Statistical Inference for High-Dimensional Time Dependent Linear Models with Knowledge Transfer

Zongqi Liu[1], Shengji Jia[2], and Xiao Guo[1]

[1] *University of Science and Technology of China and*

[2] *Shanghai Lixin University of Accounting and Finance*

Supplementary Material

Additional numerical results is presented in Section S1. Section S2 outlines the selection principles for the tuning parameters. The upper bounds established in Theorem 1 are rigorously proved in Section S3. The asymptotic normality of the transfer learning debiased estimator, developed in Section 2.3, is formally demonstrated in Section S4. Moreover, the theoretical properties of the banded estimator for the error covariance matrix, proposed in Section 3.2, are established in Section S5. Finally, the results in Theorem 4, presented in Section 4, are proved in Section S6.

S1 Numerical results

This section provides additional numerical results on the performance of transfer learning under time series settings.

First, as shown in Tables 1 and 2 in the main text, for signal features, the debiased SD Trans-Lasso method shows a clear advantage over the method of Tian and Feng (2023) under both setting (a) and (b), however for noise features, the advantage becomes less pronounced under setting (b). The phenomenon can be largely attributed to the relatively small residual bias of the debiased estimators in Tian and Feng (2023) for zero coefficients.

The method in Tian and Feng (2023) constructs debiased estimators using a nodewise Lasso procedure. However, unlike Yuan and Guo (2022), it fails to account for temporal dependence in the construction of confidence intervals. This limitation primarily affects the length of the intervals rather than the point estimates. Consequently, when the bias and variance are already small (as is the case for zero coefficients), properly modeling temporal dependence has limited impact on coverage probabilities. This observation is consistent with Li (2020), who shows that debiased estimators for nonzero coefficients typically exhibit larger residual bias than those for zero coefficients.

We further examine this issue by presenting, under both settings (a) and (b), the mean of the debiased estimators(Mean), the mean bias (Bias), the mean absolute bias (MAB), and the variance (Var) of the debiased estimators computed over 600 simulation replications. Note that for zero coefficients, the mean of the debiased estimator equals its mean bias by definition, so only the mean bias is reported

As shown in Tables 2 and 4, for zero coefficients, the debiased estimators from Tian and Feng (2023) exhibit notably smaller mean absolute bias and variance in setting (b) than in setting (a). This indicates that the residual bias and variance are already quite small in setting (b), leaving little room for improvement from accounting for temporal dependence. In contrast, for signal features (Tables 1 and 3), the point estimates from Tian and Feng (2023) exhibit non-negligible bias, and the failure to account for temporal dependence leads to narrower confidence intervals, resulting in noticeable undercoverage. This explains why the advantage of our method is more pronounced for signal features than for noise features.

Moreover, Tables 1 and 3 indicate that, for signal features, the residual bias of our debiased estimator is smaller than that of the single-task method in Yuan and Guo (2022),

Table 1: Estimation Performance for $\beta_8^* = 0.5$ in Setting (a)

Setup		Yuan and Guo (2022)				Debiased SD Trans-Lasso				Tian and Feng (2023)			
h	$\|A\|$	Mean	Bias	Var	MAB	Mean	Bias	Var	MAB	Mean	Bias	Var	MAB
1	4	0.472	-0.028	0.011	0.088	0.492	-0.008	0.008	0.074	0.493	-0.007	0.007	0.067
1	8	0.478	-0.022	0.011	0.086	0.497	-0.003	0.008	0.073	0.474	-0.026	0.007	0.070
1	12	0.482	-0.018	0.010	0.080	0.499	-0.001	0.007	0.067	0.492	-0.008	0.005	0.057
1	16	0.469	-0.031	0.011	0.088	0.492	-0.008	0.008	0.074	0.477	-0.023	0.006	0.065
1	20	0.486	-0.014	0.010	0.080	0.502	0.002	0.007	0.069	0.490	-0.010	0.005	0.058
2	4	0.482	-0.018	0.011	0.083	0.495	-0.005	0.009	0.074	0.480	-0.020	0.007	0.069
2	8	0.480	-0.020	0.010	0.083	0.493	-0.007	0.007	0.069	0.476	-0.024	0.006	0.067
2	12	0.475	-0.025	0.011	0.086	0.499	-0.001	0.008	0.069	0.478	-0.022	0.006	0.064
2	16	0.474	-0.026	0.010	0.081	0.494	-0.006	0.007	0.068	0.485	-0.015	0.005	0.060
2	20	0.474	-0.026	0.010	0.084	0.495	-0.005	0.008	0.070	0.485	-0.015	0.005	0.061
3	4	0.477	-0.023	0.009	0.080	0.498	-0.002	0.008	0.071	0.500	0.000	0.006	0.060
3	8	0.470	-0.030	0.010	0.085	0.484	-0.016	0.008	0.072	0.469	-0.031	0.006	0.066
3	12	0.483	-0.017	0.010	0.080	0.499	-0.001	0.007	0.068	0.482	-0.018	0.006	0.060
3	16	0.480	-0.020	0.010	0.083	0.496	-0.004	0.008	0.072	0.476	-0.024	0.006	0.066
3	20	0.477	-0.023	0.011	0.084	0.492	-0.008	0.009	0.075	0.474	-0.026	0.006	0.066

which helps explain the improved confidence interval coverage for nonzero regression coefficients achieved by our method.

To further evaluate the finite-sample performance of the proposed transferable source detection algorithm, we conduct an additional simulation study. Specifically, we fix the total number of auxiliary sources at 6 and vary the number of informative sources $|\mathcal{A}_h| \in \{0, 1, 2, 3, 4, 5, 6\}$, where $\mathcal{A}_h = \{1, \ldots, |\mathcal{A}_h|\}$. Both the target and auxiliary sample sizes are set to 300. For the target study, the coefficient vector is specified as

$$\boldsymbol{\beta}^* = (\underbrace{0.5, \ldots, 0.5}_{10}, \underbrace{0, \ldots, 0}_{p-10})^\top.$$

For the auxiliary regression coefficients, if $k \in \mathcal{A}_h$, the heterogeneous index set is $H_k =$

Table 2: Estimation Performance for $\beta^*_{16} = 0$ in Setting (a)

Setup		Yuan and Guo (2022)			Debiased SD Trans-Lasso			Tian and Feng (2023)		
h	$\lvert A\rvert$	Bias	Var	MAB	Bias	Var	MAB	Bias	Var	MAB
1	4	-0.001	0.008	0.069	-0.001	0.008	0.070	-0.002	0.005	0.053
1	8	0.005	0.008	0.071	0.000	0.007	0.069	0.004	0.005	0.053
1	12	-0.003	0.008	0.073	-0.001	0.008	0.073	-0.002	0.005	0.057
1	16	-0.002	0.009	0.073	0.000	0.008	0.071	-0.003	0.005	0.055
1	20	0.001	0.007	0.068	0.000	0.008	0.070	0.000	0.005	0.053
2	4	-0.001	0.008	0.070	0.001	0.007	0.069	-0.002	0.005	0.054
2	8	0.001	0.007	0.065	0.000	0.007	0.067	0.002	0.004	0.049
2	12	0.003	0.008	0.070	-0.001	0.008	0.069	0.000	0.005	0.054
2	16	-0.003	0.008	0.070	-0.006	0.007	0.069	-0.005	0.005	0.054
2	20	-0.003	0.008	0.069	-0.004	0.007	0.069	-0.004	0.005	0.053
3	4	0.002	0.008	0.070	0.001	0.008	0.073	0.008	0.006	0.059
3	8	0.001	0.007	0.068	0.002	0.007	0.069	0.001	0.004	0.050
3	12	-0.004	0.008	0.069	0.000	0.007	0.068	-0.002	0.005	0.052
3	16	0.002	0.008	0.070	0.003	0.008	0.070	0.003	0.005	0.055
3	20	0.000	0.007	0.068	0.006	0.008	0.070	0.002	0.004	0.052

$\{1, \ldots, 100\}$, and for $j = 1, \ldots, p$,

$$w_j^{(k)} = \beta_j^* + \xi_j^{(k)} \mathbb{I}(j \in H_k), \quad \xi_j^{(k)} \overset{\text{i.i.d.}}{\sim} N(0,\, h/40),$$

with $h \in \{1, 2\}$. For $k \notin \mathcal{A}_h$, we similarly set $H_k = \{1, \ldots, 100\}$ and draw

$$w_j^{(k)} = \beta_j^* + \xi_j^{(k)} \mathbb{I}(j \in H_k), \quad \xi_j^{(k)} \overset{\text{i.i.d.}}{\sim} N(0,\, 1/3), \quad j = 1, \ldots, p.$$

The error sequences $\{\epsilon_i^{(k)}\}$ and the covariate sequences $\{\boldsymbol{x}_i^{(k)}\}$ are generated in the same manner as in Section 5 of the main text. For each configuration, we conduct 300 independent replications and record: (i) the selection frequency of each auxiliary source (Freq);

Table 3: Estimation Performance for $\beta_8^* = 0.5$ in Setting (b)

Setup		Yuan and Guo (2022)				Debiased SD Trans-Lasso				Tian and Feng (2023)			
h	$\|A\|$	Mean	Bias	Var	MAB	Mean	Bias	Var	MAB	Mean	Bias	Var	MAB
1	4	0.484	-0.016	0.015	0.097	0.486	-0.014	0.011	0.084	0.474	-0.026	0.006	0.065
1	8	0.486	-0.014	0.014	0.096	0.488	-0.012	0.010	0.081	0.449	-0.051	0.004	0.070
1	12	0.491	-0.009	0.015	0.099	0.495	-0.005	0.011	0.083	0.473	-0.027	0.004	0.053
1	16	0.494	-0.006	0.014	0.095	0.497	-0.003	0.010	0.079	0.469	-0.031	0.003	0.051
1	20	0.484	-0.016	0.014	0.095	0.492	-0.008	0.011	0.083	0.467	-0.033	0.003	0.052
2	4	0.490	-0.010	0.016	0.100	0.484	-0.016	0.012	0.089	0.436	-0.064	0.007	0.087
2	8	0.495	-0.005	0.014	0.098	0.492	-0.008	0.011	0.084	0.443	-0.057	0.005	0.077
2	12	0.478	-0.022	0.015	0.101	0.487	-0.013	0.010	0.083	0.456	-0.044	0.005	0.066
2	16	0.489	-0.011	0.014	0.096	0.496	-0.004	0.010	0.080	0.475	-0.025	0.003	0.049
2	20	0.496	-0.004	0.014	0.097	0.499	-0.001	0.011	0.086	0.473	-0.027	0.003	0.051
3	4	0.485	-0.015	0.013	0.093	0.486	-0.014	0.010	0.077	0.483	-0.017	0.005	0.059
3	8	0.493	-0.007	0.014	0.093	0.487	-0.013	0.010	0.081	0.431	-0.069	0.005	0.085
3	12	0.489	-0.011	0.016	0.101	0.494	-0.006	0.011	0.086	0.462	-0.038	0.004	0.062
3	16	0.493	-0.007	0.013	0.087	0.493	-0.007	0.010	0.080	0.452	-0.048	0.004	0.066
3	20	0.489	-0.011	0.016	0.100	0.491	-0.009	0.012	0.085	0.451	-0.049	0.004	0.072

(ii) the average ℓ_1 distance between the estimated regression coefficients of each auxiliary source and the target (Mean ℓ_1); and (iii) the average value of the screening index for each auxiliary source (Mean Metric).

As shown in Tables 5–8, the algorithm selects informative auxiliary sources with high frequency, and the Mean Metric values for these sources are substantially smaller than those for non-informative sources. This empirical separation indicates that the proposed index provides a reliable finite-sample diagnostic for assessing the suitability of auxiliary datasets prior to transfer learning.

In addition to the data generation process studied in Section 5, the following setting

Table 4: Estimation Performance for $\beta^*_{16} = 0$ in Setting (b)

Setup		Yuan and Guo (2022)			Debiased SD Trans-Lasso			Tian and Feng (2023)		
h	$\lvert A\rvert$	Bias	Var	MAB	Bias	Var	MAB	Bias	Var	MAB
1	4	0.012	0.011	0.083	-0.001	0.010	0.079	-0.001	0.002	0.033
1	8	0.011	0.011	0.085	0.002	0.011	0.081	-0.001	0.002	0.034
1	12	0.015	0.012	0.089	0.006	0.012	0.086	0.003	0.003	0.036
1	16	0.013	0.011	0.083	0.001	0.010	0.081	0.000	0.002	0.033
1	20	0.006	0.011	0.082	-0.001	0.010	0.078	-0.001	0.002	0.033
2	4	0.009	0.012	0.087	-0.003	0.011	0.082	0.004	0.003	0.036
2	8	0.006	0.012	0.086	-0.002	0.010	0.080	-0.001	0.002	0.034
2	12	0.010	0.011	0.084	0.002	0.010	0.079	0.001	0.002	0.032
2	16	0.007	0.012	0.087	-0.003	0.011	0.086	-0.003	0.002	0.034
2	20	0.015	0.013	0.088	0.003	0.012	0.086	0.000	0.003	0.036
3	4	0.009	0.012	0.084	-0.003	0.011	0.082	0.004	0.003	0.035
3	8	0.012	0.012	0.084	0.001	0.010	0.082	0.000	0.002	0.033
3	12	0.017	0.012	0.089	0.004	0.011	0.081	0.003	0.002	0.034
3	16	0.014	0.012	0.090	0.005	0.011	0.085	0.002	0.003	0.036
3	20	0.016	0.012	0.088	0.008	0.011	0.084	0.004	0.003	0.036

is also considered: For each $k \in \{0, 1, \ldots, M\}$, we first independently generate $\boldsymbol{z}_i^{(k)} \sim N(\mathbf{0}, \widetilde{\boldsymbol{\Sigma}}^{(k)})$, and then construct the covariates as $\boldsymbol{x}_i^{(k)} = \boldsymbol{z}_i^{(k)} + 0.3\,\boldsymbol{z}_{i-1}^{(k)} + 0.5\,\boldsymbol{z}_{i-2}^{(k)}$. The error process follows an MA(3) model across all datasets:

$$\epsilon_i^{(k)} = \xi_i^{(k)} + 0.5\xi_{i-1}^{(k)} + 0.3\xi_{i-2}^{(k)} + 1.05\xi_{i-3}^{(k)}, \quad \xi_i^{(k)} \overset{\text{i.i.d.}}{\sim} N(0,1), \quad i = 1, 2, \ldots, n_k.$$

All other data generation settings remain the same as in the main text.

Under MA dependence setting, the performance of our transfer learning method and the benchmark approaches is largely consistent with that observed in the main text. The

Table 5: Selection accuracy performance under setting (a) with $h = 1$.

$\lvert\mathcal{A}_h\rvert$	Source (Informative)	Freq	Mean ℓ_1	Mean Metric
0	1 (−)	0.000	21.818	8.282
	2 (−)	0.007	16.115	4.016
	3 (−)	0.000	20.661	6.666
	4 (−)	0.000	22.709	7.432
	5 (−)	0.000	19.593	5.524
	6 (−)	0.000	19.430	6.505
1	1 (+)	0.957	4.646	−0.552
	2 (−)	0.000	22.382	7.398
	3 (−)	0.000	21.346	6.411
	4 (−)	0.000	23.305	7.819
	5 (−)	0.000	20.539	6.200
	6 (−)	0.003	24.349	8.013
2	1 (+)	0.973	4.612	−0.513
	2 (+)	1.000	4.245	−0.739
	3 (−)	0.000	20.745	6.847
	4 (−)	0.000	21.460	6.306
	5 (−)	0.000	21.015	7.062
	6 (−)	0.003	20.214	6.001
3	1 (+)	0.953	4.757	−0.524
	2 (+)	0.997	4.086	−0.791
	3 (+)	0.993	4.094	−0.804
	4 (−)	0.000	24.489	8.131
	5 (−)	0.003	19.095	5.175
	6 (−)	0.000	24.633	7.747
4	1 (+)	0.983	4.702	−0.567
	2 (+)	0.997	4.260	−0.793
	3 (+)	0.990	4.320	−0.817
	4 (+)	0.993	4.095	−0.821
	5 (−)	0.000	20.083	5.237
	6 (−)	0.000	21.455	5.758
5	1 (+)	0.967	4.888	−0.547
	2 (+)	1.000	4.206	−0.837
	3 (+)	0.993	4.338	−0.803
	4 (+)	0.987	4.204	−0.838
	5 (+)	1.000	4.263	−0.854
	6 (−)	0.003	21.358	5.761
6	1 (+)	0.960	4.755	−0.487
	2 (+)	0.990	4.365	−0.724
	3 (+)	0.997	4.084	−0.788
	4 (+)	1.000	4.181	−0.798
	5 (+)	0.993	4.056	−0.797
	6 (+)	0.997	4.201	−0.757

Note: "(+)" and "(−)" denote informative and non-informative sources, respectively.

Table 6: Selection accuracy performance under setting (a) with $h = 2$.

$\lvert\mathcal{A}_h\rvert$	Source (Informative)	Freq	Mean ℓ_1	Mean Metric
0	1 (−)	0.003	19.324	5.388
	2 (−)	0.000	19.486	6.326
	3 (−)	0.000	19.951	6.601
	4 (−)	0.000	20.292	5.942
	5 (−)	0.017	16.321	3.773
	6 (−)	0.003	20.429	6.491
1	1 (+)	0.943	5.120	−0.417
	2 (−)	0.000	19.138	6.814
	3 (−)	0.000	21.179	7.013
	4 (−)	0.000	19.728	5.371
	5 (−)	0.000	20.372	6.245
	6 (−)	0.000	23.133	7.462
2	1 (+)	0.947	4.962	−0.459
	2 (+)	0.983	4.314	−0.723
	3 (−)	0.003	17.448	5.043
	4 (−)	0.000	24.078	9.342
	5 (−)	0.003	18.879	5.433
	6 (−)	0.000	19.445	5.346
3	1 (+)	0.977	5.045	−0.474
	2 (+)	0.997	4.649	−0.673
	3 (+)	0.997	4.673	−0.703
	4 (−)	0.003	20.162	6.284
	5 (−)	0.000	23.283	8.077
	6 (−)	0.000	22.077	6.492
4	1 (+)	0.943	5.085	−0.383
	2 (+)	0.987	4.580	−0.706
	3 (+)	0.993	4.786	−0.676
	4 (+)	0.990	4.267	−0.686
	5 (−)	0.000	19.540	5.397
	6 (−)	0.003	20.922	6.926
5	1 (+)	0.943	5.339	−0.404
	2 (+)	0.987	4.393	−0.626
	3 (+)	0.987	4.508	−0.664
	4 (+)	0.987	4.485	−0.639
	5 (+)	0.987	4.585	−0.636
	6 (−)	0.000	21.886	6.795
6	1 (+)	0.967	4.960	−0.466
	2 (+)	0.997	4.275	−0.712
	3 (+)	0.990	4.419	−0.617
	4 (+)	0.983	4.401	−0.706
	5 (+)	0.987	4.504	−0.714
	6 (+)	0.987	4.722	−0.632

Note: "(+)" and "(−)" denote informative and non-informative sources, respectively.

Table 7: Selection accuracy performance under setting (b) with $h = 1$.

$\lvert\mathcal{A}_h\rvert$	Source (Informative)	Freq	Mean ℓ_1	Mean Metric
0	1 (−)	0.000	27.135	6.774
	2 (−)	0.067	18.757	2.854
	3 (−)	0.003	23.578	5.204
	4 (−)	0.000	25.053	6.037
	5 (−)	0.007	21.610	4.724
	6 (−)	0.000	20.624	10.097
1	1 (+)	1.000	4.650	−0.636
	2 (−)	0.000	26.932	6.858
	3 (−)	0.000	23.801	7.758
	4 (−)	0.000	27.981	14.090
	5 (−)	0.000	23.629	8.489
	6 (−)	0.000	27.608	19.492
2	1 (+)	0.990	4.503	−0.557
	2 (+)	0.997	4.618	−0.675
	3 (−)	0.000	25.388	5.703
	4 (−)	0.007	24.109	4.875
	5 (−)	0.000	23.167	22.085
	6 (−)	0.003	22.566	5.886
3	1 (+)	0.997	4.739	−0.577
	2 (+)	0.997	4.713	−0.695
	3 (+)	1.000	4.672	−0.690
	4 (−)	0.000	28.007	6.540
	5 (−)	0.000	21.333	8.699
	6 (−)	0.000	26.680	21.621
4	1 (+)	1.000	4.601	−0.653
	2 (+)	0.997	4.624	−0.653
	3 (+)	1.000	4.560	−0.630
	4 (+)	0.997	4.656	−0.646
	5 (−)	0.003	22.017	4.492
	6 (−)	0.003	23.970	4.983
5	1 (+)	0.987	4.582	−0.568
	2 (+)	1.000	4.524	−0.712
	3 (+)	0.993	4.724	−0.712
	4 (+)	1.000	4.633	−0.681
	5 (+)	1.000	4.535	−0.680
	6 (−)	0.000	22.868	4.229
6	1 (+)	0.990	4.584	−0.582
	2 (+)	1.000	4.554	−0.683
	3 (+)	0.997	4.490	−0.618
	4 (+)	0.993	4.630	−0.687
	5 (+)	1.000	4.647	−0.683
	6 (+)	1.000	4.607	−0.659

Note: "(+)" and "(−)" denote informative and non-informative sources, respectively.

Table 8: Selection accuracy performance under setting (b) with $h = 2$.

$\lvert\mathcal{A}_h\rvert$	Source (Informative)	Freq	Mean ℓ_1	Mean Metric
0	1 (−)	0.000	22.085	10.607
	2 (−)	0.000	24.430	5.780
	3 (−)	0.000	23.388	7.137
	4 (−)	0.003	23.312	4.558
	5 (−)	0.000	19.071	8.265
	6 (−)	0.000	22.306	8.375
1	1 (+)	0.977	5.265	−0.500
	2 (−)	0.000	23.913	5.414
	3 (−)	0.000	24.239	5.544
	4 (−)	0.003	22.012	4.045
	5 (−)	0.000	22.465	5.711
	6 (−)	0.000	26.102	5.784
2	1 (+)	1.000	5.134	−0.569
	2 (+)	0.973	5.014	−0.430
	3 (−)	0.007	20.550	3.896
	4 (−)	0.000	27.458	13.851
	5 (−)	0.007	21.656	4.021
	6 (−)	0.000	21.713	8.926
3	1 (+)	0.987	5.257	−0.579
	2 (+)	0.970	5.161	−0.374
	3 (+)	0.993	5.163	−0.606
	4 (−)	0.000	22.719	8.155
	5 (−)	0.000	25.213	9.106
	6 (−)	0.000	23.912	6.028
4	1 (+)	0.973	5.123	−0.480
	2 (+)	0.997	4.975	−0.639
	3 (+)	0.997	5.419	−0.616
	4 (+)	1.000	5.172	−0.595
	5 (−)	0.000	22.577	5.219
	6 (−)	0.000	24.194	5.877
5	1 (+)	0.947	5.583	−0.435
	2 (+)	0.973	4.971	−0.366
	3 (+)	0.990	5.067	−0.609
	4 (+)	0.960	5.074	−0.395
	5 (+)	1.000	5.103	−0.612
	6 (−)	0.000	23.952	5.602
6	1 (+)	0.997	4.955	−0.613
	2 (+)	0.983	4.737	−0.540
	3 (+)	0.953	4.997	−0.256
	4 (+)	1.000	5.107	−0.591
	5 (+)	0.987	4.993	−0.503
	6 (+)	0.983	5.038	−0.564

Note: "(+)" and "(−)" denote informative and non-informative sources, respectively.

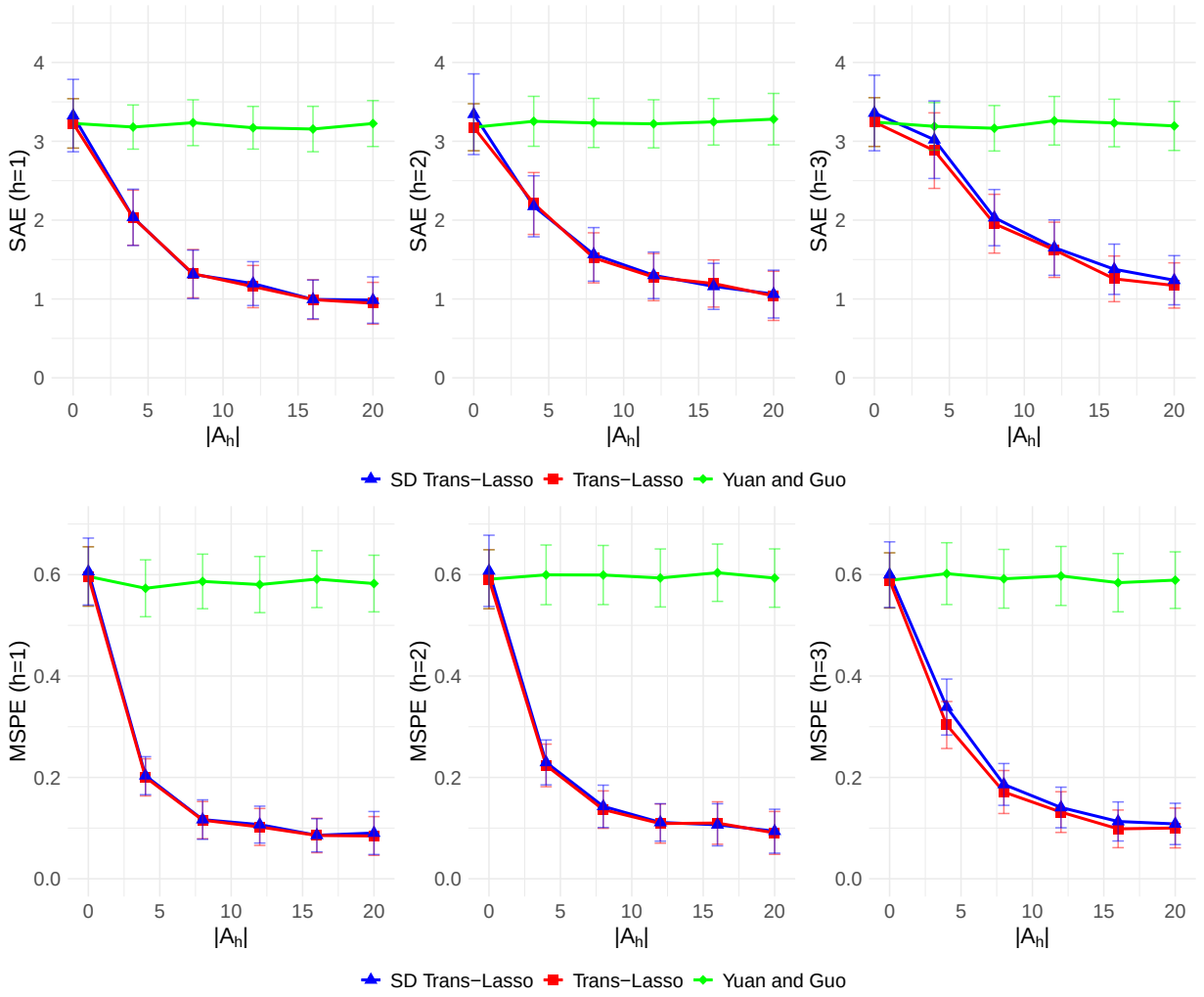

Figure 1: Estimation errors of Yuan and Guo (2022), Trans-Lasso, and SD Trans-Lasso in setting (a). The two rows correspond to SAE(i) and MSPE(ii) respectively. Error bars denote standard deviations divided by 2.5.

only notable exception occurs under setting (b) when $\beta_j^* = 0$, where the method proposed by Tian and Feng (2023) also achieves reliable confidence interval coverage. Compared with the results in the main text, this improvement is likely attributable to the weaker temporal dependence induced by the MA data-generating process, which leads to enhanced performance of the method proposed by Tian and Feng (2023).

