrm{vec}(\boldsymbol{\beta}_k\boldsymbol{\beta}_k^{\mathrm{T}})\|_0 \leq s^2$ and

$$\|\mathrm{vec}(\boldsymbol{\beta}_k\boldsymbol{\beta}_k^{\mathrm{T}})\|_2^2 = \|\boldsymbol{\beta}_k\boldsymbol{\beta}_k^{\mathrm{T}}\|_F^2 = \mathrm{Tr}(\boldsymbol{\beta}_k\boldsymbol{\beta}_k^{\mathrm{T}}\boldsymbol{\beta}_k\boldsymbol{\beta}_k^{\mathrm{T}}) = 1 \quad \text{(S39)}$$

where we use the fact that $\boldsymbol{\beta}_k^{\mathrm{T}}\boldsymbol{\beta}_k = 1$. And we have the desired conclusion. □

Throughout the rest of this section, we also repeatedly use the fact that, for a vector $\mathbf{u}$, if $\|\mathbf{u}\|_2 = 1$, $\|\mathbf{u}\|_0 \leq s$, we must have that $\|\mathbf{u}\|_1 \leq \sqrt{s}$.

S6.3 Proof for Theorem 2

In this subsection, we assume that $\boldsymbol{\Sigma}_{\mathbf{X}} = \mathbf{I}$, $\widehat{\boldsymbol{\beta}}_k$ are solutions produced by Algorithm 1 for the penalized eigen-decomposition problem, and $\widehat{\lambda}_k = (\widehat{\boldsymbol{\beta}}_k)^{\mathrm{T}}\widehat{\mathbf{M}}\widehat{\boldsymbol{\beta}}_k$. We assume all the conditions in Theorem 2. We have the following result.

Lemma S6. *If $\|\widehat{\boldsymbol{\beta}}_k\|_0 \leq s$, $\|\boldsymbol{\beta}_k\|_0 \leq s$, we have that*

$$|\hat{\lambda}_k - \lambda_k| \leq s(2\eta_k + 1)\epsilon + (\lambda_1 + \lambda_k)\eta_k^2. \quad \text{(S40)}$$

Proof of Lemma S6. Note that

$$|\hat{\lambda}_k - \lambda_k| = |\langle \widehat{\mathbf{M}}, \hat{\boldsymbol{\beta}}_k \hat{\boldsymbol{\beta}}_k^{\mathrm{T}} \rangle - \langle \mathbf{M}, \boldsymbol{\beta}_k \boldsymbol{\beta}_k^{\mathrm{T}} \rangle| \tag{S41}$$

$$\leq |\langle \widehat{\mathbf{M}} - \mathbf{M}, \hat{\boldsymbol{\beta}}_k \hat{\boldsymbol{\beta}}_k^{\mathrm{T}} - \boldsymbol{\beta}_k \boldsymbol{\beta}_k^{\mathrm{T}} \rangle| + |\langle \widehat{\mathbf{M}} - \mathbf{M}, \boldsymbol{\beta}_k \boldsymbol{\beta}_k^{\mathrm{T}} \rangle| + |\langle \mathbf{M}, \hat{\boldsymbol{\beta}}_k \hat{\boldsymbol{\beta}}_k^{\mathrm{T}} - \boldsymbol{\beta}_k \boldsymbol{\beta}_k^{\mathrm{T}} \rangle| \tag{S42}$$

$$\equiv L_1 + L_2 + L_3 \tag{S43}$$

By Lemma S4,

$$L_1 \leq \|\widehat{\mathbf{M}} - \mathbf{M}\|_{\max} \|\mathrm{vec}(\hat{\boldsymbol{\beta}}_k \hat{\boldsymbol{\beta}}_k^{\mathrm{T}} - \boldsymbol{\beta}_k \boldsymbol{\beta}_k^{\mathrm{T}})\|_1 \leq 2s\eta_k \|\widehat{\mathbf{M}} - \mathbf{M}\|_{\max} \leq 2s\eta_k \epsilon. \tag{S44}$$

By Lemma S5,

$$L_2 \leq \|\widehat{\mathbf{M}} - \mathbf{M}\|_{\max} \|\mathrm{vec}(\boldsymbol{\beta}_k \boldsymbol{\beta}_k^{\mathrm{T}})\|_1 \leq s \|\widehat{\mathbf{M}} - \mathbf{M}\|_{\max} \leq s\epsilon. \tag{S45}$$

For L_3, note that $\mathbf{M} = \sum_{j=1}^p \lambda_j \boldsymbol{\beta}_j \boldsymbol{\beta}_j^{\mathrm{T}}$, which implies that

$$\langle \mathbf{M}, \widehat{\boldsymbol{\beta}}_k \widehat{\boldsymbol{\beta}}_k^{\mathrm{T}} \rangle = \sum_{j=1}^p \lambda_j (\widehat{\boldsymbol{\beta}}_k^{\mathrm{T}} \boldsymbol{\beta}_j)^2 = \sum_{j=1}^p \lambda_j \cos^2 \boldsymbol{\Theta}(\hat{\boldsymbol{\beta}}_k, \boldsymbol{\beta}_j).$$

Also note that $\sum_{j=1}^p \cos^2 \boldsymbol{\Theta}(\hat{\boldsymbol{\beta}}_k, \boldsymbol{\beta}_j) = 1$. Hence,

$$L_3 = |\sum_{j=1}^p \lambda_j \cos^2 \boldsymbol{\Theta}(\boldsymbol{\beta}_j, \hat{\boldsymbol{\beta}}_k) - \lambda_k| \tag{S46}$$

$$\leq \sum_{j \neq k} \lambda_j \cos^2 \boldsymbol{\Theta}(\widehat{\boldsymbol{\beta}}_k, \boldsymbol{\beta}_j) + \lambda_k |\cos^2 \boldsymbol{\Theta}(\widehat{\boldsymbol{\beta}}_k, \boldsymbol{\beta}_k) - 1| \tag{S47}$$

$$\leq \lambda_1 \sum_{j \neq k} \cos^2 \boldsymbol{\Theta}(\widehat{\boldsymbol{\beta}}_k, \boldsymbol{\beta}_j) + \lambda_k (1 - \cos^2 \boldsymbol{\Theta}(\widehat{\boldsymbol{\beta}}_k, \boldsymbol{\beta}_k)) \tag{S48}$$

$$\leq (\lambda_1 + \lambda_k)(1 - \cos^2 \boldsymbol{\Theta}(\boldsymbol{\beta}_k, \hat{\boldsymbol{\beta}}_k)) \tag{S49}$$

$$= (\lambda_1 + \lambda_k)\eta_k^2 \tag{S50}$$

□

In Lemmas S7–S9, we assume that the event $\|\widehat{\mathbf{M}} - \mathbf{M}\|_{\max} \leq \epsilon$ has happened.

Lemma S7. *For the first direction $\widehat{\boldsymbol{\beta}}_1$, we have that $|\sin \Theta(\widehat{\boldsymbol{\beta}}_1, \boldsymbol{\beta}_1)| \leq Cs\epsilon$ and $|\hat{\lambda}_1 - \lambda_1| \leq Cs\epsilon$.*

Proof of Lemma S7. It is easy to see that

$$\rho(\widehat{\mathbf{M}} - \mathbf{M}, s) = \sup_{\|\mathbf{u}\|_2 = 1, \|\mathbf{u}\|_0 \leq s} s|\mathbf{u}^{\mathrm{T}}(\widehat{\mathbf{M}} - \mathbf{M})\mathbf{u}| \leq \sup_{\|\mathbf{u}\|_2 = 1, \|\mathbf{u}\|_0 \leq s} \|\mathbf{u}\|_1^2 \cdot \|\widehat{\mathbf{M}} - \mathbf{M}\|_{\max} \leq s\epsilon. \tag{S51}$$

Under our assumptions about ϵ, by Proposition S2, we have $|\sin \Theta(\widehat{\boldsymbol{\beta}}_1, \boldsymbol{\beta}_1)| \leq Cs\epsilon$ at convergence. Lemma S6 further implies that $|\hat{\lambda}_1 - \lambda_1| \leq Cs\epsilon$. □

Lemma S8. *If* $|\sin\Theta(\widehat{\boldsymbol{\beta}}_j, \boldsymbol{\beta}_j)| \le Cs\epsilon$ *for sufficiently small* ϵ*, all* $j \le k$*, then* $\rho(\widehat{\mathbf{M}}_{k+1} - \mathbf{M}_{k+1}, s) \le Cs\epsilon$.

Proof of Lemma S8. Define $\mathbf{N}_l = \hat{\lambda}_l \hat{\boldsymbol{\beta}}_l \hat{\boldsymbol{\beta}}_l^{\mathrm{T}} - \lambda_l \boldsymbol{\beta}_l \boldsymbol{\beta}_l^{\mathrm{T}}$. Then

$$\widehat{\mathbf{M}}_{k+1} - \mathbf{M}_{k+1} = (\widehat{\mathbf{M}} - \mathbf{M}) - \sum_{l \le k} \mathbf{N}_l. \tag{S52}$$

It follows that

$$\rho(\widehat{\mathbf{M}}_{k+1} - \mathbf{M}_{k+1}, s) \le \rho(\widehat{\mathbf{M}} - \mathbf{M}, s) + \sum_{l \le k} \rho(\mathbf{N}_l, s). \tag{S53}$$

According to the proof in Lemma S7, $\rho(\widehat{\mathbf{M}} - \mathbf{M}, s) \le s\epsilon$.

For any vector $\mathbf{u}$, we have

$$\begin{aligned} \mathbf{u}^{\mathrm{T}} \mathbf{N}_l \mathbf{u} &= \widehat{\lambda}_l (\mathbf{u}^{\mathrm{T}} \widehat{\boldsymbol{\beta}}_l)^2 - \lambda_l (\mathbf{u}^{\mathrm{T}} \boldsymbol{\beta}_l)^2 && \text{(S54)} \\ &= (\widehat{\lambda}_l - \lambda_l)(\mathbf{u}^{\mathrm{T}} \widehat{\boldsymbol{\beta}}_l)^2 + \lambda_l \{(\mathbf{u}^{\mathrm{T}} \boldsymbol{\beta}_l)^2 - (\mathbf{u}^{\mathrm{T}} \widehat{\boldsymbol{\beta}}_l)^2\} \equiv L_1 + L_2 && \text{(S55)} \end{aligned}$$

By Lemma S6, we have that $|\hat{\lambda}_l - \lambda_l| \le Cs\epsilon$ when $|\sin\Theta(\widehat{\boldsymbol{\beta}}_j, \boldsymbol{\beta}_j)| \le Cs\epsilon$. Also, $|\mathbf{u}^{\mathrm{T}} \boldsymbol{\beta}_l| \le \|\mathbf{u}\|_2 \cdot \|\boldsymbol{\beta}_l\|_2 = 1$. It follows that $L_1 \le Cs\epsilon$. For L_2, we assume that $\cos\boldsymbol{\Theta}(\widehat{\boldsymbol{\beta}}_l, \boldsymbol{\beta}_l) > 0$ without loss of generality, because otherwise we can consider the proof for $-\widehat{\boldsymbol{\beta}}_l$. Note that

$$\begin{aligned} L_2 &\le \lambda_l |\mathbf{u}^{\mathrm{T}}(\widehat{\boldsymbol{\beta}}_l - \boldsymbol{\beta}_l)| \cdot |\mathbf{u}^{\mathrm{T}}(\widehat{\boldsymbol{\beta}}_l + \boldsymbol{\beta}_l)| \\ &\le \lambda_l \|\mathbf{u}\|_2^2 \cdot \|\widehat{\boldsymbol{\beta}}_l - \boldsymbol{\beta}_l\|_2 (\|\widehat{\boldsymbol{\beta}}_l\|_2 + \|\boldsymbol{\beta}_l\|_2) \\ &= C\|\widehat{\boldsymbol{\beta}}_l - \boldsymbol{\beta}_l\|_2 \\ &= C\sqrt{2 - 2\widehat{\boldsymbol{\beta}}_l^{\mathrm{T}} \boldsymbol{\beta}_l} = \sqrt{2(1 - \cos\Theta(\widehat{\boldsymbol{\beta}}_l, \boldsymbol{\beta}_l))} \\ &= C\sqrt{1 - \cos^2\Theta(\widehat{\boldsymbol{\beta}}_l, \boldsymbol{\beta}_l)} / \sqrt{1 + \cos\Theta(\widehat{\boldsymbol{\beta}}_l, \boldsymbol{\beta}_l)} \\ &\le C|\sin\Theta(\widehat{\boldsymbol{\beta}}_l, \boldsymbol{\beta}_l)| \\ &\le Cs\epsilon. \end{aligned}$$

Hence, $\rho(\mathbf{N}_l, s) \le Cs\epsilon$ and the conclusion follows. □

Lemma S9. *Assume that* $|\sin\Theta(\hat{\boldsymbol{\beta}}_l, \boldsymbol{\beta}_l)| < Cs\epsilon$ *for any* $l < k$*, where* $0 < \epsilon < \min\{\frac{\Delta}{4s}, \theta\}$ *for* θ *defined in Theorem 2. We have that* $|\sin\Theta(\hat{\boldsymbol{\beta}}_k, \boldsymbol{\beta}_k)| \le Cs\epsilon$.

Proof of Lemma S9. The conclusion follows from Lemma S8 and Proposition S2. □

Proof of Theorem 2. Under the event $\|\widehat{\mathbf{M}} - \mathbf{M}\|_{\max} \le \epsilon$, we have that $|\sin\Theta(\widehat{\boldsymbol{\beta}}_k, \boldsymbol{\beta}_k)| \le Cs\epsilon$ by Lemmas S7–S9. Then by Theorem 1 we have the desired conclusion. □

S6.4 Proof for Theorem 3

In this subsection we prove Theorem 3, where $\mathbf{\Sigma_X}$ could be different from the identity matrix. We first present a simple lemma, which is a modified version of Lemma 6 in Mai & Zhang (2019).