Furthermore, as shown in Figures 3, 4 and 5, the SD Trans-Lasso method demon-

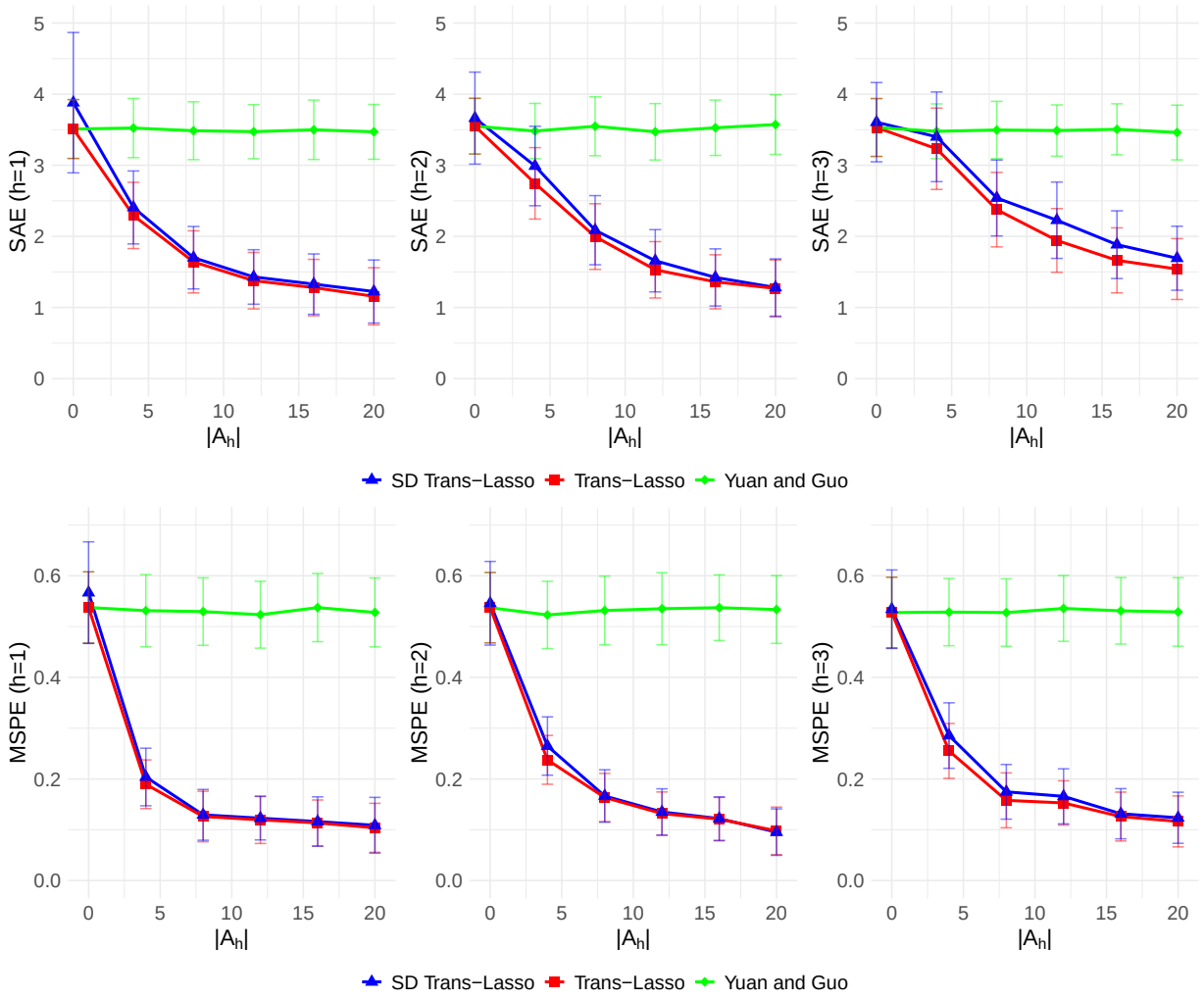

Figure 2: Estimation errors of Yuan and Guo (2022), Trans-Lasso, and SD Trans-Lasso in setting (b). The two rows correspond to SAE(i) and MSPE(ii) respectively. Error bars denote standard deviations divided by 2.5.

strates consistent performance, effectively capturing the temporal dynamics in unemployment rates. These results further demonstrate the advantages of our transfer learning method in the presence of dependence.

Table 9: Average coverage probabilities for $\beta_8^* = 0.5$ and $\beta_{16}^* = 0$ in setting (a), with average confidence interval lengths in brackets.

h	$\lvert\mathcal{A}_h\rvert$	Yuan and Guo (2022)		Debiased SD Trans-Lasso		Tian and Feng (2023)	
		β_8^*	β_{16}^*	β_8^*	β_{16}^*	β_8^*	β_{16}^*
1	4	0.907 (0.401)	0.942 (0.405)	0.937 (0.343)	0.957 (0.342)	0.868 (0.220)	0.958 (0.221)
1	8	0.933 (0.407)	0.938 (0.412)	0.938 (0.341)	0.938 (0.342)	0.857 (0.221)	0.952 (0.223)
1	12	0.922 (0.414)	0.937 (0.402)	0.943 (0.340)	0.947 (0.339)	0.930 (0.221)	0.957 (0.221)
1	16	0.910 (0.404)	0.952 (0.406)	0.945 (0.341)	0.935 (0.341)	0.912 (0.221)	0.943 (0.221)
1	20	0.925 (0.405)	0.952 (0.410)	0.943 (0.338)	0.962 (0.339)	0.933 (0.221)	0.962 (0.221)
2	4	0.938 (0.412)	0.922 (0.406)	0.940 (0.346)	0.943 (0.345)	0.827 (0.221)	0.960 (0.221)
2	8	0.927 (0.405)	0.917 (0.413)	0.933 (0.343)	0.945 (0.343)	0.865 (0.221)	0.943 (0.221)
2	12	0.927 (0.408)	0.932 (0.409)	0.960 (0.340)	0.962 (0.339)	0.895 (0.221)	0.962 (0.222)
2	16	0.943 (0.412)	0.943 (0.402)	0.955 (0.342)	0.942 (0.339)	0.940 (0.221)	0.955 (0.220)
2	20	0.925 (0.405)	0.937 (0.411)	0.933 (0.339)	0.945 (0.340)	0.935 (0.220)	0.940 (0.222)
3	4	0.902 (0.407)	0.952 (0.406)	0.928 (0.343)	0.957 (0.345)	0.870 (0.221)	0.928 (0.221)
3	8	0.928 (0.410)	0.923 (0.406)	0.942 (0.346)	0.947 (0.346)	0.847 (0.222)	0.948 (0.221)
3	12	0.917 (0.409)	0.945 (0.408)	0.932 (0.341)	0.945 (0.342)	0.907 (0.221)	0.948 (0.221)
3	16	0.942 (0.411)	0.943 (0.402)	0.952 (0.340)	0.957 (0.339)	0.885 (0.220)	0.958 (0.221)
3	20	0.907 (0.405)	0.932 (0.411)	0.933 (0.340)	0.940 (0.342)	0.893 (0.221)	0.945 (0.221)

S2 Tuning Parameter Selection

This section details the selection procedures for all tuning parameters, including those associated with the Lasso estimator, the debiased Lasso estimator, the auxiliary sample screening step, and the banding parameter in autocovariance matrix estimation.

First, we employ ten-fold cross-validation to select the tuning parameters λ_w, λ_δ, and

Table 10: Average coverage probabilities for $\beta_8^* = 0.5$ and $\beta_{16}^* = 0$ in setting (b), with average confidence interval lengths in brackets.

h	$\lvert\mathcal{A}_h\rvert$	Yuan and Guo (2022)		Debiased SD Trans-Lasso		Tian and Feng (2023)	
		β_8^*	β_{16}^*	β_8^*	β_{16}^*	β_8^*	β_{16}^*
1	4	0.907 (0.493)	0.938 (0.493)	0.930 (0.400)	0.958 (0.400)	0.752 (0.184)	0.955 (0.184)
1	8	0.942 (0.490)	0.913 (0.488)	0.957 (0.397)	0.935 (0.397)	0.752 (0.185)	0.942 (0.183)
1	12	0.940 (0.487)	0.918 (0.490)	0.945 (0.392)	0.937 (0.393)	0.915 (0.186)	0.952 (0.184)
1	16	0.907 (0.485)	0.935 (0.496)	0.915 (0.393)	0.938 (0.393)	0.880 (0.185)	0.960 (0.184)
1	20	0.917 (0.482)	0.938 (0.491)	0.933 (0.392)	0.930 (0.392)	0.903 (0.185)	0.952 (0.184)
2	4	0.913 (0.487)	0.940 (0.491)	0.927 (0.398)	0.952 (0.397)	0.587 (0.185)	0.948 (0.184)
2	8	0.937 (0.489)	0.943 (0.483)	0.947 (0.397)	0.952 (0.394)	0.643 (0.184)	0.957 (0.184)
2	12	0.935 (0.488)	0.937 (0.491)	0.940 (0.396)	0.942 (0.397)	0.848 (0.185)	0.960 (0.184)
2	16	0.932 (0.484)	0.952 (0.483)	0.952 (0.391)	0.953 (0.388)	0.930 (0.185)	0.970 (0.184)
2	20	0.920 (0.490)	0.937 (0.490)	0.945 (0.394)	0.938 (0.394)	0.920 (0.185)	0.945 (0.184)
3	4	0.927 (0.486)	0.935 (0.496)	0.938 (0.400)	0.955 (0.400)	0.793 (0.184)	0.942 (0.185)
3	8	0.940 (0.491)	0.948 (0.490)	0.957 (0.401)	0.948 (0.401)	0.480 (0.185)	0.953 (0.184)
3	12	0.928 (0.487)	0.923 (0.486)	0.947 (0.399)	0.933 (0.398)	0.840 (0.185)	0.965 (0.184)
3	16	0.928 (0.489)	0.943 (0.498)	0.932 (0.395)	0.943 (0.397)	0.785 (0.186)	0.958 (0.185)
3	20	0.937 (0.481)	0.943 (0.488)	0.948 (0.394)	0.945 (0.394)	0.768 (0.184)	0.963 (0.184)

λ_k.

For the optimization problem (2.6), we first fix $r = 1/5$ for both the simulation studies and the subsequent real data analysis. Numerical results show that the inference performance is largely insensitive to c_γ; accordingly, we set $c_\gamma = 5$ in the simulations and $c_\gamma = 8$ in the real data analysis. For the parameter $c_\Theta \lambda_\delta$, we adopt an adaptive selection strategy

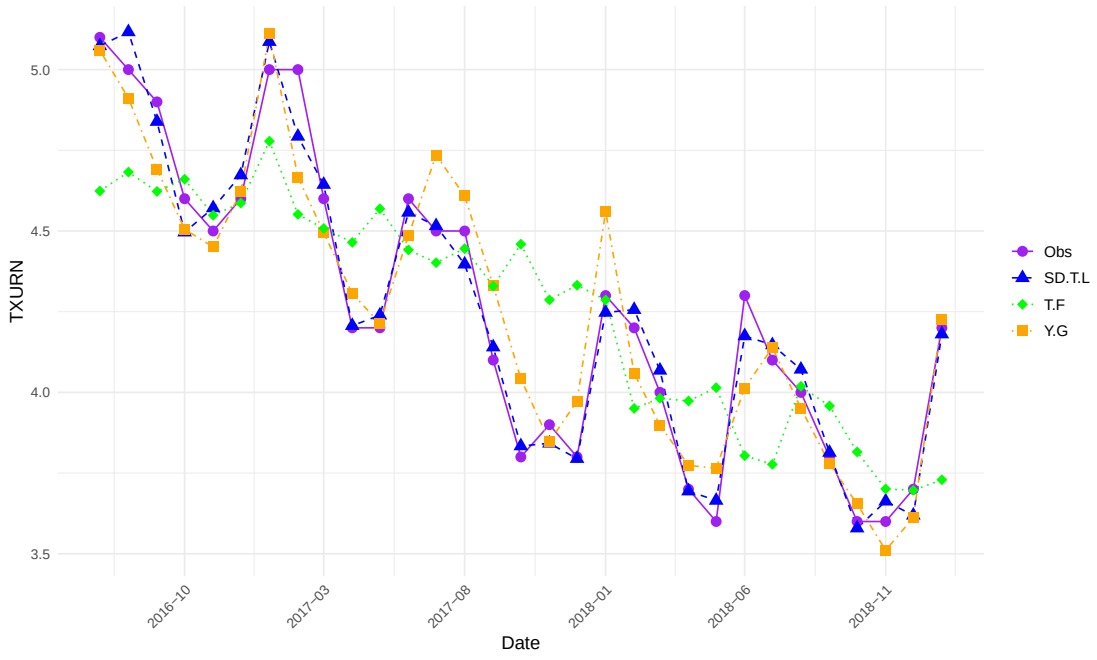

Figure 3: Unemployment rate in Texas: observed and fitted values from 2016 to 2019. Obs and SD.T.L denote the observed values and the SD Trans-Lasso method proposed in this study, respectively. T.F refers to the method of Tian and Feng (2023), while Y.G corresponds to the approach of Yuan and Guo (2022).

similar to that of Javanmard and Montanari (2014). Its initial value is set as

$$c_\Theta \lambda_\delta = \frac{1}{2\sqrt{n_0}} \Phi^{-1}\Big(1 - \frac{0.1}{p^2}\Big).$$

During the iterations, $c_\Theta \lambda_\delta$ is adjusted dynamically: it is increased if the optimization fails to converge or the constraints are violated, and decreased if convergence is achieved and the constraints are satisfied, until the direction of adjustment changes from the previous step or a maximum of 30 iterations is reached. For the threshold parameters C and C_1 in Algorithm 2, we first follow Tian and Feng (2023) by setting $C_1 = 0.1$. The parameter C is selected from the grid $C_{\text{grid}} = \{1.0, 1.2, 1.4, 1.6, 1.8, 2.0\}$ by minimizing the mean squared error (MSE) on the target training data. Specifically, for each $C \in C_{\text{grid}}$, we obtain the corresponding $\widehat{\mathcal{A}}_h$ via Algorithm 2, and then utilize the auxiliary datasets within $\widehat{\mathcal{A}}_h$ for

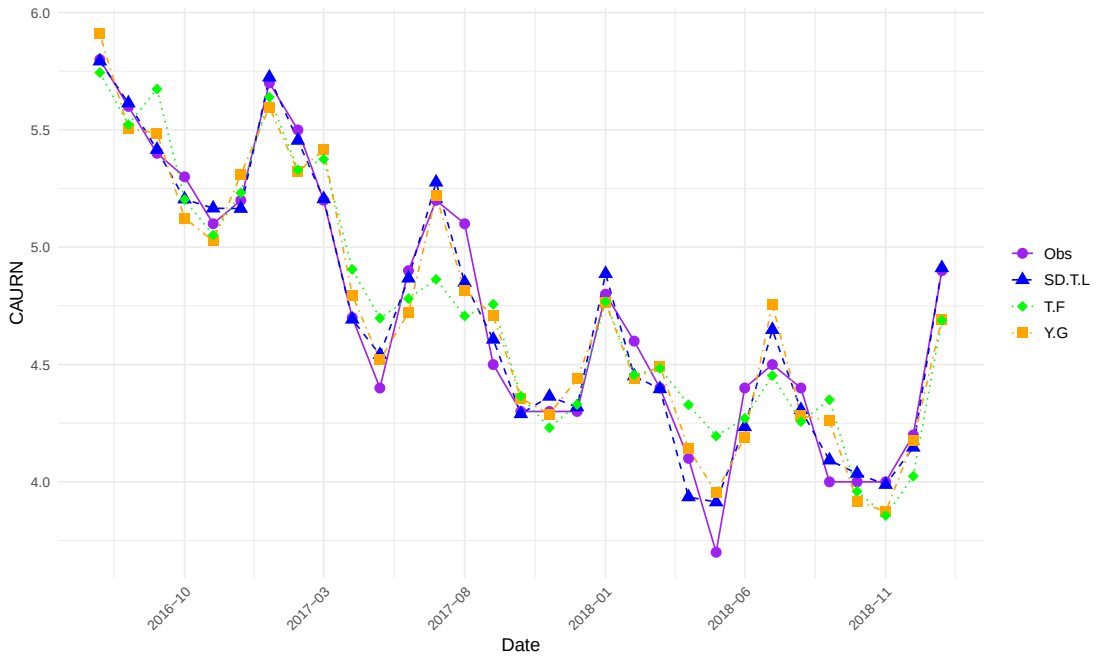

Figure 4: Unemployment Rate in California : Observed and Fitted Values from 2016 to 2019.

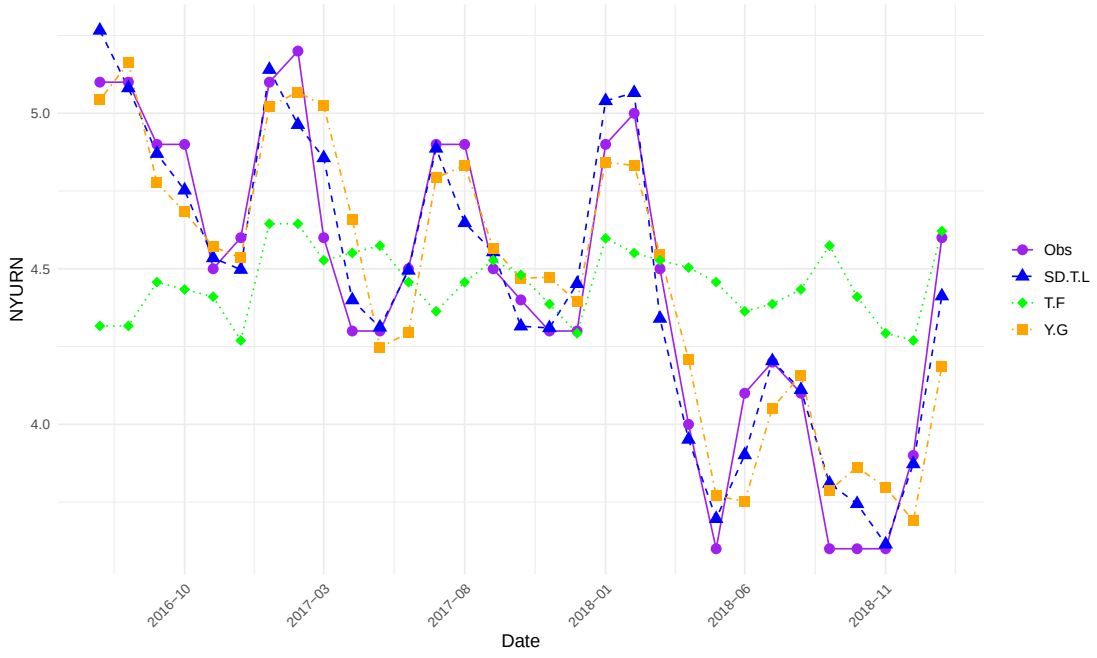

Figure 5: Unemployment Rate in New York : Observed and Fitted Values from 2016 to 2019.

transfer learning to compute the corresponding $\widehat{\boldsymbol{\beta}}$. For each candidate C, the MSE is calculated as $|\boldsymbol{y}^{(0)} - X^{(0)}\widehat{\boldsymbol{\beta}}|_2^2/n_0$. The selected C is the value in C_{grid} that minimizes this MSE. To reduce computational cost, we set $C = 2$ as specified in Section 5.2, which is the same default value adopted by Tian and Feng (2023).

For an $n \times p$ matrix $A = (a_{ij})_{i\le n, j\le p}$, we further introduce the ℓ_1 and ℓ_∞ norms defined by $|A|_1 = \max_{j\le p}\sum_{i=1}^{n}|a_{ij}|$ and $|A|_\infty = \max_{i\le n}\sum_{j=1}^{p}|a_{ij}|$, respectively. Regarding the selection of the banding parameter l, although Lemma 1 of the main text provides rate conditions that guarantee the consistency of the banded estimator $\widehat{\boldsymbol{\Sigma}}^{\epsilon}_{n_0,l}$, these theoretical results do not offer practical guidance for choosing an appropriate value of l. A natural approach is to select the banding parameter by minimizing the risk

$$R(l) = \mathbb{E}|\widehat{\boldsymbol{\Sigma}}^{\epsilon}_{n_0,l} - \boldsymbol{\Sigma}^{\epsilon}_{n_0}|_1, \tag{S2.1}$$

which measures the discrepancy between the estimated and true autocovariance matrices. However, since $\boldsymbol{\Sigma}^{\epsilon}_{n_0}$ is unknown in practice, the risk $R(l)$ cannot be evaluated directly. To address this issue, we adopt the subsampling technique proposed by Politis et al. (1999) to estimate the risk in (S2.1), thereby enabling data-driven selection of the banding parameter l. The key idea of this approach is to partition the original data into ordered and overlapping blocks. The procedure is summarized in Algorithm 1, with detailed steps outlined below.

Define $\widehat{\gamma}^{\epsilon}_k$ as

$$\widehat{\gamma}^{\epsilon}_k = \frac{1}{n_0}\sum_{i=1}^{n_0-|k|} e_i e_{i+|k|}, \quad k = 0, \pm 1, \ldots, \pm(n_0-1), \tag{S2.2}$$

where $e_i = y_i^{(0)} - (\boldsymbol{x}_i^{(0)})^\top\widehat{\boldsymbol{\beta}}$ for $i = 1, \ldots, n_0$, denotes the residuals obtained from the Trans-Lasso estimation. First, we select l from $\{0, \ldots, K-1\}$ instead of $\{0, \ldots, n_0-1\}$, where K is a predetermined integer much smaller than n_0, since $\widehat{\gamma}^{\epsilon}_k$ is not a reliable estimator of γ^{ϵ}_k for large k. In order to obtain a reliable estimate of $R(l)$ in (S2.1), we replace the

unknown autocovariance matrix $\mathbf{\Sigma}^{\epsilon}_{n_0}$ with $\widehat{\mathbf{\Sigma}}^{\epsilon}_{K}$, which is the $K \times K$ upper-left submatrix of $\widehat{\mathbf{\Sigma}}^{\epsilon}_{n_0}$. The elements of $\widehat{\mathbf{\Sigma}}^{\epsilon}_{K}$ are estimated using the entire residual sequence $\{e_1, \ldots, e_{n_0}\}$ through (S2.2). Next, we divide the entire residual sequence $\{e_i\}_{i=1}^{n_0}$ into $n_0 - s + 1$ ordered and overlapping blocks, each of length s, where s is a positive integer satisfying $s > K$. For each block, we compute the $K \times K$ l-banded sample autocovariance matrix $\widehat{\mathbf{\Sigma}}^{\epsilon}_{s,l,t}$ from the t-th subsample $\{e_t, \ldots, e_{t+s-1}\}$, and substitute it in place of the banded autocovariance matrix estimator $\widehat{\mathbf{\Sigma}}^{\epsilon}_{n_0,l}$ in (S2.1), where $t = 1, \ldots, n_0 - s + 1$.

Following these steps, we estimate $R(l)$ in (S2.1) by

$$\widehat{R}(l) = \frac{1}{n_0 - s + 1} \sum_{t=1}^{n_0 - s + 1} |\widehat{\mathbf{\Sigma}}^{\epsilon}_{s,l,t} - \widehat{\mathbf{\Sigma}}^{\epsilon}_{K}|_1,$$

and then determine l via $l = \arg\min_{l \in \{0,\ldots,K-1\}} \widehat{R}(l)$. Obviously, we need to choose the block size s. As a general guideline, Politis et al. (1999) showed that s should grow to infinity with rate $n_0^{1/3}$, provided $s > K$. For computational efficiency, we use $K = 20$ and $s = 30$ in Algorithm 1 throughout both our simulation studies and real data analysis. Furthermore, we initially perform the procedure in Algorithm 1 for 20 independent repetitions. The final value of l used in subsequent simulations is determined as the average of the estimates obtained from these repetitions.

S3 Proof of Theorem 1

We first introduce some notations. For a general vector $\boldsymbol{\beta} \in \mathbb{R}^p$ and a subset $S \subseteq \{1, \ldots, p\}$, define $\boldsymbol{\beta}_S = (\beta_{S,1}, \ldots, \beta_{S,p})^\top \in \mathbb{R}^p$, where $\beta_{S,j} := \beta_j \mathbb{I}\{j \in S\}$ for $j = 1, \ldots, p$, and $\mathbb{I}\{\cdot\}$ is the indicator function.

Let S denote the support of $\boldsymbol{\beta}^*$. Let $\widehat{\mathbf{\Sigma}}^{\mathcal{A}_h} = \sum_{k \in \mathcal{A}_h} \alpha_k \widehat{\mathbf{\Sigma}}^{(k)}$ where $\widehat{\mathbf{\Sigma}}^{(k)} = (X^{(k)})^\top X^{(k)}/n_k$,

Algorithm 1: Band Selection

Input : $(X^{(0)}, \boldsymbol{y}^{(0)}) \in \mathbb{R}^{n_0 \times p} \times \mathbb{R}^{n_0}$, $0 < K < s < n_0$

Output: l

1. Compute the Trans-Lasso estimator $\widehat{\boldsymbol{\beta}}$.
2. For $i = 1, \ldots, n_0$, compute the Trans-Lasso model residuals: $e_i = y_i^{(0)} - (\boldsymbol{x}_i^{(0)})^\top \widehat{\boldsymbol{\beta}}$.
3. Using the full residual sequence $\{e_i\}_{i=1}^{n_0}$, compute the $K \times K$ matrix $\widehat{\boldsymbol{\Sigma}}_K^\epsilon$.
4. Divide $\{e_i\}_{i=1}^{n_0}$ into overlapping blocks as follows:
   (a) For $t = 1, \ldots, n_0 - s + 1$:
      i. Define the t-th block: $\{e_t, \ldots, e_{t+s-1}\}$.
5. For a fixed integer $0 \leq l < K$, based on the t-th block $\{e_t, \ldots, e_{t+s-1}\}$, compute the $K \times K$ banded matrix with bandwidth l, $\widehat{\boldsymbol{\Sigma}}_{s,l,t}^\epsilon$.
6. Compute the estimated risk:
$$\widehat{R}(l) = \frac{1}{n_0 - s + 1} \sum_{t=1}^{n_0 - s + 1} |\widehat{\boldsymbol{\Sigma}}_{s,l,t}^\epsilon - \widehat{\boldsymbol{\Sigma}}_K^\epsilon|_1.$$
7. Select l via $l = \arg\min_{l \in \{0, \ldots, K-1\}} \widehat{R}(l)$.

and $\boldsymbol{\Sigma}^{\mathcal{A}_h} = \sum_{k \in \mathcal{A}_h} \alpha_k \boldsymbol{\Sigma}^{(k)}$.