Lemma S10. *For two vectors* $\mathbf{u}, \mathbf{v}$ *and a positive definite matrix* $\mathbf{\Sigma}$*, define* $\xi_{\mathbf{I}} = 1 - \cos \mathbf{\Theta}(\mathbf{v}, \mathbf{u}), \xi_{\mathbf{\Sigma}} = 1 - \cos \mathbf{\Theta}(\mathbf{\Sigma}^{1/2}\mathbf{v}, \mathbf{\Sigma}^{1/2}\mathbf{u})$*. We have that*

$$\frac{\lambda_{\min}(\mathbf{\Sigma})}{\lambda_{\max}(\mathbf{\Sigma})}\xi_{\mathbf{\Sigma}} \leq \xi_{\mathbf{I}} \leq \frac{\lambda_{\max}(\mathbf{\Sigma})}{\lambda_{\min}(\mathbf{\Sigma})}\xi_{\mathbf{\Sigma}}$$

Proof of Lemma S10. We first show the latter half of the desired inequality. Without loss of generality, we assume that $\mathbf{u}^{\mathrm{T}}\mathbf{\Sigma}\mathbf{u} = 1, \mathbf{v}^{\mathrm{T}}\mathbf{\Sigma}\mathbf{v} = 1$, because we can always normalize $\mathbf{u}, \mathbf{v}$ to satisfy these conditions.

Note that

$$\begin{aligned}
\lambda_{\min}(\mathbf{\Sigma})(\mathbf{u}-\mathbf{v})^{\mathrm{T}}(\mathbf{u}-\mathbf{v}) &\leq (\mathbf{u}-\mathbf{v})^{\mathrm{T}}\mathbf{\Sigma}(\mathbf{u}-\mathbf{v}) \\
&= \mathbf{u}^{\mathrm{T}}\mathbf{\Sigma}\mathbf{u} - 2\mathbf{u}^{\mathrm{T}}\mathbf{\Sigma}\mathbf{v} + \mathbf{v}^{\mathrm{T}}\mathbf{\Sigma}\mathbf{v} = 2\xi_{\mathbf{\Sigma}} \\
\lambda_{\max}(\mathbf{\Sigma})\mathbf{u}^{\mathrm{T}}\mathbf{u} &\geq \mathbf{u}^{\mathrm{T}}\mathbf{\Sigma}\mathbf{u} = 1 \\
\lambda_{\max}(\mathbf{\Sigma})\mathbf{v}^{\mathrm{T}}\mathbf{v} &\geq \mathbf{v}^{\mathrm{T}}\mathbf{\Sigma}\mathbf{v} = 1
\end{aligned}$$

Consequently,

$$\begin{aligned}
&(\mathbf{u}-\mathbf{v})^{\mathrm{T}}(\mathbf{u}-\mathbf{v}) \leq \frac{2\xi_{\mathbf{\Sigma}}}{\lambda_{\min}(\mathbf{\Sigma})} \\
&\mathbf{u}^{\mathrm{T}}\mathbf{u} \geq 1/\lambda_{\max}(\mathbf{\Sigma}), \mathbf{v}^{\mathrm{T}}\mathbf{v} \geq 1/\lambda_{\max}(\mathbf{\Sigma})
\end{aligned}$$

Now

$$\begin{aligned}
\frac{2\lambda_{\max}(\mathbf{\Sigma})\xi_{\mathbf{\Sigma}}}{\lambda_{\min}(\mathbf{\Sigma})} &\geq \frac{2\xi_{\mathbf{\Sigma}}}{\lambda_{\min}(\mathbf{\Sigma})\sqrt{\mathbf{u}^{\mathrm{T}}\mathbf{u}}\sqrt{\mathbf{v}^{\mathrm{T}}\mathbf{v}}} \\
&\geq \frac{(\mathbf{u}-\mathbf{v})^{\mathrm{T}}(\mathbf{u}-\mathbf{v})}{\sqrt{\mathbf{u}^{\mathrm{T}}\mathbf{u}}\sqrt{\mathbf{v}^{\mathrm{T}}\mathbf{v}}} = \frac{\sqrt{\mathbf{u}^{\mathrm{T}}\mathbf{u}}}{\sqrt{\mathbf{v}^{\mathrm{T}}\mathbf{v}}} + \frac{\sqrt{\mathbf{v}^{\mathrm{T}}\mathbf{v}}}{\sqrt{\mathbf{u}^{\mathrm{T}}\mathbf{u}}} - 2\frac{\mathbf{u}^{\mathrm{T}}\mathbf{v}}{\sqrt{\mathbf{u}^{\mathrm{T}}\mathbf{u}}\sqrt{\mathbf{v}^{\mathrm{T}}\mathbf{v}}} \\
&\geq 2 - 2\frac{\mathbf{u}^{\mathrm{T}}\mathbf{v}}{\sqrt{\mathbf{u}^{\mathrm{T}}\mathbf{u}}\sqrt{\mathbf{v}^{\mathrm{T}}\mathbf{v}}} = 2\xi_{\mathbf{I}},
\end{aligned}$$

and we have the second half of the inequality. For the first half of the inequality, define $\mathbf{u}^* = \mathbf{\Sigma}^{1/2}\mathbf{u}, \mathbf{v}^* = \mathbf{\Sigma}^{1/2}$. Apply the second half of the inequality to vectors $\mathbf{u}^*, \mathbf{v}^*$ and matrix $\mathbf{\Sigma}^{-1}$, and we have the desired conclusion. □

We now show the following lemma parallel to Lemma S6. Recall that $\eta_k = |\sin \mathbf{\Theta}(\widehat{\boldsymbol{\beta}}_k, \boldsymbol{\beta}_k)|$.

Lemma S11. *If $\|\widehat{\boldsymbol{\beta}}_k\|_0 \leq s, \|\boldsymbol{\beta}_k\|_0 \leq s$, we have that*

$$|\hat{\lambda}_k - \lambda_k| \leq s(2\eta_k + 1)\epsilon + 2\frac{\lambda_{\max}(\mathbf{\Sigma_X})}{\lambda_{\min}(\mathbf{\Sigma_X})}(\lambda_1 + \lambda_k)\eta_k^2. \tag{S56}$$

Proof of Lemma S11. Note that

$$\begin{aligned}
&|\hat{\lambda}_k - \lambda_k| = |\langle \widehat{\mathbf{M}}, \hat{\boldsymbol{\beta}}_k\hat{\boldsymbol{\beta}}_k^{\mathrm{T}}\rangle - \langle \mathbf{M}, \boldsymbol{\beta}_k\boldsymbol{\beta}_k^{\mathrm{T}}\rangle| && \text{(S57)}\\
\leq\ &|\langle \widehat{\mathbf{M}} - \mathbf{M}, \hat{\boldsymbol{\beta}}_k\hat{\boldsymbol{\beta}}_k^{\mathrm{T}} - \boldsymbol{\beta}_k\boldsymbol{\beta}_k^{\mathrm{T}}\rangle| + |\langle \widehat{\mathbf{M}} - \mathbf{M}, \boldsymbol{\beta}_k\boldsymbol{\beta}_k^{\mathrm{T}}\rangle| + |\langle \mathbf{M}, \hat{\boldsymbol{\beta}}_k\hat{\boldsymbol{\beta}}_k^{\mathrm{T}} - \boldsymbol{\beta}_k\boldsymbol{\beta}_k^{\mathrm{T}}\rangle| && \text{(S58)}\\
\equiv\ &L_1 + L_2 + L_3 && \text{(S59)}
\end{aligned}$$

By Lemma S4,

$$L_1 \leq \|\widehat{\mathbf{M}} - \mathbf{M}\|_{\max}\|\mathrm{vec}(\hat{\boldsymbol{\beta}}_k\hat{\boldsymbol{\beta}}_k^{\mathrm{T}} - \boldsymbol{\beta}_k\boldsymbol{\beta}_k^{\mathrm{T}})\|_1 \leq 2s\eta_k\|\widehat{\mathbf{M}} - \mathbf{M}\|_{\max} \leq 2s\eta_k\epsilon. \tag{S60}$$

By Lemma S5,

$$L_2 \leq \|\widehat{\mathbf{M}} - \mathbf{M}\|_{\max}\|\mathrm{vec}(\boldsymbol{\beta}_k\boldsymbol{\beta}_k^{\mathrm{T}})\|_1 \leq s\|\widehat{\mathbf{M}} - \mathbf{M}\|_{\max} \leq s\epsilon. \tag{S61}$$

For L_3, note that $\mathbf{M} = \mathbf{\Sigma_X}\left(\sum_{j=1}^p \lambda_j\boldsymbol{\beta}_j\boldsymbol{\beta}_j^{\mathrm{T}}\right)\mathbf{\Sigma_X}$ by the proof of Lemma 1, which implies that

$$\langle \mathbf{M}, \widehat{\boldsymbol{\beta}}_k\widehat{\boldsymbol{\beta}}_k^{\mathrm{T}}\rangle = \sum_{j=1}^p \lambda_j(\widehat{\boldsymbol{\beta}}_k^{\mathrm{T}}\mathbf{\Sigma_X}\boldsymbol{\beta}_j)^2 = \sum_{j=1}^p \lambda_j \cos^2\mathbf{\Theta}(\mathbf{\Sigma_X}^{1/2}\hat{\boldsymbol{\beta}}_k, \mathbf{\Sigma_X}^{1/2}\boldsymbol{\beta}_j)$$

Also note that $\sum_{j=1}^p \cos^2\mathbf{\Theta}(\mathbf{\Sigma_X}^{1/2}\hat{\boldsymbol{\beta}}_k, \mathbf{\Sigma_X}^{1/2}\boldsymbol{\beta}_j) = 1$. Without loss of generality, assume that $\cos\mathbf{\Theta}(\boldsymbol{\beta}_k, \hat{\boldsymbol{\beta}}_k) > 0$. We have that

$$\begin{aligned}
L_3 &= |\sum_{j=1}^p \lambda_j \cos^2\mathbf{\Theta}(\mathbf{\Sigma_X}^{1/2}\boldsymbol{\beta}_j, \mathbf{\Sigma_X}^{1/2}\hat{\boldsymbol{\beta}}_k) - \lambda_k|\\
&\leq \sum_{j\neq k} \lambda_j \cos^2\mathbf{\Theta}(\mathbf{\Sigma_X}^{1/2}\widehat{\boldsymbol{\beta}}_k, \mathbf{\Sigma_X}^{1/2}\boldsymbol{\beta}_j) + \lambda_k|\cos^2\mathbf{\Theta}(\mathbf{\Sigma_X}^{1/2}\widehat{\boldsymbol{\beta}}_k, \mathbf{\Sigma_X}^{1/2}\boldsymbol{\beta}_k) - 1|\\
&\leq \lambda_1\sum_{j\neq k} \cos^2\mathbf{\Theta}(\mathbf{\Sigma_X}^{1/2}\widehat{\boldsymbol{\beta}}_k, \mathbf{\Sigma_X}^{1/2}\boldsymbol{\beta}_j) + \lambda_k(1 - \cos^2\mathbf{\Theta}(\mathbf{\Sigma_X}^{1/2}\widehat{\boldsymbol{\beta}}_k, \mathbf{\Sigma_X}^{1/2}\boldsymbol{\beta}_k))\\
&\leq (\lambda_1 + \lambda_k)(1 - \cos^2\mathbf{\Theta}(\mathbf{\Sigma_X}^{1/2}\boldsymbol{\beta}_k, \mathbf{\Sigma_X}^{1/2}\hat{\boldsymbol{\beta}}_k))\\
&= (\lambda_1 + \lambda_k)(1 - \cos\mathbf{\Theta}(\mathbf{\Sigma_X}^{1/2}\boldsymbol{\beta}_k, \mathbf{\Sigma_X}^{1/2}\hat{\boldsymbol{\beta}}_k))(1 + \cos\mathbf{\Theta}(\mathbf{\Sigma_X}^{1/2}\boldsymbol{\beta}_k, \mathbf{\Sigma_X}^{1/2}\hat{\boldsymbol{\beta}}_k))\\
&\leq 2(\lambda_1 + \lambda_k)(1 - \cos\mathbf{\Theta}(\mathbf{\Sigma_X}^{1/2}\boldsymbol{\beta}_k, \mathbf{\Sigma_X}^{1/2}\hat{\boldsymbol{\beta}}_k))\\
&\leq 2\frac{\lambda_{\max}(\mathbf{\Sigma_X})}{\lambda_{\min}(\mathbf{\Sigma_X})}(\lambda_1 + \lambda_k)(1 - \cos\mathbf{\Theta}(\boldsymbol{\beta}_k, \hat{\boldsymbol{\beta}}_k)),
\end{aligned}$$

where the last inequality follows from Lemma S10. Further note that $1-\cos\boldsymbol{\Theta}(\boldsymbol{\beta}_k,\hat{\boldsymbol{\beta}}_k) = \dfrac{\sin^2\boldsymbol{\Theta}(\boldsymbol{\beta}_k,\hat{\boldsymbol{\beta}}_k)}{1+\cos\boldsymbol{\Theta}(\boldsymbol{\beta}_k,\hat{\boldsymbol{\beta}}_k)} \le \eta_k^2$ and we have the desired conclusion. □