Lemma 1. *Under the condition of Example 1, we have*

$$\Phi_{t,\iota} \leq c_0 |A_t|_2 \leq c_0 |A_t|_F.$$

Proof of Lemma 1: Since $\boldsymbol{\eta}_i$ are independent and identically distributed sub-Gaussian random variables, we have

$$\sup_{|v|_2 = 1} \|v^\top \boldsymbol{\eta}_0\|_{\psi_2} \leq c_1,$$

where c_1 is a constant. Let $\mathbf{z} = \boldsymbol{\eta}_0 - \boldsymbol{\eta}_0'$.

$$\sup_{|v|_2=1} \|v^\top A_t \mathbf{z}\|_{\psi_2} \leq \sup_{|v|_2=1} |A_t^\top v|_2 \|\mathbf{z}\|_{\psi_2} \leq |A_t|_2 2c_1,$$

where we used the fact that $\|\mathbf{z}\|_{\psi_2} \leq 2c_1$. Using the moment bound for sub-Gaussian random variables:

$$\|v^\top A_t \mathbf{z}\|_\iota \leq C\iota^{1/2} \|v^\top A_t \mathbf{z}\|_{\psi_2} \leq C\iota^{1/2} |A_t|_2 \cdot 2c_1,$$

where C is an constant. Hence,

$$\phi_{t,\iota} \leq 2Cc_1 \iota^{1/2} |A_t|_2 \leq 2Cc_1 \iota^{1/2} |A_t|_F.$$

Lemma 2. ***(Concentration inequalities under dependence)***

i If $\|\boldsymbol{\epsilon} \cdot\|_{q,\alpha} < \infty$, where $q > 2$, $\alpha > 0$, and $\sum_{i=1}^n a_i^2 = n$. Let $\boldsymbol{a} = (a_1, \ldots, a_n)^\top$, and ζ_n=1(resp. $\zeta_n = (\log n)^{1+2q}$ or $\zeta_n = n^{q/2-1-\alpha q}$) if $\alpha > 1/2 - 1/q$(resp. $\alpha = 0$ or $\alpha < 1/2 - 1/q$). Then for all $x > 0, S_n = \sum_{i=1}^n a_i \epsilon_i$, we have

$$\mathbb{P}(|S_n| \geq x) \leq K_1 \frac{\zeta_n |\boldsymbol{a}|_q^q \|\boldsymbol{\epsilon} \cdot\|_{q,\alpha}^q}{x^q} + K_2 \exp\Big(-\frac{K_3 x^2}{n\|\boldsymbol{\epsilon} \cdot\|_{2,\alpha}^2}\Big),$$

where K_1, K_2, K_3 are constants that only depend of q and α.

ii If $\mathcal{D}_z < \infty$ and $\sum_{i=1}^n a_i^2 = n$. Let $\alpha = 2/(1+2z)$, c_α is a constant only depending on α. Then for all $x > 0$, $S_n = \sum_{i=1}^n a_i \epsilon_i$, we have

$$\mathbb{P}(|S_n| \geq nx) \leq (2 + \sqrt{2} c_\alpha) \exp\Big(-\frac{(\sqrt{n}x/\mathcal{D}_z)^\alpha}{2e\alpha}\Big).$$

Proof: See Theorems 2 and 3 of Wu and Wu (2016). Details are omitted.

Lemma 3. *According to the definition of $C_{\boldsymbol{\Sigma}}$ in Section 3.1, we have $|\boldsymbol{\delta}^{\mathcal{A}_h}|_1 \leqslant c_1 C_{\boldsymbol{\Sigma}} h$.*

Proof: Based on the definition of $\boldsymbol{\delta}^{\mathcal{A}_h}$, we know that,

$$
\begin{aligned}
|\boldsymbol{\delta}^{\mathcal{A}_h}|_1 &= \Big|\Big(\sum_{k\in\mathcal{A}_h}\alpha_k\boldsymbol{\Sigma}^{(k)}\Big)^{-1}\sum_{k\in\mathcal{A}_h}\alpha_k\boldsymbol{\Sigma}^{(k)}\boldsymbol{\delta}^{(k)}\Big|_1 \\
&\leqslant \Big|\Big(\sum_{k\in\mathcal{A}_h}\alpha_k\boldsymbol{\Sigma}^{(k)}\Big)^{-1}\sum_{k\in\mathcal{A}_h}\alpha_k(\boldsymbol{\Sigma}^{(k)}-\boldsymbol{\Sigma}^{(0)})\boldsymbol{\delta}^{(k)}\Big|_1 + \Big|\Big(\sum_{k\in\mathcal{A}_h}\alpha_k\boldsymbol{\Sigma}^{(k)}\Big)^{-1}\sum_{k\in\mathcal{A}_h}\alpha_k\boldsymbol{\Sigma}^{(0)}\boldsymbol{\delta}^{(k)}\Big|_1.
\end{aligned}
$$

Firstly, it is easy to show that $\Big|\Big(\sum_{k\in\mathcal{A}_h}\alpha_k\boldsymbol{\Sigma}^{(k)}\Big)^{-1}\sum_{k\in\mathcal{A}_h}\alpha_k(\boldsymbol{\Sigma}^{(k)}-\boldsymbol{\Sigma}^{(0)})\boldsymbol{\delta}^{(k)}\Big|_1 \leqslant C_{\boldsymbol{\Sigma}}h$. Then for the second part, we have

$$
\begin{aligned}
&\Big|\Big(\sum_{k\in\mathcal{A}_h}\alpha_k\boldsymbol{\Sigma}^{(k)}\Big)^{-1}\sum_{k\in\mathcal{A}_h}\alpha_k\boldsymbol{\Sigma}^{(0)}\boldsymbol{\delta}^{(k)}\Big|_1 \\
\leqslant &\sum_{k\in\mathcal{A}_h}\alpha_k\Big|\Big(\sum_{k\in\mathcal{A}_h}\alpha_k\boldsymbol{\Sigma}^{(k)}\Big)^{-1}\boldsymbol{\Sigma}^{(0)}\boldsymbol{\delta}^{(k)}\Big|_1 \\
\leqslant &\sum_{k\in\mathcal{A}_h}\alpha_k\Big|\Big(\sum_{k\in\mathcal{A}_h}\alpha_k\boldsymbol{\Sigma}^{(k)}\Big)^{-1}\Big(\boldsymbol{\Sigma}^{(0)}-\sum_{j\in\mathcal{A}_h}\alpha_j\boldsymbol{\Sigma}^{(j)}\Big)\boldsymbol{\delta}^{(k)}\Big|_1 + \sum_{k\in\mathcal{A}_h}\alpha_k\Big|\Big(\sum_{k\in\mathcal{A}_h}\alpha_k\boldsymbol{\Sigma}^{(k)}\Big)^{-1}\sum_{j\in\mathcal{A}_h}\alpha_j\boldsymbol{\Sigma}^{(j)}\boldsymbol{\delta}^{(k)}\Big|_1 \\
\leqslant &C_{\boldsymbol{\Sigma}}h + h.
\end{aligned}
$$

Hence, we have $|\boldsymbol{\delta}^{\mathcal{A}_h}|_1 \leqslant c_1 C_{\boldsymbol{\Sigma}}h$.

Lemma 4. *Under the conditions of Theorem 1 , consider an index set S with cardinality $|S|$. We have*

$$
\begin{aligned}
&|\boldsymbol{\beta}_S|_1 \leq C\sqrt{|S|}\sqrt{\boldsymbol{\beta}^\top\widehat{\boldsymbol{\Sigma}}^{(0)}\boldsymbol{\beta}}, \quad |\boldsymbol{\beta}_S|_1 \leq C\sqrt{|S|}\sqrt{\boldsymbol{\beta}^\top\widehat{\boldsymbol{\Sigma}}^{\mathcal{A}_h}\boldsymbol{\beta}}, \\
&c_1|\boldsymbol{\beta}|_2^2 \leq \boldsymbol{\beta}^\top\widehat{\boldsymbol{\Sigma}}^{(0)}\boldsymbol{\beta} \leq c_2|\boldsymbol{\beta}|_2^2, \quad c_1|\boldsymbol{\beta}|_2^2 \leq \boldsymbol{\beta}^\top\widehat{\boldsymbol{\Sigma}}^{\mathcal{A}_h}\boldsymbol{\beta} \leq c_2|\boldsymbol{\beta}|_2^2,
\end{aligned}
$$

for all vectors $\{\boldsymbol{\beta}\in\mathbb{R}^p : |\boldsymbol{\beta}_{S^c}|_1 \leq 3|\boldsymbol{\beta}_S|_1\}$, with probability at least $1-O((\log p)^{-2})$ or $1-O(p^{-2})$.

Proof. For the case that $\sup_{|\alpha|_2=1}\|\alpha^\top\boldsymbol{x}^{(0)}\cdot\|_{\iota,\alpha_X} = N_X^0 \leq N_X < \infty$, by the similarly arguments in the proof of Theorem 4 (i) in Wu and Wu (2016), we have

$$
\mathbb{P}\Big(\Big|\frac{X^{(0)\top}X^{(0)}}{n_0} - \boldsymbol{\Sigma}^{(0)}\Big|_{\max} \geq a\Big) \lesssim \frac{p^2 n_0^\chi N_X^\iota}{(n_0 a)^{\iota/2}} + p^2 e^{-c_1 n_0 a^2/N_X^4}.
$$

Then, take $\lambda_X^0 = c_1 N_X^2 \max\{n_0^{-(1-2\chi/\iota)}(p\log p)^{4/\iota}, \sqrt{\log p/n_0}\}$, where c_1 is a sufficient large constant. Here, the superscript "0" indicates that the tuning parameter is constructed based on the target sample of size n_0. In addition, the superscript "k" corresponds to replacing n_0 by the k-th source sample size n_k in the definition above, and the superscript "$\mathcal{A}$" denotes the case where n_0 is replaced by the total auxiliary sample size $n_{\mathcal{A}_h}$ (the same notation applies below). Under this choice of λ_X^0, we obtain

$$\mathbb{P}\Big(\Big|\frac{X^{(0)\top}X^{(0)}}{n_0} - \mathbf{\Sigma}^{(0)}\Big|_{\max} \geq \lambda_X^0\Big) \lesssim (\log p)^{-2}.$$

For the case that $L_\varrho := \sup_{k\in\mathcal{A}_h}\sup_{q\geq 2} q^{-\varrho}\Phi_{0,q}^{(k)} < \infty$, with some $\varrho \geq 0$, by the similarly arguments in the proof of Theorem 4 (ii) in Wu and Wu (2016), we have

$$\mathbb{P}\Big(\Big|\frac{X^{(0)\top}X^{(0)}}{n_0} - \mathbf{\Sigma}^{(0)}\Big|_{\max} \geq a\Big) \lesssim p^2 e^{-C_1(\sqrt{n_0}a/L_\varrho^2)^{2/(1+4\varrho)}}.$$

Then, take $\lambda_X^0 = c_1 L_\varrho^2 (\log p)^{(1+4\varrho)/2}/\sqrt{n_0}$, where c_1 is a sufficient large constant, we can get

$$\mathbb{P}\Big(\Big|\frac{X^{(0)\top}X^{(0)}}{n_0} - \mathbf{\Sigma}^{(0)}\Big|_{\max} \geq \lambda_X^0\Big) \lesssim p^{-2}.$$

Next, on the set

$$\mathcal{CC}_0 := \Big\{\Big|\frac{X^{(0)\top}X^{(0)}}{n_0} - \mathbf{\Sigma}^{(0)}\Big|_{\max} \leq \lambda_X^0\Big\},$$

as long as

$$32\lambda_X^0 \cdot \frac{|S|}{\lambda_{\min}(\mathbf{\Sigma}^{(0)})} < 1,$$

by Corollary 6.8 in Bühlmann and van de Geer (2011), we obtain

$$|\boldsymbol{\beta}_S|_1 \leq C\sqrt{|S|}\sqrt{\boldsymbol{\beta}^\top \widehat{\mathbf{\Sigma}}^{(0)}\boldsymbol{\beta}}, \quad c_1|\boldsymbol{\beta}|_2^2 \leq c_2\lambda_{\min}(\mathbf{\Sigma}^{(0)})|\boldsymbol{\beta}|_2^2 \leq \boldsymbol{\beta}^\top\widehat{\mathbf{\Sigma}}^{(0)}\boldsymbol{\beta} \leq c_3\lambda_{\max}(\mathbf{\Sigma}^{(0)})|\boldsymbol{\beta}|_2^2 \leq c_4|\boldsymbol{\beta}|_2^2$$

for all $\{\boldsymbol{\beta} \in \mathbb{R}^p : |\boldsymbol{\beta}_{S^c}|_1 \leq 3|\boldsymbol{\beta}_S|_1\}$. For $\widehat{\boldsymbol{\Sigma}}^{\mathcal{A}_h}$, we have

$$\begin{aligned}\mathbb{P}(|\widehat{\boldsymbol{\Sigma}}^{\mathcal{A}_h} - \boldsymbol{\Sigma}^{\mathcal{A}_h}|_{\max} \geq t) =& \mathbb{P}\Big(\Big|\sum_{k\in\mathcal{A}_h}(\alpha_k\widehat{\boldsymbol{\Sigma}}^{(k)} - \alpha_k\boldsymbol{\Sigma}^{(k)})\Big|_{\max} \geq t\Big) \\ \leq& \mathbb{P}\Big(1/n_{\mathcal{A}_h}\sum_{k\in\mathcal{A}_h}\Big|\sum_{i=1}^{n_k}(\boldsymbol{x}_i^k(\boldsymbol{x}_i^{(k)})^\top - \boldsymbol{\Sigma}^{(k)})\Big|_{\max} \geq t\Big) \\ \leq& \sum_{k\in\mathcal{A}_h}\mathbb{P}\Big(1/n_{\mathcal{A}_h}\Big|\sum_{i=1}^{n_k}(\boldsymbol{x}_i^k(\boldsymbol{x}_i^{(k)})^\top - \boldsymbol{\Sigma}^{(k)})\Big|_{\max} \geq t/|\mathcal{A}_h|\Big).\end{aligned}$$

We similarly get that

$$\mathbb{P}\Big(1/n_k\Big|\sum_{i=1}^{n_k}(\boldsymbol{x}_i^k(\boldsymbol{x}_i^{(k)})^\top - \boldsymbol{\Sigma}^{(k)})\Big|_{\max} \geq \lambda_X^k\Big) = \begin{cases} O((\log p)^{-2}), & \text{case (i)}, \\ O(p^{-2}), & \text{case (ii)}. \end{cases}$$

Hence, for case (i) and $n_k/n_{\mathcal{A}_h} = a_k = o(1)$, we have

$$\begin{aligned}&\mathbb{P}\Big(1/n_{\mathcal{A}_h}\Big|\sum_{i=1}^{n_k}(\boldsymbol{x}_i^k(\boldsymbol{x}_i^{(k)})^\top - \boldsymbol{\Sigma}^{(k)})\Big|_{\max} \geq c_1\max\{n_{\mathcal{A}_h}^{-(1-2\chi/\iota)}(p\log p)^{4/\iota}, \sqrt{\log p/n_{\mathcal{A}_h}}\}\Big) \\ =&\mathbb{P}\Big(1/n_k\Big|\sum_{i=1}^{n_k}(\boldsymbol{x}_i^k(\boldsymbol{x}_i^{(k)})^\top - \boldsymbol{\Sigma}^{(k)})\Big|_{\max} \geq c_1\max\{n_k^{-(1-2\chi/\iota)}(p\log p)^{4/\iota}a_k^{-2\chi/\iota}, \sqrt{\log p/(n_k a_k)}\}\Big) \\ \lesssim&(\log p)^{-2}.\end{aligned}$$

For case (i) and $n_k/n_{\mathcal{A}_h} = a_k = O(1)$, we have

$$\mathbb{P}\Big(1/n_{\mathcal{A}_h}\Big|\sum_{i=1}^{n_k}(\boldsymbol{x}_i^k(\boldsymbol{x}_i^{(k)})^\top - \boldsymbol{\Sigma}^{(k)})\Big|_{\max} \geq c_1\max\{n_{\mathcal{A}_h}^{-(1-2\chi/\iota)}(p\log p)^{4/\iota}, \sqrt{\log p/n_{\mathcal{A}_h}}\}\Big) = O((\log p)^{-2}).$$

Hence, for case (i), we have $\mathbb{P}(|\widehat{\boldsymbol{\Sigma}}^{\mathcal{A}_h} - \boldsymbol{\Sigma}^{\mathcal{A}_h}|_{\max} \geq \lambda_X^{\mathcal{A}}) = O((\log p)^{-2})$. And for case (ii), we can similarly get that $\mathbb{P}(|\widehat{\boldsymbol{\Sigma}}^{\mathcal{A}_h} - \boldsymbol{\Sigma}^{\mathcal{A}_h}|_{\max} \geq \lambda_X^{\mathcal{A}}) = O(p^{-2})$. Since $|\mathcal{A}_h|$ is fixed and $\widehat{\boldsymbol{\Sigma}}^{\mathcal{A}_h}$ is a linear combination of $\widehat{\boldsymbol{\Sigma}}^{(k)}$, we can prove the remaining results in a similar manner. This completes the proof.

Lemma 5. *Under the conditions of Theorem 1, for* $\widehat{\boldsymbol{u}}^{\mathcal{A}_h} = \widehat{\boldsymbol{w}}^{\mathcal{A}_h} - \boldsymbol{w}^{\mathcal{A}_h}$,

- *For the case (i) in Theorem 1, with probability at least* $1 - C_1(\log p)^{-1}$, *we have*

$$|\widehat{\boldsymbol{u}}^{\mathcal{A}_h}|_2^2 \leqslant C(s_0\lambda_{\boldsymbol{w}}^2 + (C_{\boldsymbol{\Sigma}}h)^2), \text{ and } |\widehat{\boldsymbol{u}}^{\mathcal{A}_h}|_1 \leqslant C(s_0\lambda_{\boldsymbol{w}} + C_{\boldsymbol{\Sigma}}h). \tag{S3.3}$$

- *For the case (ii) in Theorem 1, with probability at least* $1-C_1p^{-2}$*, we have the bounds* (S3.3).

Proof of Lemma 5. In the event that

$$E_1 = \left\{ \frac{1}{n_{\mathcal{A}_h}} \Big| \sum_{k\in\mathcal{A}_h} (X^{(k)})^\top (\boldsymbol{y}^{(k)} - X^{(k)}\boldsymbol{w}^{\mathcal{A}_h}) \Big|_\infty \leqslant \frac{\lambda_{\boldsymbol{w}}}{2}, \right.$$
$$\left. |\widehat{\boldsymbol{\Sigma}}^{(0)} - \boldsymbol{\Sigma}^{(0)}|_{\max} \leq \lambda_X^0, \quad |\widehat{\boldsymbol{\Sigma}}^{\mathcal{A}_h} - \boldsymbol{\Sigma}^{\mathcal{A}_h}|_{\max} \leq \lambda_X^{\mathcal{A}} \right\},$$

the following oracle inequality holds for $\widehat{\boldsymbol{u}}^{\mathcal{A}_h} = \widehat{\boldsymbol{w}}^{\mathcal{A}_h} - \boldsymbol{w}^{\mathcal{A}_h}$,

$$\begin{aligned}\frac{1}{2}(\widehat{\boldsymbol{u}}^{\mathcal{A}_h})^\top \widehat{\boldsymbol{\Sigma}}^{\mathcal{A}_h} \widehat{\boldsymbol{u}}^{\mathcal{A}_h} &\leqslant \lambda_{\boldsymbol{w}}|\boldsymbol{w}^{\mathcal{A}_h}|_1 - \lambda_{\boldsymbol{w}}|\widehat{\boldsymbol{w}}^{\mathcal{A}_h}|_1 + \frac{1}{n_{\mathcal{A}_h}} \Big| (\widehat{\boldsymbol{u}}^{\mathcal{A}_h})^\top \sum_{k\in\mathcal{A}_h} (X^{(k)})^\top (\boldsymbol{y}^{(k)} - X^{(k)}\boldsymbol{w}^{\mathcal{A}_h}) \Big| \\ &\leqslant \lambda_{\boldsymbol{w}}|\boldsymbol{w}^{\mathcal{A}_h}|_1 - \lambda_{\boldsymbol{w}}|\widehat{\boldsymbol{w}}^{\mathcal{A}_h}|_1 + \frac{\lambda_{\boldsymbol{w}}}{2}|\widehat{\boldsymbol{u}}^{\mathcal{A}_h}|_1,\end{aligned}$$

where the last step is due to the first statement in E_1. As a result,

$$\begin{aligned}\frac{1}{2}(\widehat{\boldsymbol{u}}^{\mathcal{A}_h})^\top \widehat{\boldsymbol{\Sigma}}^{\mathcal{A}_h} \widehat{\boldsymbol{u}}^{\mathcal{A}_h} \leq& \lambda_w|\boldsymbol{w}_S^{\mathcal{A}_h}|_1 + \lambda_w|\boldsymbol{w}_{S^c}^{\mathcal{A}_h}|_1 - \lambda_w|\widehat{\boldsymbol{w}}_S^{\mathcal{A}_h}|_1 - \lambda_w|\widehat{\boldsymbol{w}}_{S^c}^{\mathcal{A}_h}|_1 + \frac{1}{2}\lambda_w|\widehat{\boldsymbol{u}}_S^{\mathcal{A}_h}|_1 + \frac{1}{2}\lambda_w|\widehat{\boldsymbol{u}}_{S^c}^{\mathcal{A}_h}|_1 \\ \leq& \lambda_w|\widehat{\boldsymbol{w}}_S^{\mathcal{A}_h}|_1 + \lambda_w|\widehat{\boldsymbol{u}}_S^{\mathcal{A}_h}|_1 + \lambda_w|\boldsymbol{w}_{S^c}^{\mathcal{A}_h}|_1 - \lambda_w|\widehat{\boldsymbol{w}}_S^{\mathcal{A}_h}|_1 - \lambda_w|\widehat{\boldsymbol{w}}_{S^c}^{\mathcal{A}_h}|_1 \\ &+ \frac{1}{2}\lambda_w|\widehat{\boldsymbol{u}}_S^{\mathcal{A}_h}|_1 + \frac{1}{2}\lambda_w|\widehat{\boldsymbol{u}}_{S^c}^{\mathcal{A}_h}|_1 \\ \leq& \frac{3}{2}\lambda_{\boldsymbol{w}}|\widehat{\boldsymbol{u}}_S^{\mathcal{A}_h}|_1 + \lambda_w|\boldsymbol{w}_{S^c}^{\mathcal{A}_h}|_1 - \lambda_{\boldsymbol{w}}|\widehat{\boldsymbol{w}}_{S^c}^{\mathcal{A}_h}|_1 + \frac{\lambda_{\boldsymbol{w}}}{2}|\widehat{\boldsymbol{u}}_{S^c}^{\mathcal{A}_h}|_1.\end{aligned}$$

Using the fact that $|\widehat{\boldsymbol{u}}_{S^c}^{\mathcal{A}_h}|_1 \leqslant |\widehat{\boldsymbol{w}}_{S^c}^{\mathcal{A}_h}|_1 + |w_{S^c}^{\mathcal{A}_h}|_1$, we arrive at the inequality that

$$\frac{1}{2}(\widehat{\boldsymbol{u}}^{\mathcal{A}_h})^\top \widehat{\boldsymbol{\Sigma}}^{\mathcal{A}_h} \widehat{\boldsymbol{u}}^{\mathcal{A}_h} \leqslant \frac{3}{2}\lambda_{\boldsymbol{w}}|\widehat{\boldsymbol{u}}_S^{\mathcal{A}_h}|_1 + 2\lambda_{\boldsymbol{w}}|\boldsymbol{w}_{S^c}^{\mathcal{A}_h}|_1 - \frac{\lambda_{\boldsymbol{w}}}{2}|\widehat{\boldsymbol{u}}_{S^c}^{\mathcal{A}_h}|_1.$$

(i) If $\frac{3}{2}\lambda_{\boldsymbol{w}}|\widehat{\boldsymbol{u}}_S^{\mathcal{A}_h}|_1 \geqslant 2\lambda_{\boldsymbol{w}}|\boldsymbol{w}_{S^c}^{\mathcal{A}_h}|_1$,

$$\frac{1}{2}(\widehat{\boldsymbol{u}}^{\mathcal{A}_h})^\top \widehat{\boldsymbol{\Sigma}}^{\mathcal{A}_h} \widehat{\boldsymbol{u}}^{\mathcal{A}_h} \leqslant 3\lambda_{\boldsymbol{w}}|\widehat{\boldsymbol{u}}_S^{\mathcal{A}_h}|_1 - \frac{\lambda_{\boldsymbol{w}}}{2}|\widehat{\boldsymbol{u}}_{S^c}^{\mathcal{A}_h}|_1,$$

which means that $|\widehat{\boldsymbol{u}}_{S^c}^{\mathcal{A}_h}|_1 \leq 6|\widehat{\boldsymbol{u}}_S^{\mathcal{A}_h}|_1$. By similar arguments as in Lemma 4,

$$c_1|\widehat{\boldsymbol{u}}^{\mathcal{A}_h}|_2^2 \leqslant 3\lambda_{\boldsymbol{w}}|\widehat{\boldsymbol{u}}_S^{\mathcal{A}_h}|_1,$$

Under the constraint of convex cone, $|\widehat{\boldsymbol{u}}^{\mathcal{A}_h}|_1 \leqslant |\widehat{\boldsymbol{u}}_S^{\mathcal{A}_h}|_1 + |\widehat{\boldsymbol{u}}_{S^c}^{\mathcal{A}_h}|_1 \leqslant 7|\widehat{\boldsymbol{u}}_S^{\mathcal{A}_h}|_1 \leqslant 7\sqrt{s_0}|\widehat{\boldsymbol{u}}^{\mathcal{A}_h}|_2$.