We consider the event $\|\widehat{\mathbf{M}}-\mathbf{M}\|_{\max}\le\epsilon, \|\widehat{\boldsymbol{\Sigma}}_{\mathbf{X}}-\boldsymbol{\Sigma}_{\mathbf{X}}\|_{\max}\le\epsilon$ for $\sqrt{2}s\epsilon < \min\{1/2, \lambda_{\min}(\boldsymbol{\Sigma}_{\mathbf{X}})\}$ and $\frac{\sqrt{2}s\epsilon}{\lambda_{\min}(\boldsymbol{\Sigma}_{\mathbf{X}})} + \frac{\sqrt{2}s\epsilon}{\lambda_{\min}(\boldsymbol{\Sigma}_{\mathbf{X}})\lambda_1} < a^*$ and $\dfrac{\sqrt{2}s\epsilon}{\lambda_{\min}(\boldsymbol{\Sigma}_{\mathbf{X}})} < \frac{\Delta}{2(1+\lambda_1^2)}$, where Δ is defined as in Condition (C2) and $a^* = \min\{\frac{1}{2}, \frac{\Delta}{\lambda_1+\lambda_2}, \frac{\lambda_{\min}(\boldsymbol{\Sigma}_{\mathbf{X}})}{2}\}$. As a direct consequence, $\rho(\widehat{\boldsymbol{\Sigma}}_{\mathbf{X}}-\boldsymbol{\Sigma}_{\mathbf{X}}, s)\le s\epsilon$. Also note that $\tilde{\boldsymbol{\beta}}_k = \dfrac{\widehat{\boldsymbol{\beta}}_k}{\|\widehat{\boldsymbol{\beta}}_k\|_2}$. In the RIFLE algorithm, $\boldsymbol{\beta}_k^0$ is chosen to be sufficiently close to $\boldsymbol{\beta}_k$, and the step size η satisfies that $\eta \le \frac{\lambda_{\min}(\boldsymbol{\Sigma}_{\mathbf{X}})}{2}$ and

$$\sqrt{1+2\{(d/s')^{1/2}+d/s'\}}\sqrt{1-\frac{1}{16}\eta\lambda_{\min}(\boldsymbol{\Sigma}_{\mathbf{X}})\frac{\frac{\Delta}{\lambda_K}}{\kappa(\boldsymbol{\Sigma}_{\mathbf{X}})+1}} < 1. \tag{S62}$$

Without loss of generality, in what follows we assume that $\cos\boldsymbol{\Theta}(\widehat{\boldsymbol{\beta}}_j,\boldsymbol{\beta}_k) > 0$, because otherwise we can always consider $-\widehat{\boldsymbol{\beta}}_j$, which spans the same susbspace as $\widehat{\boldsymbol{\beta}}_j$.

Lemma S12. *For the first direction $\widehat{\boldsymbol{\beta}}_1$, we have that $|\sin\Theta(\widehat{\boldsymbol{\beta}}_1,\boldsymbol{\beta}_1)| \le Cs\epsilon$ and $|\hat{\lambda}_1-\lambda_1|\le Cs\epsilon$.*

Proof of Lemma S12. It is easy to see that

$$\rho(\widehat{\mathbf{M}}-\mathbf{M},s) = \sup_{\|\mathbf{u}\|_2=1,\|\mathbf{u}\|_0\le s} s|\mathbf{u}^{\mathrm{T}}(\widehat{\mathbf{M}}-\mathbf{M})\mathbf{u}| \le \sup_{\|\mathbf{u}\|_2=1,\|\mathbf{u}\|_0\le s} \|\mathbf{u}\|_1^2\cdot\|\widehat{\mathbf{M}}-\mathbf{M}\|_{\max} \le s\epsilon. \tag{S63}$$

Similarly, $\rho(\widehat{\boldsymbol{\Sigma}}_{\mathbf{X}}-\boldsymbol{\Sigma}_{\mathbf{X}},s)\le s\epsilon$. Denote $\delta_1(s) = \sqrt{\rho^2(\widehat{\mathbf{M}}-\mathbf{M},s)+\rho^2(\widehat{\boldsymbol{\Sigma}}_{\mathbf{X}}-\boldsymbol{\Sigma}_{\mathbf{X}},s)}$. It follows that $\delta_1(s)\le\sqrt{2}s\epsilon$. Under our assumptions about ϵ, we have that $|\sin\Theta(\widehat{\boldsymbol{\beta}}_1,\boldsymbol{\beta}_1)|\le Cs\epsilon$ by Lemma S2. Lemma S11 further implies that $|\hat{\lambda}_1-\lambda_1|\le Cs\epsilon$. □

Lemma S13. *If $|\sin\Theta(\widehat{\boldsymbol{\beta}}_j,\boldsymbol{\beta}_j)|\le Cs\epsilon$ for sufficiently small ϵ, all $j\le k$, then $\delta(s)\le Cs\epsilon$.*

Proof of Lemma S13. It suffices to show that $\rho(\widehat{\mathbf{M}}_{k+1}-\mathbf{M}_{k+1},s)\le Cs\epsilon$. Define $\mathbf{N}_l = \widehat{\boldsymbol{\Sigma}}_{\mathbf{X}}\hat{\lambda}_l\hat{\boldsymbol{\beta}}_l\hat{\boldsymbol{\beta}}_l^{\mathrm{T}}\widehat{\boldsymbol{\Sigma}}_{\mathbf{X}} - \boldsymbol{\Sigma}_{\mathbf{X}}\lambda_l\boldsymbol{\beta}_l\boldsymbol{\beta}_l^{\mathrm{T}}\boldsymbol{\Sigma}_{\mathbf{X}}$. Then

$$\widehat{\mathbf{M}}_{k+1}-\mathbf{M}_{k+1} = (\widehat{\mathbf{M}}-\mathbf{M}) - \sum_{l\le k}\mathbf{N}_l. \tag{S64}$$

It follows that

$$\rho(\widehat{\mathbf{M}}_{k+1}-\mathbf{M}_{k+1},s) \le \rho(\widehat{\mathbf{M}}-\mathbf{M},s) + \sum_{l\le k}\rho(\mathbf{N}_l,s). \tag{S65}$$

According to the proof in Lemma S12, $\rho(\widehat{\mathbf{M}}-\mathbf{M},s)\le s\epsilon$.

For any vector $\mathbf{u}$, we have

$$\begin{aligned}
\mathbf{u}^{\mathrm{T}}\mathbf{N}_l\mathbf{u} &= \hat{\lambda}_l(\mathbf{u}^{\mathrm{T}}\widehat{\boldsymbol{\Sigma}}_{\mathbf{X}}\hat{\boldsymbol{\beta}}_l)^2 - \lambda_l(\mathbf{u}^{\mathrm{T}}\boldsymbol{\Sigma}_{\mathbf{X}}\boldsymbol{\beta}_l)^2 && \text{(S66)}\\
&= (\hat{\lambda}_l - \lambda_l)(\mathbf{u}^{\mathrm{T}}\widehat{\boldsymbol{\Sigma}}_{\mathbf{X}}\widehat{\boldsymbol{\beta}}_l)^2 + \lambda_l\{(\mathbf{u}^{\mathrm{T}}\widehat{\boldsymbol{\Sigma}}_{\mathbf{X}}\widehat{\boldsymbol{\beta}}_l)^2 - (\mathbf{u}^{\mathrm{T}}\boldsymbol{\Sigma}_{\mathbf{X}}\boldsymbol{\beta}_l)\} \equiv L_1 + L_2 && \text{(S67)}
\end{aligned}$$