Therefore, we have

$$|\widehat{\boldsymbol{u}}^{\mathcal{A}_h}|_2^2 \leqslant C_1 s_0 \lambda_{\boldsymbol{w}}^2 \text{ and } |\widehat{\boldsymbol{u}}^{\mathcal{A}_h}|_1 \leqslant C_2 s_0 \lambda_{\boldsymbol{w}}. \tag{S3.4}$$

(ii) If $\frac{3}{2}\lambda_{\boldsymbol{w}}|\widehat{\boldsymbol{u}}_S^{\mathcal{A}_h}|_1 \leqslant 2\lambda_{\boldsymbol{w}}|w_{S^c}^{\mathcal{A}_h}|_1$,

$$\frac{1}{2}(\widehat{\boldsymbol{u}}^{\mathcal{A}_h})^\top \widehat{\boldsymbol{\Sigma}}^{\mathcal{A}_h}\widehat{\boldsymbol{u}}^{\mathcal{A}_h} \leqslant 4\lambda_{\boldsymbol{w}}|w_{S^c}^{\mathcal{A}_h}|_1 - \frac{\lambda_{\boldsymbol{w}}}{2}|\widehat{\boldsymbol{u}}_{S^c}^{\mathcal{A}_h}|_1.$$

Therefore,

$$|\widehat{\boldsymbol{u}}^{\mathcal{A}_h}|_1 = |\widehat{\boldsymbol{u}}_S^{\mathcal{A}_h}|_1 + |\widehat{\boldsymbol{u}}_{S^c}^{\mathcal{A}_h}|_1 \leqslant 8|w_{S^c}^{\mathcal{A}_h}|_1 + 2|w_{S^c}^{\mathcal{A}_h}|_1 \leqslant 10|\boldsymbol{\delta}_{S^c}^{\mathcal{A}_h}|_1 \leqslant c_1 C_{\boldsymbol{\Sigma}} h.$$

Hence, a direct upper bound on $|\widehat{\boldsymbol{u}}^{\mathcal{A}_h}|_2$ is

$$|\widehat{\boldsymbol{u}}^{\mathcal{A}_h}|_2 \leqslant |\widehat{\boldsymbol{u}}^{\mathcal{A}_h}|_1 \leqslant c_1 C_{\boldsymbol{\Sigma}} h.$$

Together with (S3.4), we have

$$|\widehat{\boldsymbol{u}}^{\mathcal{A}_h}|_2^2 \leqslant C_1[s_0\lambda_{\boldsymbol{w}}^2 + (C_{\boldsymbol{\Sigma}}h)^2] \text{ and } |\widehat{\boldsymbol{u}}^{\mathcal{A}_h}|_1 \leqslant C_2(s_0\lambda_{\boldsymbol{w}} + C_{\boldsymbol{\Sigma}}h).$$

It is left to verify that $\mathbb{P}(E_1) \to 1$. Notice that

$$\begin{aligned}
&\frac{1}{n_{\mathcal{A}_h}}\Big|\sum_{k\in\mathcal{A}_h}(X^{(k)})^\top(\boldsymbol{y}^{(k)} - X^{(k)}\boldsymbol{w}^{\mathcal{A}_h})\Big|_\infty \\
=&\Big|\frac{1}{n_{\mathcal{A}_h}}\sum_{k\in\mathcal{A}_h}(X^{(k)})^\top\boldsymbol{\epsilon}^{(k)} + \sum_{k\in\mathcal{A}_h}\alpha_k\widehat{\boldsymbol{\Sigma}}^{(k)}\boldsymbol{w}^{(k)} - \widehat{\boldsymbol{\Sigma}}^{\mathcal{A}_h}\boldsymbol{w}^{\mathcal{A}_h}\Big|_\infty \\
\leq&\Big|\frac{1}{n_{\mathcal{A}_h}}\sum_{k\in\mathcal{A}_h}(X^{(k)})^\top\boldsymbol{\epsilon}^{(k)} + \sum_{k\in\mathcal{A}_h}\alpha_k(\boldsymbol{\Sigma}^{(k)} - \widehat{\boldsymbol{\Sigma}}^{(k)})\boldsymbol{\delta}^{(k)}\Big|_\infty + |(\widehat{\boldsymbol{\Sigma}}^{\mathcal{A}_h} - \boldsymbol{\Sigma}^{\mathcal{A}_h})\boldsymbol{\delta}^{\mathcal{A}_h}|_\infty \\
\leq&\Big|\frac{1}{n_{\mathcal{A}_h}}\sum_{k\in\mathcal{A}_h}(X^{(k)})^\top\boldsymbol{\epsilon}^{(k)}\Big|_\infty + \Big|\sum_{k\in\mathcal{A}_h}\alpha_k(\widehat{\boldsymbol{\Sigma}}^{(k)} - \boldsymbol{\Sigma}^{(k)})\boldsymbol{\delta}^{(k)}\Big|_\infty + |(\widehat{\boldsymbol{\Sigma}}^{\mathcal{A}_h} - \boldsymbol{\Sigma}^{\mathcal{A}_h})\boldsymbol{\delta}^{\mathcal{A}_h}|_\infty.
\end{aligned}$$

$$\begin{aligned}
\mathbb{P}\Big(\frac{1}{n_{\mathcal{A}_h}}\Big|\sum_{k\in\mathcal{A}_h}(X^{(k)})^\top\boldsymbol{\epsilon}^{(k)}\Big|_\infty \geq t\Big) &= \mathbb{P}\Big(\frac{1}{n_{\mathcal{A}_h}}\max_{j\leq p}\Big|\sum_{k\in\mathcal{A}_h}(\boldsymbol{X}_j^{(k)})^\top\boldsymbol{\epsilon}^{(k)}\Big| \geq t\Big) \\
&\leqslant \sum_{k\in\mathcal{A}_h}\mathbb{P}\Big(\max_{j\leq p}\frac{1}{n_{\mathcal{A}_h}}|(\boldsymbol{X}_j^{(k)})^\top\boldsymbol{\epsilon}^{(k)}| \geq t/|\mathcal{A}_h|\Big).
\end{aligned}$$

For the case (ii) in Theorem 1, by the similarlr arguments as in the proof of Theorem 4(i) in Wu and Wu (2016), we have that

$$\mathbb{P}\Big(\frac{1}{n_k}|(\boldsymbol{X}_j^{(k)})^\top \boldsymbol{\epsilon}^{(k)}| \geq a\Big) \lesssim \frac{n_k^\pi N_X^\tau N_e^\tau}{(n_k a)^\tau} + e^{-c_2 n_k a^2/(N_X^2 N_e^2)}.$$

Hence, for $n_k/n_{\mathcal{A}_h} = a_k = o(1)$, we have

$$\begin{aligned}\mathbb{P}\Big(\frac{1}{n_{\mathcal{A}_h}}|(\boldsymbol{X}_j^{(k)})^\top \boldsymbol{\epsilon}^{(k)}| \geq a\Big) \lesssim & \frac{n_k^\pi N_X^\tau N_e^\tau}{(n_k a/a_k)^\tau} + e^{-c_2 n_k a^2/(a_k^2 N_X^2 N_e^2)} \\ = & \frac{n_k^\pi N_X^\tau N_e^\tau}{(n_{\mathcal{A}_h} a)^\tau} + e^{-c_2 n_{\mathcal{A}_h} a^2/(a_k N_X^2 N_e^2)} \\ \leq & \frac{n_{\mathcal{A}_h}^\pi N_X^\tau N_e^\tau}{(n_{\mathcal{A}_h} a)^\tau} + e^{-c_2 n_{\mathcal{A}_h} a^2/(N_X^2 N_e^2)}.\end{aligned}$$

For $n_k/n_{\mathcal{A}_h} = a_k = O(1)$, we also have

$$\mathbb{P}\Big(\frac{1}{n_{\mathcal{A}_h}}|(\boldsymbol{X}_j^{(k)})^\top \boldsymbol{\epsilon}^{(k)}| \geq a\Big) \lesssim \frac{n_{\mathcal{A}_h}^\pi N_X^\tau N_e^\tau}{(n_{\mathcal{A}_h} a)^\tau} + e^{-c_2 n_{\mathcal{A}_h} a^2/(N_X^2 N_e^2)}.$$

That is to say we have

$$\mathbb{P}\Big(\frac{1}{n_{\mathcal{A}_h}}\Big|\sum_{k\in\mathcal{A}_h}(X^{(k)})^\top \boldsymbol{\epsilon}^{(k)}\Big|_\infty \geq a\Big) \lesssim p\frac{n_{\mathcal{A}_h}^\pi N_X^\tau N_e^\tau}{(n_{\mathcal{A}_h} a)^\tau} + pe^{-c_2 n_{\mathcal{A}_h} a^2/(N_X^2 N_e^2)},$$

Hence, under our choice of $a \geq A\max\{(n_{\mathcal{A}_h}^{-1}\log p)^{1/2} N_X N_e, n_{\mathcal{A}_h}^{\pi/\tau-1}(p\log p)^{1/\tau} N_x N_e\}$, where A is a sufficiently large constant, we have

$$\mathbb{P}\Big(\frac{1}{n_{\mathcal{A}_h}}\Big|\sum_{k\in\mathcal{A}_h}(X^{(k)})^\top \boldsymbol{\epsilon}^{(k)}\Big|_\infty > t\Big) \leq C_1(\log p)^{-1}.$$

For the case (ii) in Theorem 1, by the similar arguments in the proof of Theorem 4 (ii) in Wu and Wu (2016), we have

$$\mathbb{P}\Big(\frac{1}{n_k}|(\boldsymbol{X}_j^{(k)})^\top \boldsymbol{\epsilon}^{(k)}| \geq t\Big) \lesssim e^{-C_1(\sqrt{n_k}t/(K_\nu L_\varrho))^{2/(1+2\nu+2\varrho)}}$$

We can similarly obtain,

$$\mathbb{P}\Big(\frac{1}{n_{\mathcal{A}_h}}\Big|\sum_{k\in\mathcal{A}_h}(X^{(k)})^\top \boldsymbol{\epsilon}^{(k)}\Big|_\infty \geq t\Big) \leq pe^{-C_1(\sqrt{n_{\mathcal{A}_h}}t/(K_\nu L_\varrho))^{2/(1+2\nu+2\varrho)}}.$$

Let $t \geqslant An_{\mathcal{A}_h}^{-1/2}(\log p)^{(1+2\nu+2\varrho)/2}K_\nu L_\varrho$, for some sufficiently large constant A, we get

$$\mathbb{P}\Big(\frac{1}{n_{\mathcal{A}_h}}\Big|\sum_{k\in\mathcal{A}_h}(X^{(k)})^\top\boldsymbol{\epsilon}^{(k)}\Big|_\infty \geq t\Big) \leq C_1 p^{-2}.$$

And, we have

$$\begin{aligned}
&\mathbb{P}\Big(\Big|\sum_{k\in\mathcal{A}_h}\alpha_k(\widehat{\boldsymbol{\Sigma}}^{(k)}-\boldsymbol{\Sigma}^{(k)})\boldsymbol{\delta}^{(k)}\Big|_\infty \geqslant t\Big)\\
\leq&\sum_{k\in\mathcal{A}_h}\mathbb{P}\Big(\max_{j\leqslant p}\Big|\frac{1}{n_{\mathcal{A}_h}}\sum_{m\leqslant n_k}(X_{mj}^{(k)}\boldsymbol{x}_m^{(k)\top}\boldsymbol{\delta}^{(k)}-\Sigma_{j,\cdot}^{(k)}\boldsymbol{\delta}^{(k)})\Big| \geqslant t/|\mathcal{A}_h|\Big).
\end{aligned}$$

For case (i) in Theorem 1, considering the process $\{X_{mj}^{(k)}\boldsymbol{\delta}^{(k)\top}\boldsymbol{x}_m^{(k)}\}_{m\in\mathbb{Z}}$, we have

$$\begin{aligned}
&\|X_{mj}^{(k)}(\boldsymbol{\delta}^{(k)})^\top\boldsymbol{x}_m^{(k)}-X_{mj}^*(\boldsymbol{\delta}^{(k)})^\top\boldsymbol{x}_m^{(k)*}\|_{\iota/2}\\
\leq&\|(X_{mj}^{(k)}-X_{mj}^{(k)*})(\boldsymbol{\delta}^{(k)})^\top\boldsymbol{x}_m^{(k)}\|_{\iota/2}+\|X_{mj}^{(k)*}((\boldsymbol{\delta}^{(k)})^\top\boldsymbol{x}_m^{(k)}-(\boldsymbol{\delta}^{(k)})^\top\boldsymbol{x}_m^{(k)*})\|_{\iota/2}\\
\lesssim&\|X_{mj}^{(k)}-X_{mj}^{(k)*}\|_\iota\|(\boldsymbol{\delta}^{(k)})^\top\boldsymbol{x}_m^{(k)}\|_\iota+\|X_{mj}^{(k)*}\|_\iota\|(\boldsymbol{\delta}^{(k)})^\top\boldsymbol{x}_m^{(k)}-(\boldsymbol{\delta}^{(k)})^\top\boldsymbol{x}_m^{(k)*}\|_\iota\\
\lesssim&\|X_{mj}^{(k)}-X_{mj}^{(k)*}\|_\iota N_X+\|(\boldsymbol{\delta}^{(k)})^\top\boldsymbol{x}_m^{(k)}-(\boldsymbol{\delta}^{(k)})^\top\boldsymbol{x}_m^{(k)*}\|_\iota N_X
\end{aligned}\tag{S3.5}$$

where $\boldsymbol{x}_m^{(k)*}=h(\mathcal{S}'_m)$. Therefore, combining with $|\boldsymbol{\delta}^{(k)}|_2\leq c_1$ for all $k\leq M$,

$$\begin{aligned}
&\|\{X_{\cdot j}^{(k)}((\boldsymbol{\delta}^{(k)})^\top\boldsymbol{x}^{(k)})-\Sigma_{j,\cdot}^{(k)}\boldsymbol{\delta}^{(k)}\}\cdot\|_{\iota/2,\alpha_X}\\
\lesssim&\ N_X(m+1)^{\alpha_X}\sum_{i\geq M}\sup_{|\boldsymbol{v}|_2=1}\|\boldsymbol{v}^{(k)\top}\boldsymbol{x}_i^{(k)}-\boldsymbol{v}^{(k)\top}\boldsymbol{x}_i^{(k)*}\|_\iota\lesssim N_X^2.
\end{aligned}$$

By Lemma 2 (i), we have

$$\mathbb{P}\Big(\Big|\frac{1}{n_k}\sum_{m\leqslant n_k}(X_{mj}^{(k)}(\boldsymbol{x}_m^{(k)})^\top\boldsymbol{\delta}^{(k)}-\Sigma_{j,\cdot}^{(k)}\boldsymbol{\delta}^{(k)})\Big| \geqslant c_1N_X^2\max\{n_k^{-(1-2\chi/\iota)}(p\log p)^{2/\iota},\sqrt{\log p/n_k}\}\Big)=O((p\log p)^{-1}),$$

for a large enough constant c. Hence, as the similar discuss as above, we can get

$$\mathbb{P}\Big(\Big|\sum_{k\in\mathcal{A}_h}\alpha_k(\widehat{\boldsymbol{\Sigma}}^{(k)}-\boldsymbol{\Sigma}^{(k)})\boldsymbol{\delta}^{(k)}\Big|_\infty \geqslant c_1N_X^2\max\{n_{\mathcal{A}_h}^{-(1-2\chi/\iota)}(p\log p)^{2/\iota},\sqrt{\log p/n_{\mathcal{A}_h}}\}\Big)=O((\log p)^{-1}).$$

For the case (ii) in Theorem 1, similarly by (S3.5), we have

$$\sup_{\iota\geq 2}\iota^{-2\varrho}\sum_{m=0}^{\infty}\|X_{mj}^{(k)}(\boldsymbol{\delta}^{(k)})^\top\boldsymbol{x}_m^{(k)}-X_{mj}^*(\boldsymbol{\delta}^{(k)})^\top\boldsymbol{x}_m^{(k)*}\|_{\iota/2}\lesssim\sup_{\iota\geq 2}\iota^{-2\varrho}\Phi_{0,\iota}^2\lesssim L_\varrho^2.$$

Combining with Lemma 1 (ii), we can obtain that

$$\mathbb{P}\Big(\Big|\sum_{k\in\mathcal{A}_h}\alpha_k(\widehat{\boldsymbol{\Sigma}}^{(k)}-\boldsymbol{\Sigma}^{(k)})\boldsymbol{\delta}^{(k)}\Big|_\infty \geqslant c_1 L_\varrho^2(\log p)^{(1+4\varrho)/2}/\sqrt{n_{\mathcal{A}_h}}\Big)=o(p^{-2}).$$

We can similarly show that

$$\mathbb{P}(|(\widehat{\boldsymbol{\Sigma}}^{\mathcal{A}_h}-\boldsymbol{\Sigma}^{\mathcal{A}_h})\boldsymbol{\delta}^{\mathcal{A}_h}|_\infty \geqslant c_1 N_X^2\max\{n_{\mathcal{A}_h}^{-(1-2\chi/\iota)}(p\log p)^{2/\iota},(\log p/n_{\mathcal{A}_h})^{1/2}\})=O((\log p)^{-1}),$$

or

$$\mathbb{P}\Big(|(\widehat{\boldsymbol{\Sigma}}^{\mathcal{A}_h}-\boldsymbol{\Sigma}^{\mathcal{A}_h})\boldsymbol{\delta}^{\mathcal{A}_h}|_\infty \geqslant c_1 L_\varrho^2(\log p)^{(1+4\varrho)/2}/\sqrt{n_{\mathcal{A}_h}}\Big)=O(p^{-2}).$$

Define that

$$\lambda_{xx}^{\mathcal{A}}=\begin{cases} N_X^2\max\{n_{\mathcal{A}_h}^{2\chi/\iota-1}(p\log p)^{2/\iota},\sqrt{\log p/n_{\mathcal{A}_h}}\}, & \text{case (i)},\\ L_\varrho^2(\log p)^{(1+4\varrho)/2}/\sqrt{n_{\mathcal{A}_h}}, & \text{case (ii)}.\end{cases}$$

and

$$\lambda_{x\epsilon}^{\mathcal{A}}=\begin{cases} N_X N_e\max\{n_{\mathcal{A}_h}^{\pi/\tau-1}(p\log p)^{1/\tau},\sqrt{\log p/n_{\mathcal{A}_h}}\}, & \text{case (i)},\\ K_\nu L_\varrho(\log p)^{(1+2\nu+2\varrho)/2}/\sqrt{n_{\mathcal{A}_h}}, & \text{case (ii)}.\end{cases}$$

Hence it suffices to take $\lambda_{\boldsymbol{w}}\geqslant c_1\max\{\lambda_{xx}^{\mathcal{A}},\lambda_{x\epsilon}^{\mathcal{A}}\}$ for a sufficiently large constant c_1. The probabilities of the latter two terms in E_1 have been effectively controlled in Lemma 4.

Proof of Theorem 1:

$$E_1':=\Big\{\frac{1}{n_0}|(X^{(0)})^\top\boldsymbol{\epsilon}^{(0)}|_\infty\leqslant\frac{\lambda_\delta}{2},\frac{1}{n_0}|X^{(0)}\widehat{\boldsymbol{u}}^{\mathcal{A}_h}|_2^2\leqslant 2\Lambda_{\max}(\boldsymbol{\Sigma}^{(0)})|\widehat{\boldsymbol{u}}^{\mathcal{A}_h}|_2^2\Big\}\cap E_1.$$

The following oracle inequality holds for $\widehat{\boldsymbol{v}}^{\mathcal{A}_h}=\widehat{\boldsymbol{\delta}}^{\mathcal{A}_h}-\boldsymbol{\delta}^{\mathcal{A}_h}$,

$$\frac{1}{2n_0}|X^{(0)}\widehat{\boldsymbol{v}}^{\mathcal{A}_h}|_2^2\leqslant\lambda_\delta|\boldsymbol{\delta}^{\mathcal{A}_h}|_1-\lambda_\delta|\widehat{\boldsymbol{\delta}}^{\mathcal{A}_h}|_1+\frac{1}{n_0}|\langle X^{(0)}\widehat{\boldsymbol{v}}^{\mathcal{A}_h},\boldsymbol{\epsilon}^{(0)}-X^{(0)}\widehat{\boldsymbol{u}}^{\mathcal{A}_h}\rangle|.$$

For the last term, it holds that, in E_1',

$$\frac{1}{n_0}|\langle X^{(0)}\widehat{\boldsymbol{v}}^{\mathcal{A}_h},\boldsymbol{\epsilon}^{(0)}-X^{(0)}\widehat{\boldsymbol{u}}^{\mathcal{A}_h}\rangle|\leqslant\frac{\lambda_\delta}{2}|\widehat{\boldsymbol{v}}^{\mathcal{A}_h}|_1+\frac{1}{n_0}|X^{(0)}\widehat{\boldsymbol{u}}^{\mathcal{A}_h}|_2^2+\frac{1}{4n_0}|X^{(0)}\widehat{\boldsymbol{v}}^{\mathcal{A}_h}|_2^2,$$

where we use the fact that $|ab| \leqslant \frac{ca^2}{2} + \frac{b^2}{2c}$ for every $c > 0$. We arrive the following oracle inequality:

$$\begin{aligned}\frac{1}{4n_0}|X^{(0)}\widehat{\boldsymbol{v}}^{\mathcal{A}_h}|_2^2 &\leqslant \lambda_\delta|\boldsymbol{\delta}^{\mathcal{A}_h}|_1 - \lambda_\delta|\widehat{\boldsymbol{\delta}}^{\mathcal{A}_h}|_1 + \frac{1}{2}\lambda_\delta|\widehat{\boldsymbol{v}}^{\mathcal{A}_h}|_1 + \frac{1}{n_0}|X^{(0)}\widehat{\boldsymbol{u}}^{\mathcal{A}_h}|_2^2 \\ &\leqslant 2\lambda_\delta|\boldsymbol{\delta}^{\mathcal{A}_h}|_1 - \frac{1}{2}\lambda_\delta|\widehat{\boldsymbol{v}}^{\mathcal{A}_h}|_1 + \frac{1}{n_0}|X^{(0)}\widehat{\boldsymbol{u}}^{\mathcal{A}_h}|_2^2.\end{aligned}$$

(i) If $2\lambda_\delta|\boldsymbol{\delta}^{\mathcal{A}_h}|_1 \geqslant \frac{1}{n_0}|X^{(0)}\widehat{\boldsymbol{u}}^{\mathcal{A}_h}|_2^2$, then $|\widehat{\boldsymbol{v}}^{\mathcal{A}_h}|_1 \leqslant 8|\boldsymbol{\delta}^{\mathcal{A}_h}|_1 \leqslant c_1 C_{\boldsymbol{\Sigma}} h$ and

$$\frac{1}{4n_0}|X^{(0)}\widehat{\boldsymbol{v}}^{\mathcal{A}_h}|_2^2 \leqslant 4\lambda_\delta|\boldsymbol{\delta}^{\mathcal{A}_h}|_1.$$

Hence, a direct upper bound on $|\widehat{\boldsymbol{v}}^{\mathcal{A}_h}|_2$ is

$$|\widehat{\boldsymbol{v}}^{\mathcal{A}_h}|_2 \leqslant |\widehat{\boldsymbol{v}}^{\mathcal{A}_h}|_1 \leqslant c_1 C_{\boldsymbol{\Sigma}} h.$$

Then, we obtain

$$\Big(\frac{1}{n_0}|X^{(0)}\widehat{\boldsymbol{v}}^{\mathcal{A}_h}|_2^2 \vee \frac{1}{n_0}|X^{(0)}\widehat{\boldsymbol{u}}^{\mathcal{A}_h}|_2^2\Big) \leqslant c_1\lambda_\delta C_{\boldsymbol{\Sigma}} h \quad \text{ and } |\widehat{\boldsymbol{v}}^{\mathcal{A}_h}|_2^2 \leqslant c_1 (C_{\boldsymbol{\Sigma}} h)^2,$$

for some constant $c_1 > 0$.

(ii) If $2\lambda_\delta|\boldsymbol{\delta}^{\mathcal{A}_h}|_1 \leqslant \frac{1}{n_0}|X^{(0)}\widehat{\boldsymbol{u}}^{\mathcal{A}_h}|_2^2$, then

$$\lambda_\delta|\widehat{\boldsymbol{v}}^{\mathcal{A}_h}|_1 \leqslant \frac{c_2}{n_0}|X^{(0)}\widehat{\boldsymbol{u}}^{\mathcal{A}_h}|_2^2 \text{ and } \frac{1}{n_0}|X^{(0)}\widehat{\boldsymbol{v}}^{\mathcal{A}_h}|_2^2 \leqslant \frac{c_2}{n_0}|X^{(0)}\widehat{\boldsymbol{u}}^{\mathcal{A}_h}|_2^2, \tag{S3.6}$$

for some constant $c_2 > 0$. Together with the second statement in E_1', (S3.6), and Lemma 5, we arrive at

$$\frac{1}{n_0}|X^{(0)}\widehat{\boldsymbol{v}}^{\mathcal{A}_h}|_2^2 \lesssim \frac{1}{n_0}|X^{(0)}\widehat{\boldsymbol{u}}^{\mathcal{A}_h}|_2^2 \lesssim s_0\lambda_{\boldsymbol{w}}^2 + (C_{\boldsymbol{\Sigma}} h)^2,$$

so,

$$|\widehat{\boldsymbol{v}}^{\mathcal{A}_h}|_1 \leqslant \frac{c}{\lambda_\delta}(s_0{\lambda_w}^2 + (C_{\boldsymbol{\Sigma}} h)^2) \leq s_0\lambda_w + (C_{\boldsymbol{\Sigma}} h)^2/\lambda_\delta,$$

where the last inequality is due to $\lambda_\delta \geq \lambda_{\boldsymbol{w}}$.

Combine with the conclusions in (i), we have arrived our desired results,

$$\frac{1}{n_0}|X^{(0)}(\widehat{\boldsymbol{\beta}}-\boldsymbol{\beta}^*)|_2^2 \lesssim s_0\lambda_w^2+\lambda_\delta C_{\boldsymbol{\Sigma}}h+(C_{\boldsymbol{\Sigma}}h)^2, \quad |\widehat{\boldsymbol{\beta}}-\boldsymbol{\beta}^*|_1 \lesssim s_0\lambda_w+(C_{\boldsymbol{\Sigma}}h)^2/\lambda_\delta+C_{\boldsymbol{\Sigma}}h. \tag{S3.7}$$

Finally we show that $\mathbb{P}(E_1') \to 1$. As $X^{(0)}$ is independent of $\widehat{\boldsymbol{u}}^{\mathcal{A}_h}$, by by firstly conditioning on $\widehat{\boldsymbol{u}}^{\mathcal{A}_h}$,we have

$$\mathbb{P}\Big(\frac{1}{n_0}|X^{(0)}\widehat{\boldsymbol{u}}^{\mathcal{A}_h}|_2^2 \geqslant 2\Lambda_{\max}(\boldsymbol{\Sigma}^{(\mathbf{0})})|\widehat{\boldsymbol{u}}^{\mathcal{A}_h}|_2^2\Big) = \begin{cases} O((\log p)^{-1}), & \text{case (i)}, \\ O(p^{-2}), & \text{case (ii)}. \end{cases}$$

For $\mathbb{P}(1/n_0|(X^{(0)})^\top\boldsymbol{\epsilon}^{(0)}|_\infty \geqslant \lambda_\delta/2)$, we can follow the proof of lemma 5, details are omitted.