By Lemma S11, we have that $|\hat{\lambda}_l - \lambda_l| \le Cs\epsilon$ when $\sin\Theta(\widehat{\boldsymbol{\beta}}_j, \boldsymbol{\beta}_j) \le Cs\epsilon$. Also,

$$\begin{aligned}
|\mathbf{u}^{\mathrm{T}}\widehat{\boldsymbol{\Sigma}}_{\mathbf{X}}\widehat{\boldsymbol{\beta}}_l| &\le |\mathbf{u}^{\mathrm{T}}\boldsymbol{\Sigma}_{\mathbf{X}}\widehat{\boldsymbol{\beta}}_l| + \|\mathbf{u}\|_1 \cdot \|\widehat{\boldsymbol{\beta}}_l\|_1\|\widehat{\boldsymbol{\Sigma}}_{\mathbf{X}} - \boldsymbol{\Sigma}_{\mathbf{X}}\|_{\max} \le |\mathbf{u}^{\mathrm{T}}\boldsymbol{\Sigma}_{\mathbf{X}}\boldsymbol{\beta}_l| + s\epsilon && \text{(S68)}\\
&\le \sqrt{(\mathbf{u}^{\mathrm{T}}\boldsymbol{\Sigma}_{\mathbf{X}}\mathbf{u}) \cdot (\widehat{\boldsymbol{\beta}}_l^{\mathrm{T}}\boldsymbol{\Sigma}_{\mathbf{X}}\widehat{\boldsymbol{\beta}}_l)} + s\epsilon \le \lambda_{\max}(\boldsymbol{\Sigma}_{\mathbf{X}}) + s\epsilon. && \text{(S69)}
\end{aligned}$$

It follows that $L_1 \le Cs\epsilon$. For L_2, note that $(\mathbf{u}^{\mathrm{T}}\widehat{\boldsymbol{\Sigma}}_{\mathbf{X}}\widehat{\boldsymbol{\beta}}_l)^2 - (\mathbf{u}^{\mathrm{T}}\boldsymbol{\Sigma}_{\mathbf{X}}\boldsymbol{\beta}_l) = (\mathbf{u}^{\mathrm{T}}\widehat{\boldsymbol{\Sigma}}_{\mathbf{X}}\widehat{\boldsymbol{\beta}}_l + \mathbf{u}^{\mathrm{T}}\boldsymbol{\Sigma}_{\mathbf{X}}\boldsymbol{\beta}_l)(\mathbf{u}^{\mathrm{T}}\widehat{\boldsymbol{\Sigma}}_{\mathbf{X}}\widehat{\boldsymbol{\beta}}_l - \mathbf{u}^{\mathrm{T}}\boldsymbol{\Sigma}_{\mathbf{X}}\boldsymbol{\beta}_l)$. On one hand,

$$\begin{aligned}
&|\mathbf{u}^{\mathrm{T}}\widehat{\boldsymbol{\Sigma}}_{\mathbf{X}}\widehat{\boldsymbol{\beta}}_l + \mathbf{u}^{\mathrm{T}}\boldsymbol{\Sigma}_{\mathbf{X}}\boldsymbol{\beta}_l| \le |\mathbf{u}^{\mathrm{T}}\widehat{\boldsymbol{\Sigma}}_{\mathbf{X}}\widehat{\boldsymbol{\beta}}_l| + |\mathbf{u}^{\mathrm{T}}\boldsymbol{\Sigma}_{\mathbf{X}}\boldsymbol{\beta}_l| && \text{(S70)}\\
\le\ & \sqrt{\mathbf{u}^{\mathrm{T}}\widehat{\boldsymbol{\Sigma}}_{\mathbf{X}}\mathbf{u} \cdot \widehat{\boldsymbol{\beta}}_l^{\mathrm{T}}\widehat{\boldsymbol{\Sigma}}_{\mathbf{X}}\widehat{\boldsymbol{\beta}}_l} + \sqrt{\mathbf{u}^{\mathrm{T}}\boldsymbol{\Sigma}\mathbf{u} \cdot \boldsymbol{\beta}_l^{\mathrm{T}}\boldsymbol{\Sigma}\boldsymbol{\beta}_l} && \text{(S71)}\\
=\ & \sqrt{\mathbf{u}^{\mathrm{T}}\widehat{\boldsymbol{\Sigma}}_{\mathbf{X}}\mathbf{u}} + \sqrt{\mathbf{u}^{\mathrm{T}}\boldsymbol{\Sigma}_{\mathbf{X}}\mathbf{u}} = \sqrt{\mathbf{u}^{\mathrm{T}}\boldsymbol{\Sigma}_{\mathbf{X}}\mathbf{u} + \mathbf{u}^{\mathrm{T}}(\widehat{\boldsymbol{\Sigma}}_{\mathbf{X}} - \boldsymbol{\Sigma}_{\mathbf{X}})\mathbf{u}} + \sqrt{\mathbf{u}^{\mathrm{T}}\boldsymbol{\Sigma}\mathbf{u}} && \text{(S72)}\\
\le\ & \sqrt{\lambda_{\max}(\boldsymbol{\Sigma}_{\mathbf{X}}) + s\epsilon} + \sqrt{\lambda_{\max}(\boldsymbol{\Sigma}_{\mathbf{X}})} && \text{(S73)}
\end{aligned}$$

On the other hand,

$$\begin{aligned}
|\mathbf{u}^{\mathrm{T}}\widehat{\boldsymbol{\Sigma}}_{\mathbf{X}}\hat{\boldsymbol{\beta}}_l - \mathbf{u}^{\mathrm{T}}\boldsymbol{\Sigma}_{\mathbf{X}}\boldsymbol{\beta}_l| &\le |\mathbf{u}^{\mathrm{T}}(\widehat{\boldsymbol{\Sigma}}_{\mathbf{X}} - \boldsymbol{\Sigma}_{\mathbf{X}})\hat{\boldsymbol{\beta}}_l| + |\mathbf{u}^{\mathrm{T}}\boldsymbol{\Sigma}_{\mathbf{X}}(\hat{\boldsymbol{\beta}}_l - \boldsymbol{\beta}_l)| && \text{(S74)}\\
&\le s\epsilon + |\mathbf{u}^{\mathrm{T}}\boldsymbol{\Sigma}_{\mathbf{X}}(\hat{\boldsymbol{\beta}}_l - \boldsymbol{\beta}_l)| && \text{(S75)}\\
&\equiv s\epsilon + L_3. && \text{(S76)}
\end{aligned}$$

Further note that

$$\begin{aligned}
L_3 &\le \sqrt{\mathbf{u}^{\mathrm{T}}\boldsymbol{\Sigma}_{\mathbf{X}}\mathbf{u} \cdot (\hat{\boldsymbol{\beta}}_l - \boldsymbol{\beta}_l)^{\mathrm{T}}\boldsymbol{\Sigma}_{\mathbf{X}}(\hat{\boldsymbol{\beta}}_l - \boldsymbol{\beta}_l)} \le \lambda_{\max}(\boldsymbol{\Sigma}_{\mathbf{X}}) \cdot \|\hat{\boldsymbol{\beta}}_l - \boldsymbol{\beta}_l\|_2 && \text{(S77)}\\
&= \lambda_{\max}(\boldsymbol{\Sigma}_{\mathbf{X}}) \cdot \frac{1}{\sqrt{\tilde{\boldsymbol{\beta}}_l^{\mathrm{T}}\widehat{\boldsymbol{\Sigma}}_{\mathbf{X}}\tilde{\boldsymbol{\beta}}_l}}\|\tilde{\boldsymbol{\beta}}_l - \sqrt{\tilde{\boldsymbol{\beta}}^{\mathrm{T}}\widehat{\boldsymbol{\Sigma}}_{\mathbf{X}}\tilde{\boldsymbol{\beta}}_l} \cdot \boldsymbol{\beta}_l\|_2 && \text{(S78)}\\
&\le \lambda_{\max}(\boldsymbol{\Sigma}_{\mathbf{X}}) \cdot \frac{1}{\sqrt{\tilde{\boldsymbol{\beta}}_l^{\mathrm{T}}\widehat{\boldsymbol{\Sigma}}_{\mathbf{X}}\tilde{\boldsymbol{\beta}}_l}}\{\|\tilde{\boldsymbol{\beta}}_l - \boldsymbol{\beta}_l^*\|_2 + |\frac{1}{\|\boldsymbol{\beta}_l\|_2} - \sqrt{\tilde{\boldsymbol{\beta}}_l^{\mathrm{T}}\widehat{\boldsymbol{\Sigma}}_{\mathbf{X}}\tilde{\boldsymbol{\beta}}_l}| \cdot \|\boldsymbol{\beta}_l\|_2\}, && \text{(S79)}
\end{aligned}$$

where $\boldsymbol{\beta}_l^* = \dfrac{\boldsymbol{\beta}_l}{\|\boldsymbol{\beta}_l\|_2}$. By our assumption, $\|\tilde{\boldsymbol{\beta}}_l - \boldsymbol{\beta}_l^*\|_2 \le \sqrt{2}\sin\Theta(\boldsymbol{\beta}_l^*, \boldsymbol{\beta}_l) \le Cs\epsilon$. For the second

term, since $(\boldsymbol{\beta}_l^*)^{\mathrm{T}}\boldsymbol{\Sigma}_{\mathbf{X}}\boldsymbol{\beta}_l^* = \dfrac{\boldsymbol{\beta}_l^{\mathrm{T}}\boldsymbol{\Sigma}_{\mathbf{X}}\boldsymbol{\beta}_l}{\|\boldsymbol{\beta}_l\|_2^2} = \dfrac{1}{\|\boldsymbol{\beta}_l\|_2^2}$, we have $\|\boldsymbol{\beta}_l\|_2 = \dfrac{1}{\sqrt{(\boldsymbol{\beta}_l^*)^{\mathrm{T}}\boldsymbol{\Sigma}_{\mathbf{X}}\boldsymbol{\beta}_l^*}}$. It follows that,

$$
\begin{aligned}
&|\frac{1}{\|\boldsymbol{\beta}_l\|_2} - \sqrt{\tilde{\boldsymbol{\beta}}_l^{\mathrm{T}}\widehat{\boldsymbol{\Sigma}}_{\mathbf{X}}\tilde{\boldsymbol{\beta}}_l}| = |\sqrt{(\boldsymbol{\beta}_l^*)^{\mathrm{T}}\boldsymbol{\Sigma}_{\mathbf{X}}\boldsymbol{\beta}_l^*} - \sqrt{\tilde{\boldsymbol{\beta}}_l^{\mathrm{T}}\widehat{\boldsymbol{\Sigma}}_{\mathbf{X}}\tilde{\boldsymbol{\beta}}_l}| && \text{(S80)}\\
\leq\ & |\sqrt{(\boldsymbol{\beta}_l^*)^{\mathrm{T}}\boldsymbol{\Sigma}_{\mathbf{X}}\boldsymbol{\beta}_l^*} - \sqrt{\tilde{\boldsymbol{\beta}}_l^{\mathrm{T}}\boldsymbol{\Sigma}_{\mathbf{X}}\tilde{\boldsymbol{\beta}}_l}| + |\tilde{\boldsymbol{\beta}}_l^{\mathrm{T}}(\boldsymbol{\Sigma}_{\mathbf{X}} - \widehat{\boldsymbol{\Sigma}}_{\mathbf{X}})\tilde{\boldsymbol{\beta}}_l|/(\sqrt{\tilde{\boldsymbol{\beta}}^{\mathrm{T}}\boldsymbol{\Sigma}_{\mathbf{X}}\tilde{\boldsymbol{\beta}}} + \sqrt{\tilde{\boldsymbol{\beta}}^{\mathrm{T}}\widehat{\boldsymbol{\Sigma}}_{\mathbf{X}}\tilde{\boldsymbol{\beta}}}) && \text{(S81)}\\
\leq\ & \frac{|(\boldsymbol{\beta}_l^*)^{\mathrm{T}}\boldsymbol{\Sigma}_{\mathbf{X}}\boldsymbol{\beta}_l^* - \tilde{\boldsymbol{\beta}}_l^{\mathrm{T}}\widehat{\boldsymbol{\Sigma}}_{\mathbf{X}}\tilde{\boldsymbol{\beta}}_l|}{\sqrt{(\boldsymbol{\beta}_l^*)^{\mathrm{T}}\boldsymbol{\Sigma}_{\mathbf{X}}\boldsymbol{\beta}_l^*} + \sqrt{\tilde{\boldsymbol{\beta}}_l^{\mathrm{T}}\widehat{\boldsymbol{\Sigma}}_{\mathbf{X}}\tilde{\boldsymbol{\beta}}_l}} + \frac{s\epsilon}{\lambda_{\min}(\boldsymbol{\Sigma}_{\mathbf{X}})}. && \text{(S82)}
\end{aligned}
$$

Because $\sqrt{(\boldsymbol{\beta}_l^*)^{\mathrm{T}}\boldsymbol{\Sigma}_{\mathbf{X}}\boldsymbol{\beta}_l^*} + \sqrt{\tilde{\boldsymbol{\beta}}_l^{\mathrm{T}}\widehat{\boldsymbol{\Sigma}}_{\mathbf{X}}\tilde{\boldsymbol{\beta}}_l} \geq \lambda_{\min}(\boldsymbol{\Sigma}_{\mathbf{X}})$, it suffices to find a bound for $|(\boldsymbol{\beta}_l^*)^{\mathrm{T}}\boldsymbol{\Sigma}_{\mathbf{X}}\boldsymbol{\beta}_l^* - \tilde{\boldsymbol{\beta}}_l^{\mathrm{T}}\widehat{\boldsymbol{\Sigma}}_{\mathbf{X}}\tilde{\boldsymbol{\beta}}_l|$.