Remark 1. Under the conditions of Theorem 1, if we further assume that $\lambda_X^{\mathcal{A}}C_{\boldsymbol{\Sigma}}h \lesssim \lambda_w$, then the bound in (S3.7) admits the sharper bounds

$$\frac{1}{n_0}|X^{(0)}(\widehat{\boldsymbol{\beta}}-\boldsymbol{\beta}^*)|_2^2 \lesssim s_0\lambda_w^2+\lambda_\delta C_{\boldsymbol{\Sigma}}h\wedge(C_{\boldsymbol{\Sigma}}h)^2, \quad |\widehat{\boldsymbol{\beta}}-\boldsymbol{\beta}^*|_1 \lesssim s_0\lambda_w+C_{\boldsymbol{\Sigma}}h. \tag{S3.8}$$

Proof of Remark 1: Using

$$(\widehat{\boldsymbol{u}}^{\mathcal{A}_h})^\top\widehat{\boldsymbol{\Sigma}}^{\mathcal{A}_h}\widehat{\boldsymbol{u}}^{\mathcal{A}_h} \lesssim |\widehat{\boldsymbol{u}}^{\mathcal{A}_h}|_2^2+|\widehat{\boldsymbol{u}}^{\mathcal{A}_h}|_1^2\lambda_X^{\mathcal{A}},$$

and proceeding as in the proof of Theorem 1, we obtain the desired results.

Lemma 6. *Suppose $\widehat{\Gamma}$ is an estimator of $\boldsymbol{\Sigma}_x$. We say that $\widehat{\Gamma}$ satisfies the Restricted Strong Convexity (RSC) condition if it satisfies both the lower-RE condition*

$$\boldsymbol{v}^\top\widehat{\Gamma}\boldsymbol{v} \geq \frac{\lambda_{\min}(\boldsymbol{\Sigma}_x)}{2}|\boldsymbol{v}|_2^2-\frac{c_1\log p}{n}|\boldsymbol{v}|_1^2. \tag{S3.9}$$

and the upper-RE condition

$$\boldsymbol{v}^\top\widehat{\Gamma}\boldsymbol{v} \leq \frac{3}{2}\lambda_{\max}(\boldsymbol{\Sigma}_x)|\boldsymbol{v}|_2^2+\frac{c_2\log p}{n}|\boldsymbol{v}|_1^2. \tag{S3.10}$$

Furthermore, under the conditions of Proposition 1, inequalities (S3.9) *and* (S3.10) *hold for both $\widehat{\boldsymbol{\Sigma}}^{(0)}$ and $\widehat{\boldsymbol{\Sigma}}^{\mathcal{A}_h}$ with probability at least $1-c_1p^{-2}$, where n is replaced by n_0 and $n_{\mathcal{A}_h}$, respectively.*

Proof of Lemma 6: We only prove the case of $\widehat{\mathbf{\Sigma}}^{\mathcal{A}_h}$, as the results for $\widehat{\mathbf{\Sigma}}^{(0)}$ can be established in a similar manner. Without loss of generality, we assume that $\mathcal{A}_h = \{1, \ldots, |\mathcal{A}_h|\}$. Define $X^{\mathcal{A}_h} \in \mathbb{R}^{n_{\mathcal{A}_h} \times p}$ as the design matrix obtained by stacking $X^{(k)}$, $k \in \mathcal{A}_h$, row-wise. Next, for any vector $\boldsymbol{v}$ with $|\boldsymbol{v}|_2 = 1$ and any constant $t > 0$, we control the term $\mathbb{P}(|\boldsymbol{v}^\top(\widehat{\mathbf{\Sigma}}^{\mathcal{A}_h} - \mathbf{\Sigma}^{\mathcal{A}_h})\boldsymbol{v}| \geq 2\pi M t)$. Define $\boldsymbol{y}^{\mathcal{A}_h} = X^{\mathcal{A}_h}\boldsymbol{v}$, then $\boldsymbol{y}^{\mathcal{A}_h} \in \mathbb{R}^{n_{\mathcal{A}_h}}$ is the mean zero Gaussian random vector with the covariance $Q^{\mathcal{A}_h} = \mathrm{diag}(Q^1, \ldots, Q^{|\mathcal{A}_h|})$, where

$$Q^k_{i,j} = \boldsymbol{v}^\top \mathrm{cov}(\boldsymbol{x}^{(k)}_i, \boldsymbol{x}^{(k)}_j)\boldsymbol{v}.$$

Hence, we have

$$\boldsymbol{v}^\top(\widehat{\mathbf{\Sigma}}^{\mathcal{A}_h})\boldsymbol{v} = \boldsymbol{y}^{\mathcal{A}_h\top}\boldsymbol{y}^{\mathcal{A}_h}/n_{\mathcal{A}_h} = \boldsymbol{z}^{\mathcal{A}_h\top}Q^{\mathcal{A}_h}\boldsymbol{z}^{\mathcal{A}_h}/n_{\mathcal{A}_h}, \quad v^\top\mathbf{\Sigma}^{\mathcal{A}_h}v = \mathbb{E}(\boldsymbol{z}^{\mathcal{A}_h\top}Q^{\mathcal{A}_h}\boldsymbol{z}^{\mathcal{A}_h})/n_{\mathcal{A}_h},$$

where $\boldsymbol{z}^{\mathcal{A}_h} \sim N(\mathbf{0}, \mathbf{I}_{n_{\mathcal{A}_h}})$. So, by the Hanson-Wright inequality of Rudelson and Vershynin (2013), with $\|z_i^{\mathcal{A}_h}\|_{\psi_2} \leq 1$ since $z_i \sim N(0,1)$, we get

$$\begin{aligned}\mathbb{P}[|\boldsymbol{v}^\top(\widehat{\mathbf{\Sigma}}^{\mathcal{A}_h} - \mathbf{\Sigma}^{\mathcal{A}_h})\boldsymbol{v}| > t] &= \mathbb{P}[|\boldsymbol{z}^{\mathcal{A}_h\top}Q^{\mathcal{A}_h}\boldsymbol{z}^{\mathcal{A}_h} - \mathrm{E}[\boldsymbol{z}^{\mathcal{A}_h\top}Q^{\mathcal{A}_h}\boldsymbol{z}]| > n_{\mathcal{A}_h}t] \\ &\leq 2\exp\Big[-cn_{\mathcal{A}_h}\min\Big\{\frac{n^2_{\mathcal{A}_h}t^2}{|Q^{\mathcal{A}_h}|_F^2}, \frac{n_{\mathcal{A}_h}t}{|Q^{\mathcal{A}_h}|_2}\Big\}\Big].\end{aligned}$$

Since $|Q^{\mathcal{A}_h}|_F^2/n_{\mathcal{A}_h} \leq |Q^{\mathcal{A}_h}|_2^2$, setting $t = |Q^{\mathcal{A}_h}|_2 t_1$, we obtain

$$\mathbb{P}[|\boldsymbol{v}^\top(\widehat{\mathbf{\Sigma}}^{\mathcal{A}_h} - \mathbf{\Sigma}^{\mathcal{A}_h})\boldsymbol{v}| > t_1|Q^{\mathcal{A}_h}|_2] \leq 2\exp(-cn_{\mathcal{A}_h}\min\{t_1, t_1^2\}).$$

By the similarly arguments as in Basu and Michailidis (2015), we can get that $|Q^{\mathcal{A}_h}|_2 \leq 2\pi M$, where $M < \infty$ is a constant since $\sum_{i=1}^\infty |A_i^{(k)}|_F < \infty$, for all $k \in \mathcal{A}_h$. Then applying Supplementary Lemma F.2 in Basu and Michailidis (2015), for any integer $s \geq 1$, we have

$$\mathbb{P}\Big[\sup_{\boldsymbol{v}\in\mathcal{K}(s)} |\boldsymbol{v}^\top(\widehat{\mathbf{\Sigma}}^{\mathcal{A}_h} - \mathbf{\Sigma}^{\mathcal{A}_h})\boldsymbol{v}| > 2\pi M t] \leq 2\exp\Big[-cn_{\mathcal{A}_h}\min\{t, t^2\} + s\log p\Big].$$

Then set $t = \lambda_{\min}(\mathbf{\Sigma}^{\mathcal{A}_h})/(108\pi M)$, and $s = c_1 n_{\mathcal{A}_h}/\log p$ for some sufficient small constant

c_1, we get

$$\mathbb{P}[\sup_{v\in\mathcal{K}(s)}|v(\widehat{\mathbf{\Sigma}}^{\mathcal{A}_h}-\mathbf{\Sigma}^{\mathcal{A}_h})v|>\lambda_{\min}(\mathbf{\Sigma}^{\mathcal{A}_h})/54]\leq 2\exp[-cn_{\mathcal{A}_h}].$$

Hence, applying supplementary Lemma 13 in Loh and Wainwright (2012), we can get the desired results.

Proof of Proposition 1: Combining Lemma 6 with arguments similar to those used in the proofs of Theorem 1 in Li et al. (2022) and Theorem 1, it suffices to show that one can choose $\lambda_w \asymp \sqrt{\log p/n_{\mathcal{A}_h}}$ and $\lambda_\delta \asymp \sqrt{\log p/n_0}$. Based on Propositions 3.2 and 2.4 in Basu and Michailidis (2015), we obtain that for all $k \in \mathcal{A}_h \cup \{0\}$,

$$\mathbb{P}\Big(\frac{1}{n_k}|(X^{(k)})^\top\boldsymbol{\epsilon}^{(k)}|\geq t\Big)\leq c_1 p\exp(-c_2 n_k\min\{t,t^2\}),$$

$$\mathbb{P}\Big(\sup_{j\leq p}\Big|\frac{1}{n_k}\sum_{m\leqslant n_k}(X^{(k)}_{mj}(\boldsymbol{x}^{(k)}_m)^\top\boldsymbol{\delta}^{(k)}-\Sigma^{(k)}_{j,\cdot}\boldsymbol{\delta}^{(k)})\Big|\geqslant t\Big)\leq c_1 p\exp(-c_2 n_k\min\{t,t^2\}).$$

$$\mathbb{P}\Big(\sup_{j\leq p}\Big|\frac{1}{n_k}\sum_{m\leqslant n_k}(X^{(k)}_{mj}(\boldsymbol{x}^{(k)}_m)^\top\boldsymbol{\delta}^{\mathcal{A}_h}-\Sigma^{(k)}_{j,\cdot}\boldsymbol{\delta}^{\mathcal{A}_h})\Big|\geqslant t\Big)\leq c_1 p\exp(-c_2 n_k\min\{t,t^2\}).$$

Then by the similar arguments as in the proof of Lemma 5 and Theorem 1, it is sufficient to take $\lambda_w \asymp \sqrt{\log p/n_{\mathcal{A}_h}}$ and $\lambda_\delta \asymp \sqrt{\log p/n_0}$.

S4 Proof of Theorem 2

Recall that $\epsilon_i = g(\ldots,\xi_{i-1},\xi_i) = g(\mathcal{F}_i)$ in (2.2), define

$$\epsilon_i^* := \epsilon_{i,m} = \mathbb{E}(\boldsymbol{\epsilon}_i \mid \xi_{i-m},\ldots,\xi_i) = \mathbb{E}(\epsilon_i \mid \mathcal{F}_{i-m,i}),\quad m\geq 0,$$

where $\mathcal{F}_{i-m,i} = (\xi_{i-m},\ldots,\xi_i)$ is a σ algebra. Then $\{\epsilon_i^*\}$ are m-dependent random variables with mean zero.

For convenience and without causing ambiguity, we will abbreviate n_0 as n, X^0 as X, $\boldsymbol{\epsilon}^{(0)}$ as $\boldsymbol{\epsilon}$ and $\mathbf{\Sigma}^{(0)}$ as $\mathbf{\Sigma}$ in the following text.

Lemma 7. *Suppose that* $\Delta_{0,q} < \infty$ *for* $q \geq 2$. *Let* $a_1, a_2, \ldots, \in \mathbb{R}$, $A_n = (\sum_{i=1}^n a_i^2)^{1/2}$, *and* $C_q = 18q^{3/2}(q-1)^{-1/2}$. *Then (i)* $\| \sum_{i=1}^n a_i \boldsymbol{\epsilon}_i \|_q \leqslant C_q A_n \Delta_{0,q}$, *(ii)* $\| \sum_{i=1}^n a_i \boldsymbol{\epsilon}_i^* \|_q \leqslant C_q A_n \Delta_{0,q}$ *and (iii)* $\| \sum_{i=1}^n a_i (\boldsymbol{\epsilon}_i - \boldsymbol{\epsilon}_i^*) \|_q \leqslant C_q A_n \Delta_{m+1,q}$.

Proof: This result can be proved by using the arguments in Liu and Wu (2010) Lemma 1.

Lemma 8. *Suppose* $\Delta_{0,q} < \infty$ *for* $2 < q \leq 4$. *Let* $d = q/2$, *then for any* $j \in \mathbb{Z}$,

$$\Big\| \sum_{i=1}^n \boldsymbol{\epsilon}_i \boldsymbol{\epsilon}_{i+j} - n\gamma_j^{\boldsymbol{\epsilon}} \Big\|_d \leqslant 2B_d n^{1/d} \|\boldsymbol{\epsilon}_1\|_q \Delta_{0,q},$$

where

$$B_d = \begin{cases} \frac{18d^{3/2}}{(d-1)^{1/2}}, & \text{if} \quad d \neq 2, \\ 1, & \text{if} \quad d = 2. \end{cases}$$

Proof: See Lemma 1 of Wu and Pourahmadi (2009).

Lemma 9. *Under the conditions of Theorem 2, for all* $1 \leq i \leq p$, $\boldsymbol{\Sigma}_{\cdot,i}^{-1}$ *satisfies the constraints in* (2.6) *with probability tending to 1.*

Proof: By Condition 1, it holds that $|\boldsymbol{\Sigma}_{\cdot,i}^{-1}|_2^2 \leq 1/\lambda_{\min}^2(\boldsymbol{\Sigma}) \leq c_1$, then by the similar arguments as in the proof of Lemma 5, we have:

$$\mathbb{P}(|\widehat{\boldsymbol{\Sigma}}\boldsymbol{\Sigma}_{\cdot,i}^{-1} - \boldsymbol{e}_i|_\infty \geq \lambda_\delta) = \begin{cases} O((\log p)^{-1}), & \text{case (i)}, \\ O(p^{-2}), & \text{case (ii)}, \end{cases}$$

for all $1 \leq i \leq p$. Regarding the second constraint in (2.6), for case (i), we have

$$\begin{aligned} \mathbb{P}(|X\boldsymbol{\Sigma}_{\cdot,i}^{-1}|_\infty \geq c_\gamma n^r) &\leq n\mathbb{P}(|\boldsymbol{x}_j^\top \boldsymbol{\Sigma}_{\cdot,i}^{-1}| \geq c_\gamma n^r) \\ &\lesssim n\Phi_{0,\iota}^\iota / n^{r\iota} \\ &\lesssim n^{1-r\iota}, \end{aligned}$$

where the penultimate inequality follows from Markov's inequality. Therefore, in case (i), with probability at least $1 - c_1 \log(p)^{-1} - c_2 n^{1-r\iota}$, $\mathbf{\Sigma}_{\cdot,i}^{-1}$ satisfies the constraints in (2.6). Similarly, in case (ii), for any $q > 1/r$, with probability at least $1 - c_1 n^{1-rq}$, $\mathbf{\Sigma}_{\cdot,i}^{-1}$ satisfies the constraints in (2.6).

Lemma 10. *Assume the conditions of Theorem 2 holds, for some positive constants c_0, c_1. we have that*

$$c_0 - o_{\mathbb{P}}(1) \leqslant \widehat{\mathbf{\Theta}}_j^\top \widehat{\mathbf{\Sigma}} \widehat{\mathbf{\Theta}}_j \leqslant c_1 + o_{\mathbb{P}}(1),$$

where $\widehat{\mathbf{\Theta}}_j$ is obtained from (2.6).

Proof: As $1 - \boldsymbol{e}_j^\top \widehat{\mathbf{\Sigma}} \widehat{\Theta}_j \leq c_\Theta \lambda_\delta = \lambda$, for any $c \geq 0$,

$$\begin{aligned}
\widehat{\Theta}_j^\top \widehat{\mathbf{\Sigma}} \widehat{\Theta}_j &\geq \widehat{\Theta}_j^\top \widehat{\mathbf{\Sigma}} \widehat{\Theta}_j + c(1 - \lambda - \boldsymbol{e}_j^\top \widehat{\mathbf{\Sigma}} \widehat{\Theta}_j) \\
&\geq \min_{\boldsymbol{w}} \boldsymbol{w}^\top \widehat{\mathbf{\Sigma}} \boldsymbol{w} + c(1 - \lambda - \boldsymbol{e}_j^\top \widehat{\mathbf{\Sigma}} \boldsymbol{w}) \\
&= c(1 - \lambda) - \frac{c^2}{4} \widehat{\mathbf{\Sigma}}_{j,j}.
\end{aligned}$$

Optimizing this bound over c, we have

$$\widehat{\mathbf{\Theta}}_j^\top \widehat{\mathbf{\Sigma}} \widehat{\mathbf{\Theta}}_j \geq \frac{(1-\lambda)^2}{\widehat{\mathbf{\Sigma}}_{j,j}}.$$

By the arguments as in Lemma 4, we can easily arrive at $\widehat{\mathbf{\Sigma}}_{j,j} \leq \mathbf{\Sigma}_{j,j} + o_{\mathbb{P}}(1) \leq \lambda_{max}(\mathbf{\Sigma}) + o_{\mathbb{P}}(1)$. Based on Condition 1, We have proven the lower bound.

Define that

$$\lambda_{xx}^0 = \begin{cases} N_X^2 \max\{n_{\mathcal{A}_h}^{2\chi/\iota - 1} (p \log p)^{2/\iota}, \sqrt{\log p / n_0}\}, & \text{case (i)}, \\ L_\varrho^2 (\log p)^{(1+4\varrho)/2} / \sqrt{n_0}, & \text{case (ii)}. \end{cases}$$

For the upper bound, by Lemma 9, $\mathbf{\Sigma}_{\cdot,i}^{-1}$ satisfies the constraints in (2.6) with probability at least $1 - o(1)$, then on this event we have,

$$\widehat{\mathbf{\Theta}}_j^\top \widehat{\mathbf{\Sigma}} \widehat{\mathbf{\Theta}}_j \leq \mathbf{\Sigma}_{j,\cdot}^{-1} \widehat{\mathbf{\Sigma}} \mathbf{\Sigma}_{\cdot,j}^{-1} = (\mathbf{\Sigma}_{j,\cdot}^{-1} \widehat{\mathbf{\Sigma}} \mathbf{\Sigma}_{\cdot,j}^{-1} - \mathbf{\Sigma}_{jj}^{-1}) + \mathbf{\Sigma}_{j,j}^{-1} = \frac{1}{n} \sum_{i=1}^{n} (V_i^2 - \Sigma_{jj}^{-1}) + \Sigma_{j,j}^{-1},$$

where $V_i = \boldsymbol{\Sigma}_{j,\cdot}^{-1}\boldsymbol{x}_i$, with $\mathbb{E}(V_i^2) = \boldsymbol{\Sigma}_{jj}^{-1}$. Then, we have,

$$\mathbb{P}\Big(\widehat{\boldsymbol{\Theta}}_j^\top \widehat{\boldsymbol{\Sigma}}\widehat{\boldsymbol{\Theta}}_j \geq \boldsymbol{\Sigma}_{jj}^{-1} + \lambda_{xx}^0\Big) \leq \mathbb{P}\Big(\frac{1}{n}\sum_{i=1}^n (V_i^2 - \boldsymbol{\Sigma}_{jj}^{-1}) \geq \lambda_{xx}^0\Big) + o(1)$$

$$\leq o(1).$$

Hence, we have obtained that $\widehat{\boldsymbol{\Theta}}_j^\top \widehat{\boldsymbol{\Sigma}}\widehat{\boldsymbol{\Theta}}_j \leq c_1 + o_{\mathbb{P}}(1)$.

Proof of Theorem 2.

$$|\Lambda_j| = \sqrt{n}|(\boldsymbol{e}_j^\top - \widehat{\boldsymbol{\Theta}}_j^\top \widehat{\boldsymbol{\Sigma}})(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*)| \leqslant \sqrt{n}|(\boldsymbol{e}_j^\top - \widehat{\boldsymbol{\Theta}}_j^\top \widehat{\boldsymbol{\Sigma}})|_\infty |\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*|_1.$$

By the first constraint in (2.6), we get

$$|\Lambda_j| \leq C\sqrt{n}\lambda_\delta |\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*|_1.$$

For case (i) in Theorem 1, by Theorem 1 (i),

$$\mathbb{P}(|\Lambda_k| \geq C\sqrt{n}(s_0\lambda_w\lambda_\delta + \lambda_\delta C_{\boldsymbol{\Sigma}} h + (C_{\boldsymbol{\Sigma}} h)^2)) \leqslant C_1(\log p)^{-1}.$$

For case (ii) in Theorem 1, then using Theorem 1 (ii),

$$\mathbb{P}(|\Lambda_k| \geq C\sqrt{n}(s_0\lambda_w\lambda_\delta + \lambda_\delta C_{\boldsymbol{\Sigma}} h + (C_{\boldsymbol{\Sigma}} h)^2)) \leqslant C_1 p^{-2}.$$

For $R_{0,j}$, we prove its asymptotic normality. The following proof is applicable to both situations considered in Theorem 2, and we will use the m-dependent approximation to prove the asymptotic normality.

For some constants $2 < d < \infty$ and $2r < \eta < 1/2$, where r is defined in (2.6) of the main text, let $\zeta_n = \lfloor n^\eta/(\log n)^d \rfloor, t_n = \lfloor n^{1/2-\eta} \rfloor$ and $w_n = \lfloor n/\zeta_n - t_n \rfloor$. Define $\boldsymbol{\epsilon}^* = (\epsilon_1^*, \ldots, \epsilon_n^*)^\top$ with

$$\boldsymbol{\epsilon}_i^* := \epsilon_{i,t_n} = \mathbb{E}(\epsilon_i \mid \xi_{i-t_n}, \ldots, \xi_i) = \mathbb{E}(\epsilon_i \mid \mathcal{F}_{i-t_n,i}),$$

thus, $\{\epsilon_i^*\}$ are t_n-dependent sequence.

Define the event,

$$\mathbf{E_2} = \{c_0 \leq \widehat{\mathbf{\Theta}}_k^\top \widehat{\mathbf{\Sigma}} \widehat{\mathbf{\Theta}}_k \leq c_k\}.$$

By Lemma 10, there exist constants c_0 and c_k such that $\mathbb{P}(\mathbf{E_2}) \to 1$. Note that on the set $\mathbf{E_2}$, we have

$$\sigma_k^2 = |n^{-1}\widehat{\mathbf{\Theta}}_k^\top X^\top \mathbf{\Sigma}_n^{\boldsymbol{\epsilon}} X \widehat{\mathbf{\Theta}}_k| \geq n^{-1}|X\widehat{\mathbf{\Theta}}_k|_2^2 \cdot \lambda_{\min}(\mathbf{\Sigma}_n^{\boldsymbol{\epsilon}}) \geq c_0 \lambda_{\min}(\mathbf{\Sigma}_n^{\boldsymbol{\epsilon}}) = O(1),$$

$$\sigma_k^2 = |n^{-1}\widehat{\mathbf{\Theta}}_k^\top X^\top \mathbf{\Sigma}_n^{\boldsymbol{\epsilon}} X \widehat{\mathbf{\Theta}}_k| \leq n^{-1}|X\widehat{\mathbf{\Theta}}_k|_2^2 \cdot \lambda_{\max}(\mathbf{\Sigma}_n^{\boldsymbol{\epsilon}}) \leq c_k \lambda_{\max}(\mathbf{\Sigma}_n^{\boldsymbol{\epsilon}}) = O(1).$$

where $\lambda_{\min}(\mathbf{\Sigma}_n^{\boldsymbol{\epsilon}})$ and $\lambda_{\max}(\mathbf{\Sigma}_n^{\boldsymbol{\epsilon}})$ and the minimum and maximum eigenvalue of $\mathbf{\Sigma}_n^{\boldsymbol{\epsilon}}$, since the eigenvalues of $\mathbf{\Sigma}_n^{\epsilon} = \mathrm{Cov}(\boldsymbol{\epsilon})$ are bounded away from zero and infinity (see Section 5.2 in Grenander and Szegö (1958)). For any $k \in \{1, \ldots, p\}$,

$$\begin{aligned}
&n^{-1/2}\sum_{i=1}^n \widehat{\mathbf{\Theta}}_k^\top \boldsymbol{x}_i(\epsilon_i - \epsilon_i^*)/\sigma_k \\
&= n^{-1/2}\sum_{i=1}^n \widehat{\mathbf{\Theta}}_k^\top \boldsymbol{x}_i(\epsilon_i - \epsilon_i^*)\mathbb{I}(\mathbf{E_2})/\sigma_k + n^{-1/2}\sum_{i=1}^n \widehat{\mathbf{\Theta}}_k^\top \boldsymbol{x}_i(\epsilon_i - \epsilon_i^*)\mathbb{I}(\mathbf{E}_2^c)/\sigma_k.
\end{aligned}$$

It is obviously to obtain that

$$n^{-1/2}\sum_{i=1}^n \widehat{\mathbf{\Theta}}_k^\top \boldsymbol{x}_i(\epsilon_i - \epsilon_i^*)\mathbb{I}(\mathbf{E}_2^c)/\sigma_k = o_{\mathbb{P}}(1),$$

since $\mathbb{P}(\mathbf{E_2}) \to 1$, as $n \to \infty$. According to Lemma 7 (iii)

$$\begin{aligned}
&\mathrm{Var}\Big\{n^{-1/2}\sum_{i=1}^n \widehat{\mathbf{\Theta}}_k^\top \boldsymbol{x}_i(\epsilon_i - \epsilon_i^*)\mathbb{I}(\mathbf{E_2})/\sigma_k\Big\} \\
=&\mathbb{E}\Big(\mathrm{Var}\Big\{n^{-1/2}\sum_{i=1}^n \widehat{\mathbf{\Theta}}_k^\top \boldsymbol{x}_i(\epsilon_i - \epsilon_i^*)\mathbb{I}(\mathbf{E_2})/\sigma_k\Big|\{\boldsymbol{x}_i\}_{i=1}^n\Big\}\Big) \\
&\leq C\mathbb{E}(|X\widehat{\mathbf{\Theta}}_k\mathbb{I}(\mathbf{E_2})|_2^2/n\Delta_{t_n+1,2}^2) \\
&\leq O(\Delta_{t_n+1,2}^2) = o(1).
\end{aligned}$$

Hence, $n^{-1/2}\sum_{i=1}^n \widehat{\mathbf{\Theta}}_k^\top \boldsymbol{x}_i(\epsilon_i - \epsilon_i^*)/\sigma_k = o_{\mathbb{P}}(1)$, and show

$$V_k = \frac{1}{\sqrt{n}\sigma_k}\sum_{i=1}^n \widehat{\mathbf{\Theta}}_k^\top \boldsymbol{x}_i\epsilon_i \xrightarrow{D} N(0,1),$$

is equivalent to show

$$V_k^* = \frac{1}{\sqrt{n}\sigma_k}\sum_{i=1}^{n}\widehat{\boldsymbol{\Theta}}_k^\top \boldsymbol{x}_i\epsilon_i^* \xrightarrow{D} N(0,1).$$

Denote by $\boldsymbol{v} = n^{-1/2}\widehat{\boldsymbol{\Theta}}_k^\top X^\top/\sigma_k = (v_1,\ldots,v_n)$, where $v_j = n^{-1/2}\widehat{\boldsymbol{\Theta}}_k^\top \boldsymbol{x}_j/\sigma_k$. Define

$$\boldsymbol{\Xi}_i = \sum_{j=(i-1)(w_n+t_n)+1}^{(i-1)(w_n+t_n)+w_n} v_j\epsilon_j^*, \quad i = 1,\ldots,\zeta_n.$$

$$\boldsymbol{\Xi}_i' = \begin{cases} \displaystyle\sum_{j=(i-1)(w_n+t_n)+w_n+1}^{i(w_n+t_n)} v_j\epsilon_j^*, & i = 1,\ldots,\zeta_n-1, \\ \displaystyle\sum_{j=(i-1)(w_n+t_n)+w_n+1}^{n} v_j\epsilon_j^*, & i = \zeta_n. \end{cases}$$

Therefore

$$\sum_{i=1}^{\zeta_n}\boldsymbol{\Xi}_i + \sum_{i=1}^{\zeta_n}\boldsymbol{\Xi}_i' = \boldsymbol{v}\boldsymbol{\epsilon}^* = V_k^*.$$

For n large enough, $\{\boldsymbol{\Xi}_i\}_{i=1}^{\zeta_n}$ are independent and $\{\boldsymbol{\Xi}_i'\}_{i=1}^{\zeta_n}$ are independent when given $\{\boldsymbol{x}_i\}_{i=1}^n$, since $\boldsymbol{\epsilon}^*$ are t_n-dependent. Combining with that $\mathbb{P}(\mathbf{E}_2)\to 1$, we first show that $V_k^*\mathbb{I}(\mathbf{E_2}) \xrightarrow{D} N(0,1)$. And in the following, we will prove that,

$$\sum_{i=1}^{\zeta_n}\boldsymbol{\Xi}_i\mathbb{I}(\mathbf{E_2}) \xrightarrow{D} N(0,1) \text{ and } \sum_{i=1}^{\zeta_n}\boldsymbol{\Xi}_i'\mathbb{I}(\mathbf{E_2}) = o_{\mathbb{P}}(1).$$

On the set $\mathbf{E}_2$, we also have

$$\max_j |v_j\mathbb{I}(\mathbf{E_2})| = |n^{-1/2}\sigma_k^{-1}\widehat{\boldsymbol{\Theta}}_k^\top X^\top\mathbb{I}(\mathbf{E_2})|_\infty \le Cn^{-1/2}|X\widehat{\boldsymbol{\Theta}}_k|_\infty$$

$$\le C_1 n^{-1/2}n^r = O(n^{-1/2+r}),$$

and

$$|\mathbf{v}\mathbb{I}(\mathbf{E_2})|_2 = |n^{-1/2}\sigma_k^{-1}\widehat{\boldsymbol{\Theta}}_k^\top X^\top\mathbb{I}(\mathbf{E_2})|_2 \le Cn^{-1/2}|X\widehat{\boldsymbol{\Theta}}_k\mathbb{I}(\mathbf{E_2})|_2 \le C_1.$$

Therefore,

$$\mathrm{Var}\Big(\sum_{i=1}^{\zeta_n}\boldsymbol{\Xi}_i'\mathbb{I}(\mathbf{E_2})\Big|\{\boldsymbol{x}_i\}_{i=1}^n\Big) = \sum_{i=1}^{\zeta_n}\mathrm{Var}(\boldsymbol{\Xi}_i'\mathbb{I}(\mathbf{E_2})|\{\boldsymbol{x}_i\}_{i=1}^n)$$

$$
\begin{aligned}
&\leq C\zeta_n \mathrm{Var}\Big(\sum_{j=w_n+1}^{w_n+t_n} v_j \mathbb{I}(\mathbf{E_2})\epsilon_j^* \Big| \{\boldsymbol{x}_i\}_{i=1}^n\Big)\\
&\leq C\zeta_n t_n^2 \mathrm{Var}(v_j \mathbb{I}(\mathbf{E_2})\epsilon_j^* | \{\boldsymbol{x}_i\}_{i=1}^n)\\
&\leq C\frac{\zeta_n \mathbb{I}(\mathbf{E_2}) t_n^2}{n^{1-2r}} \mathrm{Var}(\epsilon_j^*)\\
&= O\Big(\frac{\mathbb{I}(\mathbf{E_2})\zeta_n t_n^2}{n^{1-2r}}\Big)\\
&= o\Big(\frac{n^{2r}}{n^{\eta}(\log n)^d}\mathbb{I}(\mathbf{E_2})\Big)\\
&= o(\mathbb{I}(\mathbf{E_2})).
\end{aligned}
$$

Hence, we have $\mathrm{Var}\Big(\sum_{i=1}^{\zeta_n} \boldsymbol{\Xi}_i' \mathbb{I}(\mathbf{E_2})\Big) = \mathrm{E}\Big(\mathrm{Var}\Big(\sum_{i=1}^{\zeta_n} \boldsymbol{\Xi}_i' \mathbb{I}(\mathbf{E_2}) \Big| \{\boldsymbol{x}_i\}_{i=1}^n\Big)\Big) + \mathrm{Var}\Big(\mathrm{E}\Big(\sum_{i=1}^{\zeta_n} \boldsymbol{\Xi}_i' \mathbb{I}(\mathbf{E_2}) \Big| \{\boldsymbol{x}_i\}_{i=1}^n\Big)\Big)$ $o(1)$, which together with $\mathbb{E}(\sum_{i=1}^{\zeta_n} \boldsymbol{\Xi}_i' \mathbb{I}(\mathbf{E_2})) = 0$ implies $\sum_{i=1}^{\zeta_n} \boldsymbol{\Xi}_i' \mathbb{I}(\mathbf{E_2}) = o_{\mathbb{P}}(1)$. By Lemma 7 (iii), we have

$$
\mathrm{Var}\{\mathbf{v}(\boldsymbol{\epsilon} - \boldsymbol{\epsilon}^*)\mathbb{I}(\mathbf{E_2}) | \{\boldsymbol{x}_i\}_{i=1}^n\} \leq C|\mathbf{v}\mathbb{I}(\mathbf{E_2})|_2^2 \Delta_{t_n+1,2}^2 \leq O(\Delta_{t_n+1,2}^2) = o(1), \quad \text{as} \quad n \to \infty.
$$

Then, we can get,

(i) If $\mathbb{I}(\mathbf{E_2}) = 1$ when given $\{\boldsymbol{x}_i\}_{i=1}^n$,

$$
\begin{aligned}
&\lim_{n\to\infty} \mathrm{Var}\Big\{\sum_{i=1}^{\zeta_n} \boldsymbol{\Xi}_i \mathbb{I}(\mathbf{E_2}) + \sum_{i=1}^{\zeta_n} \boldsymbol{\Xi}_i' \mathbb{I}(\mathbf{E_2}) \Big| \{\boldsymbol{x}_i\}_{i=1}^n\Big\}\\
=&\lim_{n\to\infty} \mathrm{Var}(\mathbf{v}\boldsymbol{\epsilon}^* \mathbb{I}(\mathbf{E_2}) | \{\boldsymbol{x}_i\}_{i=1}^n)\\
=&\lim_{n\to\infty} \mathrm{Var}(\mathbf{v}\boldsymbol{\epsilon} \mathbb{I}(\mathbf{E_2}) | \{\boldsymbol{x}_i\}_{i=1}^n)\\
=&1,
\end{aligned}
$$

implying that $\lim_{n\to\infty} \mathrm{Var}(\sum_{i=1}^{\zeta_n} \boldsymbol{\Xi}_i \mathbb{I}(\mathbf{E_2}) | \{\boldsymbol{x}_i\}_{i=1}^n) = 1$, since $\lim_{n\to\infty} \mathrm{Var}(\sum_{i=1}^{\zeta_n} \boldsymbol{\Xi}_i' \mathbb{I}(\mathbf{E_2}) | \{\boldsymbol{x}_i\}_{i=1}^n) = 0$. According to Lemma 7(ii), the Liapounov condition follows

$$
\frac{1}{\mathrm{Var}(\sum_{i=1}^{\zeta_n} \boldsymbol{\Xi}_i \mathbb{I}(\mathbf{E_2}) | \{\boldsymbol{x}_i\}_{i=1}^n)^{1+c/2}} \sum_{i=1}^{\zeta_n} \mathbb{E}(|\boldsymbol{\Xi}_i \mathbb{I}(\mathbf{E_2})|^{2+c} | \{\boldsymbol{x}_i\}_{i=1}^n)
$$

$$
\begin{aligned}
\leq & \frac{C}{\mathrm{Var}(\sum_{i=1}^{\zeta_n} \boldsymbol{\Xi}_i \mathbb{I}(\mathbf{E_2}) | \{\boldsymbol{x}_i\}_{i=1}^n)^{1+c/2}} \sum_{i=1}^{\zeta_n} \Big\{ \sum_{j=(i-1)(w_n+t_n)+1}^{(i-1)(w_n+t_n)+w_n} (v_j \mathbb{I}(\mathbf{E_2}))^2 \Big\}^{1+c/2} \Delta_{0,2+c}^{2+c} \\
= & \frac{C}{\mathrm{Var}(\sum_{i=1}^{\zeta_n} \boldsymbol{\Xi}_i \mathbb{I}(\mathbf{E_2}) | \{\boldsymbol{x}_i\}_{i=1}^n)^{1+c/2}} \frac{\zeta_n w_n^{1+c/2}}{n^{1+c/2}} \cdot (n^r)^{2+c} \\
= & O\Big(\frac{n^{2r+cr} (\log n)^{dc/2}}{n^{c\eta/2}} \Big) \\
= & o(1),
\end{aligned}
$$

for some constant $c > 0$. Hence, by central limit theorem,

$$
\sup_{\nu \in \mathbb{R}} \Big| \mathbb{P}\Big(\sum_{i=1}^{\zeta_n} \boldsymbol{\Xi}_i \mathbb{I}(\mathbf{E_2}) \leq \nu \Big| \{\boldsymbol{x}_i\}_{i=1}^n \Big) - \Phi(\nu) \Big| = o(1).
$$

(ii) If $\mathbb{I}(\mathbf{E_2}) = 0$ when given $\{\boldsymbol{x}_i\}_{i=1}^n$, we have

$$
\sup_{\nu \in \mathbb{R}} \Big| \mathbb{P}\Big(\sum_{i=1}^{\zeta_n} \boldsymbol{\Xi}_i \mathbb{I}(\mathbf{E_2}) \leq \nu \Big| \{\boldsymbol{x}_i\}_{i=1}^n \Big) - \Phi(\nu) \Big| \leq 1.
$$

Hence

$$
\begin{aligned}
& \int \sup_{\nu \in \mathbb{R}} \Big| \mathbb{P}\Big(\sum_{i=1}^{\zeta_n} \boldsymbol{\Xi}_i \mathbb{I}(\mathbf{E_2}) \leq \nu \Big| \{\boldsymbol{x}_i\}_{i=1}^n \Big) - \Phi(\nu) \Big| d\mathbb{F}(\{\boldsymbol{x}_i\}_{i=1}^n) \\
\leq & \int_{\mathbf{E_2}} \sup_{\nu \in \mathbb{R}} \Big| \mathbb{P}\Big(\sum_{i=1}^{\zeta_n} \boldsymbol{\Xi}_i \mathbb{I}(\mathbf{E_2}) \leq \nu \Big| \{\boldsymbol{x}_i\}_{i=1}^n \Big) - \Phi(\nu) \Big| d\mathbb{F}(\{\boldsymbol{x}_i\}_{i=1}^n) \\
+ & \int_{\mathbf{E}_2^c} \sup_{\nu \in \mathbb{R}} \Big| \mathbb{P}\Big(\sum_{i=1}^{\zeta_n} \boldsymbol{\Xi}_i \mathbb{I}(\mathbf{E_2}) \leq \nu \Big| \{\boldsymbol{x}_i\}_{i=1}^n \Big) - \Phi(\nu) \Big| d\mathbb{F}(\{\boldsymbol{x}_i\}_{i=1}^n) \\
\leq & o(1) + \mathbb{P}(\mathbf{E}_2^c) = o(1),
\end{aligned}
$$

which implies that

$$
\sup_{\nu \in \mathbb{R}} \Big| \mathbb{P}\Big(\sum_{i=1}^{\zeta_n} \boldsymbol{\Xi}_i \mathbb{I}(\mathbf{E_2}) \leq \nu \Big) - \Phi(\nu) \Big| = o(1).
$$

This is because

$$\begin{aligned}
&\sup_{\nu\in\mathbb{R}}\Big|\mathbb{P}\Big(\sum_{i=1}^{\zeta_n}\boldsymbol{\Xi}_i\mathbb{I}(\mathbf{E_2})\le\nu\Big)-\Phi(\nu)\Big|\\
=&\sup_{\nu\in\mathbb{R}}\Big|\int\mathbb{P}\Big(\sum_{i=1}^{\zeta_n}\boldsymbol{\Xi}_i\mathbb{I}(\mathbf{E_2})\le\nu\Big|\{\boldsymbol{x}_i\}_{i=1}^n\Big)d\mathbb{F}(\{\boldsymbol{x}_i\}_{i=1}^n)-\Phi(\nu)\Big|\\
\le&\sup_{\nu\in\mathbb{R}}\int\Big|\mathbb{P}\Big(\sum_{i=1}^{\zeta_n}\boldsymbol{\Xi}_i\mathbb{I}(\mathbf{E_2})\le\nu\Big|\{\boldsymbol{x}_i\}_{i=1}^n\Big)-\Phi(\nu)\Big|d\mathbb{F}(\{\boldsymbol{x}_i\}_{i=1}^n)\\
\le&\int\sup_{\nu\in\mathbb{R}}\Big|\mathbb{P}\Big(\sum_{i=1}^{\zeta_n}\boldsymbol{\Xi}_i\mathbb{I}(\mathbf{E_2})\le\nu\Big|\{\boldsymbol{x}_i\}_{i=1}^n\Big)-\Phi(\nu)\Big|d\mathbb{F}(\{\boldsymbol{x}_i\}_{i=1}^n).
\end{aligned}$$

Combing with $\sum_{i=1}^{\zeta_n}\boldsymbol{\Xi}_i'\mathbb{I}(\mathbf{E_2})=o_{\mathbb{P}}(1)$, we obtain that

$$V_k^*\mathbb{I}(\mathbf{E_2})\xrightarrow{D}N(0,1).$$

Using the fact that $\mathbb{P}(X\leqslant t)=\mathbb{P}(X\leqslant t,A)+\mathbb{P}(X\leqslant t|A^c)\mathbb{P}(A^c)$ and $\mathbb{P}(\mathbf{E_2})\to 1$, we obtain

$$V_k^*=\frac{1}{\sqrt{n}\sigma_k}\sum_{i=1}^n\widehat{\boldsymbol{\Theta}}_k^\top\boldsymbol{x}_i\epsilon_i^*\xrightarrow{D}N(0,1),$$

which implies that

$$V_k=\frac{1}{\sqrt{n}\sigma_k}\sum_{i=1}^n\widehat{\boldsymbol{\Theta}}_k^\top\boldsymbol{x}_i\epsilon_i\xrightarrow{D}N(0,1),$$

and the proofs are completed.

S5 Proof of Theorem 3

Proof of Lemma 1: By using the fact that for any matrix $\mathbf{A}\in\mathbb{R}^{m\times n}$,

$$|\mathbf{A}|_2^2\le|\mathbf{A}|_1|\mathbf{A}|_\infty,$$

we have $|\widehat{\boldsymbol{\Sigma}}^{\boldsymbol{\epsilon}}_{n,l} - \boldsymbol{\Sigma}^{\boldsymbol{\epsilon}}_n|_2 \le |\widehat{\boldsymbol{\Sigma}}^{\boldsymbol{\epsilon}}_{n,l} - \boldsymbol{\Sigma}^{\boldsymbol{\epsilon}}_n|_1$, since $\widehat{\boldsymbol{\Sigma}}^{\boldsymbol{\epsilon}}_{n,l} - \boldsymbol{\Sigma}^{\boldsymbol{\epsilon}}_n$ is symmetric. Thus

$$\begin{aligned}
|\widehat{\boldsymbol{\Sigma}}^{\boldsymbol{\epsilon}}_{n,l} - \boldsymbol{\Sigma}^{\boldsymbol{\epsilon}}_n|_2 &\le \max_{1\le j\le n} \sum_{i=1}^{n} |\widehat{\gamma}^{\boldsymbol{\epsilon}}_{i-j}\mathbb{I}_{|i-j|\le l} - \gamma^{\boldsymbol{\epsilon}}_{i-j}| \\
&\le \sum_{i=1-n}^{n-1} |\widehat{\gamma}^{\boldsymbol{\epsilon}}_{i}\mathbb{I}_{|i|\le l} - \gamma^{\boldsymbol{\epsilon}}_{i}| \\
&\le 2\sum_{i=0}^{l} |\widehat{\gamma}^{\boldsymbol{\epsilon}}_{i} - \gamma^{\boldsymbol{\epsilon}}_{i}| + 2\sum_{i=l+1}^{n} |\gamma^{\boldsymbol{\epsilon}}_{i}| \\
&:= T_1 + T_2.
\end{aligned}$$

Note that for $i \ge 0$,

$$\begin{aligned}
&\Big|\frac{1}{n}\sum_{j=1}^{n-i} e_j e_{j+i} - \gamma^{\boldsymbol{\epsilon}}_i\Big| = \Big|\frac{1}{n}\sum_{j=1}^{n-i}[\epsilon_j - \boldsymbol{x}_j^\top(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*)][\epsilon_{j+i} - \boldsymbol{x}_{j+i}^\top(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*)] - \gamma^{\boldsymbol{\epsilon}}_i\Big| \\
&\le \frac{1}{n}\Big|\sum_{j=1}^{n-i}\epsilon_j\epsilon_{j+i} - (n-i)\gamma^{\boldsymbol{\epsilon}}_i\Big| + \frac{1}{n}\Big|\sum_{j=1}^{n-i}[\epsilon_j\boldsymbol{x}_{j+i}^\top + \epsilon_{j+i}\boldsymbol{x}_j^\top](\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*)\Big| \\
&+ \frac{1}{n}\sum_{j=1}^{n-i}|\boldsymbol{x}_j^\top(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*)\cdot\boldsymbol{x}_{j+i}^\top(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*)| + \frac{i}{n}|\gamma^{\boldsymbol{\epsilon}}_i| \\
&:= D_1 + D_2 + D_3 + D_4. \qquad \text{(S5.1)}
\end{aligned}$$

Using Lemma 8, there exists a constant $c_{q'}$ depending only on q' with q' defined in Lemma 1, such that

$$\|D_1\|_{q'/2} \le c_{q'}\|\epsilon_1\|_{q'}\Delta_{0,q'}\frac{(n-i)^{2/q'}}{n} \le Cn^{2/q'-1},$$

thus $D_1 = O_{\mathbb{P}}(n^{2/q'-1})$.

Note that

$$\begin{aligned}
\frac{1}{n}\Big|\sum_{j=1}^{n-i}\epsilon_j\boldsymbol{x}_{i+j}^\top(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*)\Big| &\le \frac{1}{n}\Big|\sum_{j=1}^{n-i}\epsilon_j\boldsymbol{x}_{i+j}\Big|_\infty|\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*|_1 \\
&= \max_{k\le p}\frac{1}{n}\Big|\sum_{j=1}^{n-i}\epsilon_j\boldsymbol{X}_{j+i,k}\Big|\cdot|\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*|_1,
\end{aligned}$$

By the same arguments as in the proof of Theorem 1 and combining with $l = o(n)$,

$$\mathbb{P}\Big(\max_k\frac{1}{n}\Big|\sum_{j=1}^{n-i}\boldsymbol{X}_{j+i,k}\epsilon_j\Big| \ge \frac{n-i}{n}\tilde{\lambda}_\delta\Big) \le C_1(\log p)^{-1},$$

or

$$\mathbb{P}\Big(\max_k \frac{1}{n}\Big|\sum_{j=1}^{n-i} \boldsymbol{X}_{j+i,k}\epsilon_j\Big| \geq \frac{n-i}{n}\tilde{\lambda}_\delta\Big) \leq C_1 p^{-2},$$

where $\tilde{\lambda}_\delta$ be the result of replacing n in λ_δ with $n-i$. Combining with Theorem 1 and $l = o(n)$, yields $D_2 = O_{\mathbb{P}}(s_0\lambda_{\boldsymbol{w}}\lambda_\delta + \lambda_\delta C_{\boldsymbol{\Sigma}} h + (C_{\boldsymbol{\Sigma}} h)^2)$. Moreover,

$$\begin{aligned} D_3 =& \frac{1}{n}\sum_{j=1}^{n-i} |\boldsymbol{x}_j^\top(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*)| \cdot |\boldsymbol{x}_{j+i}^\top(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*)| \\ &\leq \frac{1}{n}\sum_{j=1}^{n-i} \frac{|\boldsymbol{x}_j^\top(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*)|^2 + |\boldsymbol{x}_{j+i}^\top(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*)|^2}{2} \\ &\leq |X(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*)|_2^2/n. \end{aligned}$$

Combining with (S3.7), we obtain $D_3 = O_{\mathbb{P}}(s_0{\lambda_w}^2 + \lambda_\delta C_{\boldsymbol{\Sigma}} h + (C_{\boldsymbol{\Sigma}} h)^2)$. Then we get

$$T_1 = O_{\mathbb{P}}(l(n^{2/q'-1} + s_0\lambda_{\boldsymbol{w}}\lambda_\delta + \lambda_\delta C_{\boldsymbol{\Sigma}} h + (C_{\boldsymbol{\Sigma}} h)^2 + l/n)),$$

since $D_4 = O(ln^{-1})$ and $\lambda_w \leq \lambda_\delta$. Therefore

$$|\widehat{\boldsymbol{\Sigma}}_{n,l}^\epsilon - \boldsymbol{\Sigma}_n^\epsilon|_2 = O_{\mathbb{P}}(l(n^{2/q'-1} + s_0\lambda_{\boldsymbol{w}}\lambda_\delta + \lambda_\delta C_{\boldsymbol{\Sigma}} h + (C_{\boldsymbol{\Sigma}} h)^2 + l/n)) + 2\sum_{i=l+1}^{n} |\gamma_i^\epsilon|.$$

Define the projection operator $\mathcal{P}_k$ as $\mathcal{P}_k \cdot = E(\cdot \mid \mathcal{F}_k) - E(\cdot \mid \mathcal{F}_{k-1})$. Since $\mathcal{P}_k$, $k \in \mathbb{Z}$ are orthogonal projections, we have $\epsilon_i = \sum_{k\in\mathbb{Z}} \mathcal{P}_k \epsilon_i$. Hence, we have

$$\gamma_k^\epsilon = \mathbb{E}\Big(\sum_{i\in\mathbb{Z}} \mathcal{P}_i\epsilon_0 \sum_{j\in\mathbb{Z}} \mathcal{P}_j\epsilon_k\Big) = \mathbb{E}\Big(\sum_{i\in\mathbb{Z}} \mathcal{P}_i\epsilon_0\mathcal{P}_i\epsilon_k\Big).$$

Then by Schwarz's inequality and stationarity, we have

$$\sum_{k\in\mathbb{Z}} |\gamma_k^\epsilon| \leq \sum_{k\in\mathbb{Z}}\sum_{i\in\mathbb{Z}} \mathbb{E}|\mathcal{P}_i\epsilon_0\mathcal{P}_i\epsilon_k| \leq \sum_{k\in\mathbb{Z}}\sum_{i\in\mathbb{Z}} \|\mathcal{P}_i\epsilon_0\|_q \|\mathcal{P}_i\epsilon_k\|_q \leq \Delta_{0,q}^2 < \infty.$$

Let $l \to \infty, l(n^{2/q'-1} + s_0\lambda_{\boldsymbol{w}}\lambda_\delta + \lambda_\delta C_{\boldsymbol{\Sigma}} h + (C_{\boldsymbol{\Sigma}} h)^2 + l/n) = o(1)$, yields

$$|\widehat{\boldsymbol{\Sigma}}_{n,l}^\epsilon - \boldsymbol{\Sigma}_n^\epsilon|_2 = o_{\mathbb{P}}(1),$$

since $\sum_{i=l+1}^{n} |\gamma_i^\epsilon| \to 0$, as $l \to \infty$.

Lemma 11. *Under the conditions of Theorem 1, for any fixed $k > 0$, we have $\widehat{\gamma}_k^{\epsilon} \xrightarrow{\mathbb{P}} \gamma_k^{\epsilon}$.*

Proof of Lemma 11. A detailed proof has already been provided in the proof of Lemma 1; see Equation S5.1 for reference.

Proof of Theorem 3: Since $|\widehat{\boldsymbol{\Sigma}}_{n,l}^{\epsilon} - \boldsymbol{\Sigma}_n^{\boldsymbol{\epsilon}}|_2 = o_{\mathbb{P}}(1)$ by Lemma 1, combining with Lemma 10, we get

$$\begin{aligned} |\widehat{\sigma}_k^2 - \sigma_k^2| &= |n^{-1}\widehat{\boldsymbol{\Theta}}_k^{\top} X^{\top}(\widehat{\boldsymbol{\Sigma}}_{n,l} - \boldsymbol{\Sigma}_n^{\boldsymbol{\epsilon}})X\widehat{\boldsymbol{\Theta}}_k| \\ &\leq n^{-1}|X\widehat{\boldsymbol{\Theta}}_k|_2^2 \cdot |\widehat{\boldsymbol{\Sigma}}_{n,l}^{\epsilon} - \boldsymbol{\Sigma}_n^{\epsilon}|_2 \\ &= o_{\mathbb{P}}(1). \end{aligned}$$

Thus $\widehat{\sigma}_k^2$ is a consistent estimator of σ_k^2. Combining with Theorem 2, Lemma 1 and Slutsky's Theorem, the proofs are completed.