$$
\begin{aligned}
&|(\boldsymbol{\beta}_l^*)^{\mathrm{T}}\boldsymbol{\Sigma}_{\mathbf{X}}\boldsymbol{\beta}_l^* - \tilde{\boldsymbol{\beta}}_l^{\mathrm{T}}\widehat{\boldsymbol{\Sigma}}_{\mathbf{X}}\tilde{\boldsymbol{\beta}}_l| \leq |(\boldsymbol{\beta}_l^*)^{\mathrm{T}}\boldsymbol{\Sigma}_{\mathbf{X}}\boldsymbol{\beta}_l^* - \tilde{\boldsymbol{\beta}}_l^{\mathrm{T}}\boldsymbol{\Sigma}_{\mathbf{X}}\tilde{\boldsymbol{\beta}}_l| + s\epsilon && \text{(S83)}\\
\leq\ & |(\boldsymbol{\beta}_l^* - \tilde{\boldsymbol{\beta}}_l)^{\mathrm{T}}\boldsymbol{\Sigma}_{\mathbf{X}}\boldsymbol{\beta}_l^*| + |\tilde{\boldsymbol{\beta}}_l^{\mathrm{T}}\boldsymbol{\Sigma}_{\mathbf{X}}(\boldsymbol{\beta}_l^* - \tilde{\boldsymbol{\beta}}_l)| + s\epsilon && \text{(S84)}\\
\leq\ & \sqrt{(\boldsymbol{\beta}_l^* - \tilde{\boldsymbol{\beta}}_l)^{\mathrm{T}}\boldsymbol{\Sigma}_{\mathbf{X}}(\boldsymbol{\beta}_l^* - \tilde{\boldsymbol{\beta}}_l)\cdot(\boldsymbol{\beta}_l^*)^{\mathrm{T}}\boldsymbol{\Sigma}_{\mathbf{X}}\boldsymbol{\beta}_l^*} + \sqrt{\tilde{\boldsymbol{\beta}}_l^{\mathrm{T}}\boldsymbol{\Sigma}_{\mathbf{X}}\tilde{\boldsymbol{\beta}}_l\cdot(\boldsymbol{\beta}_l^* - \tilde{\boldsymbol{\beta}}_l)^{\mathrm{T}}\boldsymbol{\Sigma}_{\mathbf{X}}(\boldsymbol{\beta}_l^* - \tilde{\boldsymbol{\beta}}_l)}\\
&+s\epsilon && \text{(S85)}\\
\leq\ & 2\lambda_{\max}(\boldsymbol{\Sigma}_{\mathbf{X}})\cdot\|\boldsymbol{\beta}_l^* - \tilde{\boldsymbol{\beta}}_l\|_2 + s\epsilon && \text{(S86)}\\
\leq\ & \sqrt{2\{1-\cos\Theta(\boldsymbol{\beta}_l^*, \tilde{\boldsymbol{\beta}}_l)\}} + s\epsilon \leq \sqrt{2}|\sin\Theta(\boldsymbol{\beta}_l^*, \tilde{\boldsymbol{\beta}}_l)| + s\epsilon \leq Cs\epsilon, && \text{(S87)}
\end{aligned}
$$

which also implies that $\dfrac{1}{\sqrt{\tilde{\boldsymbol{\beta}}_l^{\mathrm{T}}\widehat{\boldsymbol{\Sigma}}_{\mathbf{X}}\tilde{\boldsymbol{\beta}}_l}} \leq \dfrac{1}{\boldsymbol{\beta}_l^{*}\boldsymbol{\Sigma}_{\mathbf{X}}\boldsymbol{\beta}_l^{*} - Cs\epsilon} \leq \dfrac{2}{\lambda_{\min}(\boldsymbol{\Sigma}_{\mathbf{X}})}$ if $s\epsilon \leq \dfrac{\lambda_{\min(\boldsymbol{\Sigma}_{\mathbf{X}})}}{2}$. Finally, by (S65) we have the desired conclusion. □

Now we define $cr_k(s) = \inf_{F:|F|\leq s} cr(\mathbf{M}_{k,F}, \boldsymbol{\Sigma}_F)$. We have the following lemma.

Lemma S14. *Assume that* $|\sin\Theta(\hat{\boldsymbol{\beta}}_l, \boldsymbol{\beta}_l)| < Cs\epsilon$ *for any* $l < k$. *If* $\dfrac{2s\epsilon}{\lambda_{\min}(\boldsymbol{\Sigma}_{\mathbf{X}})} + \dfrac{2s\epsilon}{\lambda_{\min}(\boldsymbol{\Sigma}_{\mathbf{X}})\lambda_1} < \min\{\frac{1}{2}, \dfrac{\Delta}{\lambda_1+\lambda_2}, \dfrac{\lambda_{\min}(\boldsymbol{\Sigma}_{\mathbf{X}})}{2}, \dfrac{\Delta}{2(1+\lambda_1^2)}cr_k(s)\}$, *we have that* $\sin\Theta(\hat{\boldsymbol{\beta}}_k, \boldsymbol{\beta}_k) \leq Cs\epsilon$.

Proof of Lemma S14. Combine Lemma S13 with Lemma S2 and we have the desired conclusion. □

Proof of Theorem 3. Combining Lemma S12 with Lemma S14, we have that if $\|\widehat{\mathbf{M}} - \mathbf{M}\|_{\max} \leq \epsilon$, $\|\widehat{\boldsymbol{\Sigma}}_{\mathbf{X}} - \boldsymbol{\Sigma}_{\mathbf{X}}\|_{\max} \leq \epsilon$, then $|\sin\Theta(\hat{\boldsymbol{\beta}}_k, \boldsymbol{\beta}_k)| \leq Cs\epsilon, k = 1, \ldots, K$. By Theorem 1 we have the desired conclusion. □

References

Cruz-Cano, R. & Lee, M.-L. T. (2014), 'Fast regularized canonical correlation analysis', *Computational Statistics & Data Analysis* **70**, 88–100.

Huo, X. & Székely, G. J. (2016), 'Fast computing for distance covariance', *Technometrics* **58**(4), 435–447.

Mai, Q. & Zhang, X. (2019), 'An iterative penalized least squares approach to sparse canonical correlation analysis', *Biometrics* **75**(3), 734–744.

Serfling, R. J. (1980), *Approximation Theorems of Mathematical Statistics*, Wiley Series in Probability and Mathematical Statistics.

Székely, G. J., Rizzo, M. L. & Bakirov, N. K. (2007), 'Measuring and testing dependence by correlation of distances', *Annals of Statistics* **35**, 2769–2794.

Tan, K. M., Wang, Z., Liu, H. & Zhang, T. (2018), 'Sparse Generalized Eigenvalue Problem: Optimal Statistical Rates via Truncated Rayleigh Flow', *Journal of the Royal Statistical Society Series B: Statistical Methodology* **80**(5), 1057–1086.

Vershynin, R. (2018), *High-dimensional probability: An introduction with applications in data science*, Vol. 47, Cambridge university press.

Yuan, X.-T. & Zhang, T. (2013), 'Truncated power method for sparse eigenvalue problems', *Journal of Machine Learning Research* **14**(Apr), 899–925.