S6 Proof of Theorem 4

Lemma 12. *Under the conditions of Theorem 1, with probability at least $1 - C_1(\log p)^{-1}$ for case (i) in Theorem 1 or with probability at least $1 - C_1 p^{-2}$ for case (ii) in Theorem 1, the single task Lasso estimator $\widehat{\boldsymbol{\beta}}^l$ satisfies*

$$|\widehat{\boldsymbol{\beta}}^l - \boldsymbol{\beta}^*|_2^2 \leq c_1 s_0 \lambda_\delta^2, \quad |\widehat{\boldsymbol{\beta}}^l - \boldsymbol{\beta}^*|_1 \leq c_2 s_0 \lambda_\delta.$$

Proof of Lemma 12 Since $\widehat{\boldsymbol{\beta}}^l$ minimizes (4.15), we have

$$\frac{1}{n}|X(\widehat{\boldsymbol{\beta}}^l - \boldsymbol{\beta}^*)|_2^2 + 2\lambda|\widehat{\boldsymbol{\beta}}^l|_1 \leq 2\lambda|\boldsymbol{\beta}^*|_1 + \frac{2}{n}\boldsymbol{\epsilon}^{\top}X(\widehat{\boldsymbol{\beta}}^l - \boldsymbol{\beta}^*). \tag{S6.1}$$

Define the event $\boldsymbol{\Gamma} := \{\max_{1\leq j\leq p}|\Gamma_j| \leq \lambda/2\}$ for some constant $c > 0$, where $\Gamma_j = n^{-1}\sum_{i=1}^n \boldsymbol{X}_{ij}\boldsymbol{\epsilon}_i$, define the event $\boldsymbol{\Gamma}_1 := |\widehat{\boldsymbol{\Sigma}} - \boldsymbol{\Sigma}|_{\max} \leq \lambda_X^0\}$, for some constant $c_1 > 0$.

Then on the event $\boldsymbol{\Gamma} \cap \boldsymbol{\Gamma}_1$, using the fact

$$\frac{1}{n}|2\boldsymbol{\epsilon}^{\top}X(\widehat{\boldsymbol{\beta}}^l - \boldsymbol{\beta}^*)|_\infty \leq 2(\max_{1\leq j\leq p}|\Gamma_j|) \cdot |\widehat{\boldsymbol{\beta}}^l - \boldsymbol{\beta}^*|_1,$$

inequality (S6.1) implies that

$$\frac{1}{n}|X(\widehat{\boldsymbol{\beta}}^l - \boldsymbol{\beta}^*)|_2^2 + 2\lambda|\widehat{\boldsymbol{\beta}}^l|_1 \leq 2\lambda|\boldsymbol{\beta}^*|_1 + \lambda|\widehat{\boldsymbol{\beta}}^l - \boldsymbol{\beta}^*|_1. \tag{S6.2}$$

On the left hand side in (S6.2), using the triangle inequality,

$$|\widehat{\boldsymbol{\beta}}^l|_1 = |\widehat{\boldsymbol{\beta}}^l_S|_1 + |\widehat{\boldsymbol{\beta}}^l_{S^c}|_1 \geq |\boldsymbol{\beta}^*_S|_1 - |\widehat{\boldsymbol{\beta}}^l_S - \boldsymbol{\beta}^*_S|_1 + |\widehat{\boldsymbol{\beta}}^l_{S^c}|_1,$$

whereas on the right hand side in (S6.2), we can use.

$$|\widehat{\boldsymbol{\beta}}^l - \boldsymbol{\beta}^*|_1 = |\widehat{\boldsymbol{\beta}}^l_S - \boldsymbol{\beta}^*_S|_1 + |\widehat{\boldsymbol{\beta}}^l_{S^c}|_1.$$

Thus, we have

$$\frac{1}{n}|X(\widehat{\boldsymbol{\beta}}^l - \boldsymbol{\beta}^*)|_2^2 + \lambda|\widehat{\boldsymbol{\beta}}^l_{S^c}|_1 \leq 3\lambda|\widehat{\boldsymbol{\beta}}^l_S - \boldsymbol{\beta}^*_S|_1. \tag{S6.3}$$

In particular, the above implies

$$|\widehat{\boldsymbol{\beta}}^l_{S^c}|_1 \leq 3|\widehat{\boldsymbol{\beta}}^l_S - \boldsymbol{\beta}^*_S|_1. \tag{S6.4}$$

Based on Corollary 6.8 of Bühlmann and van de Geer (2011), we obtain

$$\begin{aligned}|\widehat{\boldsymbol{\beta}}^l_S - \boldsymbol{\beta}^*_S|_1^2 \leq & s_0|\widehat{\boldsymbol{\beta}}^l_S - \boldsymbol{\beta}^*_S|_2^2 \leq s_0|\widehat{\boldsymbol{\beta}}^l - \boldsymbol{\beta}^*|_2^2 \\ \leq & cs_0(\widehat{\boldsymbol{\beta}}^l - \boldsymbol{\beta}^*)^\top \boldsymbol{\Sigma}(\widehat{\boldsymbol{\beta}}^l - \boldsymbol{\beta}^*) \leq c_1\frac{s_0}{n}|X(\widehat{\boldsymbol{\beta}}^l - \boldsymbol{\beta}^*)|_2^2.\end{aligned}$$

Combining with (S6.3) again, we get

$$|\widehat{\boldsymbol{\beta}}^l_S - \boldsymbol{\beta}^*_S|_1 \leq c_2 s_0 \lambda.$$

Furthermore, $|\widehat{\boldsymbol{\beta}}^l - \boldsymbol{\beta}^*|_1 = |\widehat{\boldsymbol{\beta}}^l_S - \boldsymbol{\beta}^*_S|_1 + |\widehat{\boldsymbol{\beta}}^l_{S^c}|_1 \leq c_3|\widehat{\boldsymbol{\beta}}^l_S - \boldsymbol{\beta}^*_S|_1$, we have

$$|\widehat{\boldsymbol{\beta}}^l - \boldsymbol{\beta}^*|_1 \leq cs_0\lambda,$$

combining with (S6.3),

$$|\widehat{\boldsymbol{\beta}}^l - \boldsymbol{\beta}^*|_2^2 \leq C_1|X(\widehat{\boldsymbol{\beta}}^l - \boldsymbol{\beta}^*)|_2^2/n \leq Cs_0\lambda^2.$$

Now, we need to control the probability $\mathbb{P}(\mathbf{\Gamma} \cap \mathbf{\Gamma_1})$. We have already provided the corresponding proof in Section S3.

Proof of Theorem 4 Remark that

$$\widehat{L}(\widehat{\boldsymbol{w}}^{(k)}) - \widehat{L}(\widehat{\boldsymbol{\beta}}^l) = \frac{\widehat{L}_1(\widehat{\boldsymbol{w}}^{(k)}) - \widehat{L}_1(\widehat{\boldsymbol{\beta}}^l_{(2)})}{2} + \frac{\widehat{L}_2(\widehat{\boldsymbol{w}}^{(k)}) - \widehat{L}_2(\widehat{\boldsymbol{\beta}}^l_{(1)})}{2}.$$

Note that

$$\begin{aligned}
&\widehat{L}_i(\widehat{\boldsymbol{w}}^{(k)}) - \widehat{L}_i(\widehat{\boldsymbol{\beta}}^l_j) \\
=&\Big(\frac{1}{n_0/2}|\boldsymbol{y}^{(0)}_{(i)} - X^{(0)}_{(i)}\widehat{\boldsymbol{w}}^{(k)}|_2^2 - \frac{1}{n_0/2}|\boldsymbol{y}^{(0)}_{(i)}|_2^2\Big) - \Big(\frac{1}{n_0/2}|\boldsymbol{y}^{(0)}_{(i)} - X^{(0)}_{(i)}\widehat{\boldsymbol{\beta}}^l_j|_2^2 - \frac{1}{n_0/2}|\boldsymbol{y}^{(0)}_{(i)}|_2^2\Big),
\end{aligned}$$

where $j = 2$ when $i = 1$, and $j = 1$ when $i = 2$, respectively. Hence, we first treat $\widehat{L}_i(\boldsymbol{w})$ as

$$\frac{1}{n_0/2}|\boldsymbol{y}^{(0)}_{(i)} - X^{(0)}_{(i)}\boldsymbol{w}|_2^2 - \frac{1}{n_0/2}|\boldsymbol{y}^{(0)}_{(i)}|_2^2,$$

when controlling the magnitude of $\widehat{L}(\widehat{\boldsymbol{w}}^{(k)}) - \widehat{L}(\widehat{\boldsymbol{\beta}}^l)$. We define that

$$L(\boldsymbol{w}) = -2{\boldsymbol{\beta}^*}^\top \mathbf{\Sigma}\boldsymbol{w} + \boldsymbol{w}^\top \mathbf{\Sigma}\boldsymbol{w},$$

which is the population version of $\widehat{L}(\boldsymbol{w})$. We first bound the term $\frac{\widehat{L}_1(\widehat{\boldsymbol{w}}^{(k)}) - \widehat{L}_1(\widehat{\boldsymbol{\beta}}^l_{(2)})}{2}$, and the second term can be handled similarly. For simplicity, in the following text, we denote $\widehat{L}_1(\widehat{\boldsymbol{w}}^{(k)}) - \widehat{L}_1(\widehat{\boldsymbol{\beta}}^l_{(2)})$ as $\widehat{L}(\widehat{\boldsymbol{w}}^{(k)}) - \widehat{L}(\widehat{\boldsymbol{\beta}}^l)$, and abbreviate $X^{(0)}_{(1)}, \boldsymbol{y}^{(0)}_{(1)}, \widehat{\boldsymbol{\beta}}^l_{(2)}, n_0/2$ as $X, \boldsymbol{y}, \widehat{\boldsymbol{\beta}}^l, n$, respectively. Then

$$\begin{aligned}
&\widehat{L}(\widehat{\boldsymbol{w}}^{(k)}) - \widehat{L}(\widehat{\boldsymbol{\beta}}^l) \\
\leq&(\widehat{L}(\widehat{\boldsymbol{w}}^{(k)}) - \widehat{L}(\boldsymbol{w}^{(k)})) + (\widehat{L}(\boldsymbol{w}^{(k)}) - L(\boldsymbol{w}^{(k)})) + (L(\boldsymbol{w}^{(k)}) - L(\boldsymbol{\beta}^*)) \\
&+(L(\boldsymbol{\beta}^*) - \widehat{L}(\boldsymbol{\beta}^*)) + (\widehat{L}(\boldsymbol{\beta}^*) - \widehat{L}(\widehat{\boldsymbol{\beta}}^l)) \\
:=&R1 + R2 + R3 + R4 + R5.
\end{aligned}$$

We know that

$$\begin{aligned}|\widehat{L}(\widehat{\boldsymbol{w}}^{(k)}) - \widehat{L}(\boldsymbol{w}^{(k)})| &= |(|\boldsymbol{y} - X\widehat{\boldsymbol{w}}^{(k)}|_2^2 - |\boldsymbol{y} - X\boldsymbol{w}^{(k)}|_2^2)/n| \\ &= |(\widehat{\boldsymbol{w}}^{(k)} - \boldsymbol{w}^{(k)})^\top X^\top X(\widehat{\boldsymbol{w}}^{(k)} - \boldsymbol{w}^{(k)})/n + 2(\boldsymbol{w}^{(k)} - \widehat{\boldsymbol{w}}^{(k)})^\top X^\top(\boldsymbol{y} - X\boldsymbol{w}^{(k)})/n| \\ &= |(\widehat{\boldsymbol{w}}^{(k)} - \boldsymbol{w}^{(k)})^\top X^\top X(\widehat{\boldsymbol{w}}^{(k)} - \boldsymbol{w}^{(k)})/n + 2(\boldsymbol{w}^{(k)} - \widehat{\boldsymbol{w}}^{(k)})^\top X^\top(X\boldsymbol{\delta}^{(k)} + \boldsymbol{\epsilon})/n|.\end{aligned}$$

Combining with Lemma 12, we have $|\widehat{\boldsymbol{w}}^{(k)} - \boldsymbol{w}^{(k)}|_2^2 \leq s_k\lambda_k^2$ with probability at least $1-\delta_1$, and $\widehat{\boldsymbol{w}}^{(k)} - \boldsymbol{w}^{(k)}$ is independent of X, hence we can obtain that

$$(\widehat{\boldsymbol{w}}^{(k)} - \boldsymbol{w}^{(k)})^\top X^\top X(\widehat{\boldsymbol{w}}^{(k)} - \boldsymbol{w}^{(k)})/n \leq c(\widehat{\boldsymbol{w}}^{(k)} - \boldsymbol{w}^{(k)})^\top \boldsymbol{\Sigma}(\widehat{\boldsymbol{w}}^{(k)} - \boldsymbol{w}^{(k)}) \leq c_1 s_k\lambda_k^2,$$

with probability tending to 1 as $n \to \infty$. And,

$$\begin{aligned}&|(\boldsymbol{w}^{(k)} - \widehat{\boldsymbol{w}}^{(k)})^\top X^\top(X\boldsymbol{\delta}^{(k)} + \boldsymbol{\epsilon})/n| \\ \leq&\sqrt{[(\boldsymbol{w}^{(k)} - \widehat{\boldsymbol{w}}^{(k)})^\top X^\top X(\boldsymbol{w}^{(k)} - \widehat{\boldsymbol{w}}^{(k)})]/n}\sqrt{({\boldsymbol{\delta}^{(k)}}^\top X^\top X\boldsymbol{\delta}^{(k)})/n} + |(\boldsymbol{w}^{(k)} - \widehat{\boldsymbol{w}}^{(k)})^\top X^\top\boldsymbol{\epsilon}/n| \\ \leq&c\sqrt{s_k}\lambda_k|\boldsymbol{\delta}^{(k)}|_2 + c\sqrt{s_k}\lambda_k^2 + |\boldsymbol{w}^{(k)} - \widehat{\boldsymbol{w}}^{(k)}|_1|X^\top\boldsymbol{\epsilon}|_\infty/n \\ \leq&c\sqrt{s_k}\lambda_k|\boldsymbol{\delta}^{(k)}|_2 + +c\sqrt{s_k}\lambda_k^2 + c_1\lambda_\delta s_k\lambda_k,\end{aligned}$$

with probability at least $1-\delta_2$, $\delta_2 \to 0$ as $n_{\min} \to \infty$, where the last second inequality can be similarly proved as in Section S3 and the last inequality is obviously.

So we have that $|R_1| \leq c\sqrt{s_k}\lambda_k|\boldsymbol{\delta}^{(k)}|_2 + c_1\lambda_\delta s_k\lambda_k + cs_k\lambda_k^2$, with probability at least $1-\gamma_1$, $\gamma_1 \to 0$ as $n_{\min} \to \infty$.

Secondly,

$$\begin{aligned}|R_2| =&| - 2\boldsymbol{y}^\top X\boldsymbol{w}^{(k)}/n + (\boldsymbol{w}^{(k)})^\top X^\top X\boldsymbol{w}^{(k)}/n - (\boldsymbol{w}^{(k)})^\top\boldsymbol{\Sigma}\boldsymbol{w}^{(k)} + 2{\boldsymbol{\beta}^*}^\top\boldsymbol{\Sigma}\boldsymbol{w}^{(k)}| \\ \leq&|(\boldsymbol{w}^{(k)})^\top X^\top X\boldsymbol{w}^{(k)}/n - (\boldsymbol{w}^{(k)})^\top\boldsymbol{\Sigma}\boldsymbol{w}^{(k)}| + 2|\boldsymbol{y}^\top X\boldsymbol{w}^{(k)}/n - {\boldsymbol{\beta}^*}^\top\boldsymbol{\Sigma}\boldsymbol{w}^{(k)}| \\ \leq&c\lambda_\delta + 2|{\boldsymbol{\beta}^*}^\top X^\top X\boldsymbol{w}^{(k)}/n - {\boldsymbol{\beta}^*}^\top\boldsymbol{\Sigma}\boldsymbol{w}^{(k)}| + 2|\boldsymbol{\epsilon}^\top X\boldsymbol{w}^{(k)}|/n \\ \leq&c_1\lambda_\delta + c|\boldsymbol{w}^{(k)}|_2\lambda_\delta,\end{aligned}$$

with probability at least at least $1-\gamma_2$, $\gamma_2 \to 0$ as $n_{\min} \to \infty$.

Thirdly, for R_3, we can get that

$$\begin{aligned} R_3 =&(\boldsymbol{w}^{(k)})^\top \boldsymbol{\Sigma} \boldsymbol{w}^{(k)} - 2\boldsymbol{\beta}^{*\top} \boldsymbol{\Sigma} \boldsymbol{w}^{(k)} + \boldsymbol{\beta}^{*\top} \boldsymbol{\Sigma} \boldsymbol{\beta}^* \\ =&(\boldsymbol{w}^{(k)} - \boldsymbol{\beta}^*)^\top \boldsymbol{\Sigma} (\boldsymbol{w}^{(k)} - \boldsymbol{\beta}^*), \end{aligned}$$

then for $k \in \mathcal{A}_h$, we have $R_3 \leq \lambda_{\max}(\boldsymbol{\Sigma}) |\boldsymbol{w}^{(k)} - \boldsymbol{\beta}^*|_2^2$, and for $k \notin \mathcal{A}_h$, we have $R_3 \geq \lambda_{\min}(\boldsymbol{\Sigma}) |\boldsymbol{w}^{(k)} - \boldsymbol{\beta}^*|_2^2$

Fourthly, we have

$$\begin{aligned} |R_4| =&| - 2\boldsymbol{y}^\top X \boldsymbol{\beta}^*/n + \boldsymbol{\beta}^{*\top} X^\top X \boldsymbol{\beta}^*/n + \boldsymbol{\beta}^{*\top} \boldsymbol{\Sigma} \boldsymbol{\beta}^*| \\ \leq&|\boldsymbol{\beta}^{*\top} X^\top X \boldsymbol{\beta}^*/n - \boldsymbol{\beta}^{*\top} \boldsymbol{\Sigma} \boldsymbol{\beta}^*| + 2|\boldsymbol{\epsilon}^\top X \boldsymbol{\beta}^*|/n \\ \leq&c\lambda_\delta + c|\boldsymbol{\beta}^*|_2 \lambda_\delta, \end{aligned}$$

with probability at least $1 - \gamma_4$, $\gamma_4 \to 0$ as $n_{\min} \to \infty$.

Finally, for R_5, we have that

$$\begin{aligned} |R_5| &= ||\boldsymbol{y} - X\widehat{\boldsymbol{\beta}}^l|_2^2 - |\boldsymbol{y} - X\boldsymbol{\beta}^*|_2^2|/n \\ &= |(\widehat{\boldsymbol{\beta}}^l - \boldsymbol{\beta}^*)^\top X^\top X (\widehat{\boldsymbol{\beta}}^l - \boldsymbol{\beta}^*)/n + 2(\boldsymbol{\beta}^* - \widehat{\boldsymbol{\beta}}^l)^\top X^\top (\boldsymbol{y} - X\boldsymbol{\beta}^*)/n| \\ &= |(\widehat{\boldsymbol{\beta}}^l - \boldsymbol{\beta}^*)^\top X^\top X (\widehat{\boldsymbol{\beta}}^l - \boldsymbol{\beta}^*)/n + 2(\boldsymbol{\beta}^* - \widehat{\boldsymbol{\beta}}^l)^\top X^\top \boldsymbol{\epsilon}/n|, \end{aligned}$$

then on the event $\{|\widehat{\boldsymbol{\beta}}^l - \boldsymbol{\beta}^*|_2^2 \leq s_0\lambda_\delta^2, |\widehat{\boldsymbol{\beta}}^l_{S^c}|_1 \leq 3|\widehat{\boldsymbol{\beta}}^l_S - \boldsymbol{\beta}^*_S|_1\} \cap \{|X^\top X/n - \boldsymbol{\Sigma}|_{\max} \leq c\lambda_X^0\}$, combining with corollary 6.8 in Bühlmann and van de Geer (2011), we obtain that

$$(\boldsymbol{\beta}^* - \widehat{\boldsymbol{\beta}}^l)^\top X^\top X (\boldsymbol{\beta}^* - \widehat{\boldsymbol{\beta}}^l)/n \leq c(\boldsymbol{\beta}^* - \widehat{\boldsymbol{\beta}}^l)^\top \boldsymbol{\Sigma} (\boldsymbol{\beta}^* - \widehat{\boldsymbol{\beta}}^l) \leq c_1 |\boldsymbol{\beta}^* - \widehat{\boldsymbol{\beta}}^l|_2 \leq c_2 s_0 \lambda_\delta^2.$$

And

$$|(\boldsymbol{\beta}^* - \widehat{\boldsymbol{\beta}}^l)^\top X^\top \boldsymbol{\epsilon}/n| \leq |\boldsymbol{\beta}^* - \widehat{\boldsymbol{\beta}}^l|_1 |X^\top \boldsymbol{\epsilon}/n|_\infty \leq s_0 \lambda_\delta^2,$$

with probability tends to one. Hence we have $|R_5| \leq c_1 s_0 \lambda_\delta^2$ with probability at least $1 - \gamma_5$, $\gamma_5 \to 0$ as $n_{\min} \to \infty$.

To summarize, we have that for $k \in \mathcal{A}_h$,

$$\widehat{L}(\widehat{\boldsymbol{w}}^{(k)}) - \widehat{L}(\widehat{\boldsymbol{\beta}}^{l}) \leq c\sqrt{s_k}\lambda_k|\boldsymbol{\delta}^{(k)}|_2 + c\lambda_\delta s_k\lambda_k + cs_k\lambda_k^2 + c\lambda_\delta + c\lambda_{\max}(\boldsymbol{\Sigma})|\boldsymbol{w}^{(k)} - \boldsymbol{\beta}^*|_2^2 + cs_0\lambda_\delta^2, \tag{S6.5}$$

and for $k \notin \mathcal{A}_h$

$$\widehat{L}(\widehat{\boldsymbol{w}}^{(k)}) - \widehat{L}(\widehat{\boldsymbol{\beta}}^{l}) \geq c\lambda_{\min}(\boldsymbol{\Sigma})|\boldsymbol{w}^{(k)} - \boldsymbol{\beta}^*|_2^2 - c\sqrt{s_k}\lambda_k|\boldsymbol{\delta}^{(k)}|_2 - c\lambda_\delta s_k\lambda_k - cs_k\lambda_k^2 - cs_0\lambda_\delta^2 - c\lambda_\delta,$$

with probability at least $1-\gamma_6$, $\gamma_6 \to 0$ as $n_{\min} \to \infty$.

Next,

$$\begin{aligned}\Delta =& \big||\boldsymbol{y}_{(1)}^{(0)} - \boldsymbol{X}_{(1)}\widehat{\boldsymbol{\beta}}_{(2)}^{l}|_2^2 - |\boldsymbol{y}_{(2)}^{(0)} - \boldsymbol{X}_{(2)}\widehat{\boldsymbol{\beta}}_{(1)}^{l}|_2^2\big|/n \\ =& \big||\boldsymbol{X}_{(1)}(\boldsymbol{\beta}^* - \widehat{\boldsymbol{\beta}}_{(2)}^{l})|_2^2 - 2\boldsymbol{\epsilon}_{(1)}^{(0)\top}\boldsymbol{X}_{(1)}(\boldsymbol{\beta}^* - \widehat{\boldsymbol{\beta}}_{(2)}^{l}) + |\boldsymbol{\epsilon}_{(1)}^{(0)}|_2^2 \\ &-|\boldsymbol{X}_{(2)}(\boldsymbol{\beta}^* - \widehat{\boldsymbol{\beta}}_{(1)}^{l})|_2^2 + 2\boldsymbol{\epsilon}_{(2)}^{(0)\top}\boldsymbol{X}_{(2)}(\boldsymbol{\beta}^* - \widehat{\boldsymbol{\beta}}_{(1)}^{l}) - |\boldsymbol{\epsilon}_{(2)}^{(0)}|_2^2\big|/n.\end{aligned}$$

As shown in R_5, we have

$$|\boldsymbol{X}_{(1)}(\boldsymbol{\beta}^* - \widehat{\boldsymbol{\beta}}_{(2)}^{l})|_2^2/n \leq s_0\lambda_\delta^2, \quad |\boldsymbol{\epsilon}_{(1)}^{(0)\top}\boldsymbol{X}_{(1)}(\boldsymbol{\beta}^* - \widehat{\boldsymbol{\beta}}_{(2)}^{l})|/n \leq s_0\lambda_\delta^2,$$

with probability tending to 1. And

$$\big||\boldsymbol{\epsilon}_{(1)}^{(0)}|_2^2 - |\boldsymbol{\epsilon}_{(2)}^{(0)}|_2^2\big|/n \leq \big||\boldsymbol{\epsilon}_{(1)}^{(0)}|_2^2/n - \gamma_0^\epsilon\big| + \big||\boldsymbol{\epsilon}_{(2)}^{(0)}|_2^2/n - \gamma_0^\epsilon\big| \leq cn^{2/q_1-1},$$

with probability tending to 1, where q_1 satisfies $2 < q_1 < q'$, and the last inequality have been shown in Section S5. Then, $\Delta \leq cs_0\lambda_\delta^2 + +n^{2/q_1-1}$, with probability at least $1-\gamma_7$, $\gamma_7 \to 0$ as $n_{\min} \to \infty$. Hence, combining with the condition that M is a constant, we obtain that,

$$\begin{aligned}&\mathbb{P}(\max_{k\in\mathcal{A}_h}\widehat{L}(\widehat{\boldsymbol{w}}^{(k)}) - \widehat{L}(\widehat{\boldsymbol{\beta}}^{l}) \geq C(\Delta \vee C_1)) \\ \leq& M\mathbb{P}(\widehat{L}(\widehat{\boldsymbol{w}}^{(k)}) - \widehat{L}(\widehat{\boldsymbol{\beta}}^{l}) \geq C(\Delta \vee C_1)) \\ \leq& M\gamma_6 \to 0, \quad \text{as} \quad n_{\min} \to \infty,\end{aligned}$$

for some h, C and C_1, since the right side of the Equation (S6.5) is sufficient small when $n_{\min}$ tends to infinity. Then, we obtain that $\max_{k\in\mathcal{A}_h}\widehat{L}(\widehat{\boldsymbol{w}}^{(k)})-\widehat{L}(\widehat{\boldsymbol{\beta}}^{l})\leq C(\Delta\vee C_1)$ with probability tending to one. We can similarly prove that $\min_{k\notin\mathcal{A}_h}\widehat{L}(\widehat{\boldsymbol{w}}^{(k)})-\widehat{L}(\widehat{\boldsymbol{\beta}}^{l})\geq C(\Delta\vee C_1)$, with probability tending to one.

Proof of Remark 2: Combining the result of Theorem 1 with the condition $C_{\boldsymbol{\Sigma}}h\lesssim\sqrt{s_0}\lambda_\delta$ and $n_{\mathcal{A}_h}\gg n_0$, we obtain that

$$\frac{1}{n_0}\big|X^{(0)}(\widehat{\boldsymbol{\beta}}-\boldsymbol{\beta}^*)\big|_2^2\ \lesssim\ \big(s_0\lambda_w^2+\lambda_\delta C_{\boldsymbol{\Sigma}}h+(C_{\boldsymbol{\Sigma}}h)^2\big)\wedge s_0\lambda_\delta^2,$$

and

$$|\widehat{\boldsymbol{\beta}}-\boldsymbol{\beta}^*|_1\ \lesssim\ \big(s_0\lambda_w+\tfrac{(C_{\boldsymbol{\Sigma}}h)^2}{\lambda_\delta}+C_{\boldsymbol{\Sigma}}h\big)\wedge s_0\lambda_\delta,$$

with probability tending to 1. By Theorem 4, we also have that $\mathbb{P}(\widehat{\boldsymbol{\beta}}^{\mathrm{SD}}=\widehat{\boldsymbol{\beta}})\to 1$, as $n_{\min}\to\infty$. Then, it is obviously to get that

$$\frac{1}{n_0}\big|X^{(0)}(\widehat{\boldsymbol{\beta}}^{\mathrm{SD}}-\boldsymbol{\beta}^*)\big|_2^2\ \lesssim\ \big(s_0\lambda_w^2+\lambda_\delta C_{\boldsymbol{\Sigma}}h+(C_{\boldsymbol{\Sigma}}h)^2\big)\wedge s_0\lambda_\delta^2,$$

and

$$|\widehat{\boldsymbol{\beta}}^{\mathrm{SD}}-\boldsymbol{\beta}^*|_1\ \lesssim\ \big(s_0\lambda_w+\tfrac{(C_{\boldsymbol{\Sigma}}h)^2}{\lambda_\delta}+C_{\boldsymbol{\Sigma}}h\big)\wedge s_0\lambda_\delta,$$

with probability tending to 1.

S7 An Alternative Estimation Approach

We focus on the following optimization problem, which is the linear-model formulation of the estimator proposed in Li et al. (2024). Without loss of generality, we first assume that $\mathcal{A}_h=[M]$.

$$(\widehat{\boldsymbol{\beta}}^{o},\widehat{\boldsymbol{\delta}}^{(1)},\ldots,\widehat{\boldsymbol{\delta}}^{(M)})=\arg\min_{\boldsymbol{\beta},|\boldsymbol{\delta}^{(k)}|_2\leq C}\Big\{\lambda_\beta|\boldsymbol{\beta}|_1+\sum_{k=1}^{M}\lambda_k|\boldsymbol{\delta}^{(k)}|_1\Big\},$$

$$\text{subject to} \quad \begin{cases} |\sum_{i=1}^{n_k} \boldsymbol{x}_i^{(k)}\{y_i^{(k)} - (\boldsymbol{x}_i^{(k)})^\top(\boldsymbol{\beta} - \boldsymbol{\delta}^{(k)})\}|_\infty \le \lambda_k, \quad \forall\, 0 \le k \le M, \\ |\sum_{i=1}^{n_0} \boldsymbol{x}_i^{(0)}(y_i^{(0)} - (\boldsymbol{x}_i^{(0)})^\top\boldsymbol{\beta}) + \sum_{k=1}^{M}\sum_{i=1}^{n_k} \boldsymbol{x}_i^{(k)}\{y_i^{(k)} - (\boldsymbol{x}_i^{(k)})^\top(\boldsymbol{\beta} - \boldsymbol{\delta}^{(k)})\}|_\infty \le \lambda_\beta. \end{cases} \tag{S7.6}$$

The tuning parameters λ_β and λ_k, $0 \le k \le M$, are defined as follows. For case (i) in Theorem 1, let

$$\lambda_X^{\mathcal{A}} \asymp N_X^2 \max\{n_{\mathcal{A}_h}^{2\chi/\iota-1}(p\log p)^{4/\iota}, (\log p/n_{\mathcal{A}_h})^{1/2}\},$$

$$\lambda_{x\epsilon}^{\mathcal{A}} \asymp N_X N_e \max\{n_{\mathcal{A}_h}^{\pi/\tau-1}(p\log p)^{1/\tau}, (\log p/n_{\mathcal{A}_h})^{1/2}\},$$

and define $\tilde{\lambda}_\beta = \max\{\lambda_X^{\mathcal{A}}, \lambda_{x\epsilon}^{\mathcal{A}}\}$. For case (ii) in Theorem 1, define

$$\lambda_X^{\mathcal{A}} \asymp L_\varrho^2(\log p)^{(1+4\varrho)/2}/\sqrt{n_{\mathcal{A}_h}}, \quad \lambda_{x\epsilon}^{\mathcal{A}} \asymp K_\nu L_\varrho(\log p)^{(1+2\nu+2\varrho)/2}/\sqrt{n_{\mathcal{A}_h}},$$

$\tilde{\lambda}_\beta = \max\{\lambda_X^{\mathcal{A}}, \lambda_{x\epsilon}^{\mathcal{A}}\}$. The tuning parameter $\tilde{\lambda}_k$ is defined analogously to $\tilde{\lambda}_\beta$ with $n_{\mathcal{A}_h}$ replaced by $n_0 \wedge n_k$. We further define $\lambda_\beta = n_{\mathcal{A}_h}\tilde{\lambda}_\beta$ and $\lambda_k = n_k\tilde{\lambda}_k$ for $0 \le k \le M$. We refer to the transfer learning method that computes $\widehat{\boldsymbol{\beta}}^o$ via (S7.6) as O-Trans-Lasso.

We next present the theoretical guarantees for $\widehat{\boldsymbol{\beta}}^o$. We further assume that $n_k \asymp n_j$ for all $1 \le k, j \le M$, in order to simplify the analysis.

Proposition 1. *Assume that Conditions 1 and 2 hold, and that $n_0 \ll n_{\mathcal{A}_h}$. Further suppose that $s_0\lambda_X^0 \le c$ for some sufficiently small constant $c > 0$ and $\tilde{\lambda}_0\lambda_X^0 h/\tilde{\lambda}_\beta^2 + s\tilde{\lambda}_\beta + h\tilde{\lambda}_0/\tilde{\lambda}_\beta = o(1)$. Then the following asymptotic properties hold.*

(i) For case (i) in Theorem 1, with probability at least $1 - C_1(\log p)^{-1}$, we have

$$|X(\widehat{\boldsymbol{\beta}}^o - \boldsymbol{\beta}^*)|_2^2/n_0 \lesssim h\tilde{\lambda}_0 + s_0\tilde{\lambda}_\beta^2, \tag{S7.7}$$

$$|\widehat{\boldsymbol{\beta}}^o - \boldsymbol{\beta}^*|_1 \lesssim s_0\tilde{\lambda}_\beta + h\tilde{\lambda}_0/\tilde{\lambda}_\beta. \tag{S7.8}$$

(ii) For case (ii) in Theorem 1, (S7.7) *and* (S7.8) *hold with probability at least* $1 - c_1 p^{-2}$.

We compare the convergence rates in Remark 1 and Proposition 1. First, note that $\lambda_w \lesssim \tilde{\lambda}_\beta$ and $\lambda_\delta \lesssim \tilde{\lambda}_0$, since $\lambda_{xx}^{\mathcal{A}} \lesssim \lambda_X^{\mathcal{A}}$ and $\lambda_{xx}^0 \lesssim \lambda_X^0$, respectively. This is due to the fact that, under the dependence structure in case (i), the tail probability bounds established in Wu and Wu (2016) are not exponentially decaying, in contrast to those in the independent case. For the ℓ_1 error rate, the term $h\tilde{\lambda}_0/\tilde{\lambda}_\beta$ on the right-hand side of (S7.8) is strictly larger than h. This additional factor, $\tilde{\lambda}_0/\tilde{\lambda}_\beta$, arises due to differences in the methods employed. For example, even in the independent setting, the ℓ_1 error rates in Li et al. (2024) and Tian and Feng (2023) take different forms, with the latter considering a two-step transfer learning estimator similar to Trans-Lasso. Hence, when $C_{\mathbf{\Sigma}} < \infty$, the convergence rates in (S3.8) are at least as good as those in (S7.7) and (S7.8), and can be strictly better in certain regimes. For instance, when $s_0\lambda_w \ll h$, the ℓ_1 error rate in (S3.8) is strictly smaller than that in (S7.8). When $C_{\mathbf{\Sigma}} < \infty$ does not hold, the comparison depends on specific regimes, and we omit further discussion.

Moreover, it is clear that, compared with the Trans-Lasso method, O-Trans-Lasso incurs a substantially higher computational cost. Specifically, Trans-Lasso pools all auxiliary data to obtain a rough Lasso estimator and then uses the target data for bias correction, involving only two Lasso fits. In contrast, the estimation procedure in O-Trans-Lasso requires jointly estimating the regression coefficients for all auxiliary samples as well as the target sample, leading to a significantly increased computational burden.

Proof of Proposition 1: In addition, we introduce $\tilde{\lambda}_{\max}$ and $\tilde{\lambda}_{\min}$, which are defined analogously to $\tilde{\lambda}_\beta$, with $n_{\mathcal{A}_h}$ replaced by $n_0 \wedge n_{\min}$ and $n_0 \wedge n_{\max}$, respectively, where $n_{\min} = \min_{1\le k\le M} n_k$ and $n_{\max} = \max_{1\le k\le M} n_k$. We also introduce $\lambda_{x\epsilon}^k$ and λ_X^k, defined

analogously to $\lambda_{x\epsilon}^{\mathcal{A}}$ and $\lambda_X^{\mathcal{A}}$, respectively, with $n_{\mathcal{A}_h}$ replaced by n_k for $k \geq 0$. It follows that $\tilde{\lambda}_{\min} \leq \tilde{\lambda}_k \leq \tilde{\lambda}_{\max}$ for all $1 \leq k \leq M$. For notational convenience, we write $\widehat{\boldsymbol{\beta}}^o$ as $\widehat{\boldsymbol{\beta}}$ throughout the remainder of the proof.

Let $H_k = \{j : |\delta_j^{(k)}| \geq \lambda_k/n_k\}$. Notice that $|H_k| \lesssim h(\lambda_k/n_k)^{-1}$ and $|\boldsymbol{\delta}_{H_k^c}^{(k)}|_1 \leq h$. Let

$$Q_\delta = \sum_{k=1}^{M} n_k |\widehat{\boldsymbol{\delta}}^{(k)} - \boldsymbol{\delta}^{(k)}|_2^2 \quad \text{and} \quad \mathcal{R}_\delta = \sum_{k=1}^{M} \lambda_k |(\widehat{\boldsymbol{\delta}}^{(k)} - \boldsymbol{\delta}^{(k)})_{H_k}|_1.$$

Define $\widehat{\boldsymbol{u}} = \widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*$, $\widehat{\boldsymbol{v}}^{(k)} = \widehat{\boldsymbol{\delta}}^{(k)} - \boldsymbol{\delta}^{(k)}$, and $\widehat{\boldsymbol{w}}^{(k)} = \widehat{\boldsymbol{\beta}} - \widehat{\boldsymbol{\delta}}^{(k)}$. Notice that

$$\begin{aligned}
&\Big|\sum_{k=1}^{M} (\widehat{\boldsymbol{u}} - \widehat{\boldsymbol{v}}^{(k)})^\top \sum_{i=1}^{n_k} \boldsymbol{x}_i^{(k)} \{y_i^{(k)} - (\boldsymbol{x}_i^{(k)})^\top (\widehat{\boldsymbol{\beta}} - \widehat{\boldsymbol{\delta}}^{(k)})\}\Big| \\
\leq &\Big|\sum_{k=1}^{M} \widehat{\boldsymbol{u}}^\top \sum_{i=1}^{n_k} \boldsymbol{x}_i^{(k)} \{y_i^{(k)} - (\boldsymbol{x}_i^{(k)})^\top (\widehat{\boldsymbol{\beta}} - \widehat{\boldsymbol{\delta}}^{(k)})\}\Big| + \Big|\sum_{k=1}^{M} (\widehat{\boldsymbol{v}}^{(k)})^\top \sum_{i=1}^{n_k} \boldsymbol{x}_i^{(k)} \{y_i^{(k)} - (\boldsymbol{x}_i^{(k)})^\top (\widehat{\boldsymbol{\beta}} - \widehat{\boldsymbol{\delta}}^{(k)})\}\Big| \\
\leq &|\widehat{\boldsymbol{u}}|_1 \Big|\sum_{k=1}^{M} \sum_{i=1}^{n_k} \boldsymbol{x}_i^{(k)} \{y_i^{(k)} - (\boldsymbol{x}_i^{(k)})^\top (\widehat{\boldsymbol{\beta}} - \widehat{\boldsymbol{\delta}}^{(k)})\}\Big|_\infty + \sum_{k=1}^{M} |(\widehat{\boldsymbol{v}}^{(k)})^\top|_1 \Big|\sum_{i=1}^{n_k} \boldsymbol{x}_i^{(k)} \{y_i^{(k)} - (\boldsymbol{x}_i^{(k)})^\top (\widehat{\boldsymbol{\beta}} - \widehat{\boldsymbol{\delta}}^{(k)})\}\Big|_\infty \\
\lesssim &\lambda_\beta |\widehat{\boldsymbol{u}}|_1 + \sum_{k=1}^{M} \lambda_k |\widehat{\boldsymbol{v}}^{(k)}|_1,
\end{aligned}$$

where the last inequality is due to the fact that $\lambda_0 \ll \lambda_\beta$.

In what follows, we establish the proof for case (i) only, as case (ii) follows by analogous arguments. We first show that $\boldsymbol{\beta}^*$ and $\{\boldsymbol{\delta}^{(k)}\}_{1 \leq k \leq M}$ are feasible for the optimization problem (S7.6) with probability at least $1 - (\log p)^{-1}$. For the first constraint, we have

$$\begin{aligned}
\mathbb{P}\Big(\Big|\sum_{i=1}^{n_k} \boldsymbol{x}_i^{(k)} \{y_i^{(k)} - (\boldsymbol{x}_i^{(k)})^\top (\boldsymbol{\beta}^* - \boldsymbol{\delta}^{(k)})\}\Big|_\infty \geq \lambda_k\Big) =& \mathbb{P}\Big(|\sum_{i=1}^{n_k} \boldsymbol{x}_i^{(k)} \epsilon_i^{(k)}|_\infty \geq n_k \tilde{\lambda}_k\Big) \\
\leq& \mathbb{P}\Big(|\sum_{i=1}^{n_k} \boldsymbol{x}_i^{(k)} \epsilon_i^{(k)}|_\infty \geq n_k \lambda_{x\epsilon}^k\Big) \lesssim (\log p)^{-1}.
\end{aligned}$$

For the second constraint, by the similarly arguments as in the proof of Lemma 5,

$$\begin{aligned}
&\mathbb{P}\Big(\Big|\sum_{i=1}^{n_0} \boldsymbol{x}_i^{(0)} (y_i^{(0)} - (\boldsymbol{x}_i^{(0)})^\top \boldsymbol{\beta}^*) + \sum_{k=1}^{M} \sum_{i=1}^{n_k} \boldsymbol{x}_i^{(k)} \{y_i^{(k)} - (\boldsymbol{x}_i^{(k)})^\top (\boldsymbol{\beta}^* - \boldsymbol{\delta}^{(k)})\}\Big|_\infty \leq \lambda_\beta\Big) \\
=&\mathbb{P}\Big(\Big|\sum_{k=0}^{M} X^{(k)\top} \boldsymbol{\epsilon}^{(k)}\Big|_\infty \leq \lambda_\beta\Big) \geq 1 - (\log p)^{-1}.
\end{aligned}$$

Using the feasibility of $\boldsymbol{\beta}^*$ and $\{\boldsymbol{\delta}^{(k)}\}_{1\le k\le M}$, the following oracle inequality holds:

$$\Big|\sum_{k=1}^{M}(\widehat{\boldsymbol{w}}^{(k)}-\boldsymbol{w}^{(k)})^\top\sum_{i=1}^{n_k}\boldsymbol{x}_i^{(k)}(\boldsymbol{x}_i^{(k)})^\top(\widehat{\boldsymbol{w}}^{(k)}-\boldsymbol{w}^{(k)})\Big|\lesssim\lambda_\beta|\widehat{\boldsymbol{u}}|_1+\sum_{k=1}^{M}\lambda_k|\widehat{\boldsymbol{v}}^{(k)}|_1. \tag{S7.9}$$

For the left-hand side of (S7.9),

$$\begin{aligned}\text{LHS of (S7.9)}=&\sum_{k=1}^{M}(\widehat{\boldsymbol{w}}^{(k)}-\boldsymbol{w}^{(k)})^\top n_k\widehat{\boldsymbol{\Sigma}}^{(k)}(\widehat{\boldsymbol{w}}^{(k)}-\boldsymbol{w}^{(k)})\\ \ge&c_1\sum_{k=1}^{M}\{n_k|\widehat{\boldsymbol{w}}^{(k)}-\boldsymbol{w}^{(k)}|_2^2-n_k\lambda_X^k|\widehat{\boldsymbol{w}}^{(k)}-\boldsymbol{w}^{(k)}|_1^2\}\\ \ge&c_1\sum_{k=1}^{M}n_k|\widehat{\boldsymbol{w}}^{(k)}-\boldsymbol{w}^{(k)}|_2^2\\ &\underbrace{-c_2\max_{k\le k}|\widehat{\boldsymbol{w}}^{(k)}-\boldsymbol{w}^{(k)}|_1\Big(\sum_{k=1}^{M}|\widehat{\boldsymbol{v}}^{(k)}|_1\lambda_X^k n_k+\sum_{k=1}^{M}\lambda_X^k n_k|\widehat{\boldsymbol{u}}|_1\Big)}_{D_1},\end{aligned} \tag{S7.10}$$

with probability at least $1-(\log p)^{-2}$. By the definition of $\lambda_k\ge n_k\lambda_X^k$, $\lambda_\beta\ge n_{\mathcal{A}_h}\lambda_X^{\mathcal{A}}$ and M is fixed, we have $\lambda_X^k n_k\le\lambda_k$ and $\sum_{k=1}^{M}\lambda_X^k n_k\lesssim\lambda_\beta$. Hence we have $D_1\lesssim\max_{k\le k}|\widehat{\boldsymbol{w}}^{(k)}-\boldsymbol{w}^{(k)}|_1(\sum_{k=1}^{M}|\widehat{\boldsymbol{v}}^{(k)}|_1\lambda_k+\lambda_\beta|\widehat{\boldsymbol{u}}|_1)$. Then,

$$\begin{aligned}&\sum_{k=1}^{M}n_k|\widehat{\boldsymbol{w}}^{(k)}-\boldsymbol{w}^{(k)}|_2^2\\ \lesssim&\lambda_\beta|\widehat{\boldsymbol{u}}|_1+\sum_{k=1}^{M}\lambda_k|\widehat{\boldsymbol{v}}^{(k)}|_1+\max_{k\le k}|\widehat{\boldsymbol{w}}^{(k)}-\boldsymbol{w}^{(k)}|_1\Big(\sum_{k=1}^{M}|\widehat{\boldsymbol{v}}^{(k)}|_1\lambda_k+\lambda_\beta|\widehat{\boldsymbol{u}}|_1\Big)\\ \lesssim&(\max_{k\le M}|\widehat{\boldsymbol{w}}^{(k)}-\boldsymbol{w}^{(k)}|_1\vee 1)\Big(\sum_{k=1}^{M}|\widehat{\boldsymbol{v}}^{(k)}|_1\lambda_k+\lambda_\beta|\widehat{\boldsymbol{u}}|_1\Big).\end{aligned} \tag{S7.11}$$

We also have

$$\begin{aligned}&\sum_{k=1}^{M}n_k(\widehat{\boldsymbol{w}}^{(k)}-\boldsymbol{w}^{(k)})^\top\widehat{\boldsymbol{\Sigma}}^{(0)}(\widehat{\boldsymbol{w}}^{(k)}-\boldsymbol{w}^{(k)})\\ \lesssim&\sum_{k=1}^{M}n_k|\widehat{\boldsymbol{w}}^{(k)}-\boldsymbol{w}^{(k)}|_2^2+\sum_{k=1}^{M}n_k\lambda_X^0|\widehat{\boldsymbol{w}}^{(k)}-\boldsymbol{w}^{(k)}|_1^2\\ \lesssim&\sum_{k=1}^{M}n_k|\widehat{\boldsymbol{w}}^{(k)}-\boldsymbol{w}^{(k)}|_2^2+n_{\mathcal{A}_h}\lambda_X^0|\widehat{\boldsymbol{u}}|_1^2+\sum_{k=1}^{M}n_k\lambda_X^0|\widehat{\boldsymbol{v}}^{(k)}|_1^2\end{aligned}$$

$$
\begin{aligned}
\lesssim & (\max_{k\le M}|\widehat{\boldsymbol{w}}^{(k)}-\boldsymbol{w}^{(k)}|_1\vee 1)\Big(\sum_{k=1}^M|\widehat{\boldsymbol{v}}^{(k)}|_1\lambda_k+\lambda_\beta|\widehat{\boldsymbol{u}}|_1\Big)+n_{\mathcal{A}_h}\lambda_X^0|\widehat{\boldsymbol{u}}|_1^2+\sum_{k=1}^M n_k\lambda_X^0|\widehat{\boldsymbol{v}}^{(k)}|_1^2\\
\le & (\max_{k\le M}|\widehat{\boldsymbol{w}}^{(k)}-\boldsymbol{w}^{(k)}|_1\vee 1)\Big(\sum_{k=1}^M|\widehat{\boldsymbol{v}}^{(k)}|_1\lambda_k+\lambda_\beta|\widehat{\boldsymbol{u}}|_1\Big)+n_{\mathcal{A}_h}\lambda_X^0|\widehat{\boldsymbol{u}}|_1^2+\max_{k\le K}|\widehat{\boldsymbol{v}}^{(k)}|_1\sum_{k=1}^M n_k\tilde{\lambda}_k|\widehat{\boldsymbol{v}}^{(k)}|_1\\
\le & (\max_{k\le M}|\widehat{\boldsymbol{w}}^{(k)}-\boldsymbol{w}^{(k)}|_1\vee 1)\Big(\sum_{k=1}^M|\widehat{\boldsymbol{v}}^{(k)}|_1\lambda_k+\lambda_\beta|\widehat{\boldsymbol{u}}|_1\Big)+n_{\mathcal{A}_h}\lambda_X^0|\widehat{\boldsymbol{u}}|_1^2+\max_{k\le K}|\widehat{\boldsymbol{v}}^{(k)}|_1\sum_{k=1}^M c_1\lambda_k|\widehat{\boldsymbol{v}}^{(k)}|_1\\
\lesssim & (|\widehat{\boldsymbol{u}}|_1\vee\max_{k\le M}|\widehat{\boldsymbol{v}}^{(k)}|_1\vee 1)\Big(\sum_{k=1}^M|\widehat{\boldsymbol{v}}^{(k)}|_1\lambda_k+\lambda_\beta|\widehat{\boldsymbol{u}}|_1\Big)+n_{\mathcal{A}_h}\lambda_X^0|\widehat{\boldsymbol{u}}|_1^2
\end{aligned}
$$

whose LHS is indeed

$$
\begin{aligned}
& n_{\mathcal{A}_h}\widehat{\boldsymbol{u}}^T\widehat{\boldsymbol{\Sigma}}^{(0)}\widehat{\boldsymbol{u}}+\sum_{k=1}^M n_k(\widehat{\boldsymbol{v}}^{(k)})^T\widehat{\boldsymbol{\Sigma}}^{(0)}\widehat{\boldsymbol{v}}^{(k)}-\underbrace{2\sum_{k=1}^M n_k\widehat{\boldsymbol{u}}^T\widehat{\boldsymbol{\Sigma}}^{(0)}\widehat{\boldsymbol{v}}^{(k)}}_{D_2}\\
\ge & c_1 n_{\mathcal{A}_h}|\widehat{\boldsymbol{u}}|_2^2-c_2 n_{\mathcal{A}_h}\lambda_X^0|\widehat{\boldsymbol{u}}|_1^2+c_3\sum_{k=1}^M n_k|\widehat{\boldsymbol{v}}^{(k)}|_2^2-c_4\sum_{k=1}^M n_k\lambda_X^0|\widehat{\boldsymbol{v}}^{(k)}|_1^2-D_2. \qquad \text{(S7.12)}
\end{aligned}
$$

For D_2, we use the fact that

$$
|D_2|\lesssim \lambda_0/n_0\sum_{k=1}^M n_k|\widehat{\boldsymbol{v}}^{(k)}|_1\le\sum_{k=1}^M\lambda_k|\widehat{\boldsymbol{v}}^{(k)}|_1,
$$

where the last inequality is due to $\lambda_0/n_0\le\lambda_k/n_k$. To summarize, we obtain that

$$
n_{\mathcal{A}_h}|\widehat{\boldsymbol{u}}|_2^2+\sum_{k=1}^M n_k|\widehat{\boldsymbol{v}}^{(k)}|_2^2\lesssim(|\widehat{\boldsymbol{u}}|_1\vee\max_{k\le M}|\widehat{\boldsymbol{v}}^{(k)}|_1\vee 1)\Big(\sum_{k=1}^M|\widehat{\boldsymbol{v}}^{(k)}|_1\lambda_k+\lambda_\beta|\widehat{\boldsymbol{u}}|_1\Big)+n_{\mathcal{A}_h}\lambda_X^0|\widehat{\boldsymbol{u}}|_1^2. \qquad \text{(S7.13)}
$$

Based on the oracle inequality, we have

$$
\begin{aligned}
& \lambda_\beta|\widehat{\boldsymbol{\beta}}|_1+\sum_{k=1}^M\lambda_k|\widehat{\boldsymbol{\delta}}^{(k)}|_1\le\lambda_\beta|\boldsymbol{\beta}^*|_1+\sum_{k=1}^M\lambda_k|\boldsymbol{\delta}^{(k)}|_1\\
\Longrightarrow & \lambda_\beta|(\widehat{\boldsymbol{\beta}}-\boldsymbol{\beta}^*)_{S^c}|_1+\sum_{k=1}^M\lambda_k|(\widehat{\boldsymbol{\delta}}^{(k)}-\boldsymbol{\delta}^{(k)})_{H_k^c}|_1\\
& \le\lambda_\beta|(\widehat{\boldsymbol{\beta}}-\boldsymbol{\beta}^*)_S|_1+\sum_{k=1}^M\lambda_k|(\widehat{\boldsymbol{\delta}}^{(k)}-\boldsymbol{\delta}^{(k)})_{H_k}|_1+2\sum_{k=1}^M\lambda_k|\boldsymbol{\delta}^{(k)}_{H_k^c}|_1\\
& \le\lambda_\beta|(\widehat{\boldsymbol{\beta}}-\boldsymbol{\beta}^*)_S|_1+\mathcal{R}_{\boldsymbol{\delta}}+2h\sum_{k=1}^M\lambda_k
\end{aligned}
$$

Case (i). If $\mathcal{R}_{\boldsymbol{\delta}} \geq \lambda_\beta |\widehat{\boldsymbol{u}}_S|_1 + 2h\sum_{k=1}^M \lambda_k$, then

$$\text{RHS of (S7.9)} \lesssim R_\delta.$$

From (S7.13), we have

$$\begin{aligned} n_{\mathcal{A}_h}|\widehat{\boldsymbol{u}}|_2^2 + \sum_{k=1}^M n_k |\widehat{\boldsymbol{v}}^{(k)}|_2^2 &\lesssim (|\widehat{\boldsymbol{u}}|_1 \vee \max_{k\leq M}|\widehat{\boldsymbol{v}}^{(k)}|_1 \vee 1)\mathcal{R}_\delta + n_{\mathcal{A}_h}\lambda_X^0 |\widehat{\boldsymbol{u}}|_1^2 \\ &\lesssim (|\widehat{\boldsymbol{u}}|_1 \vee \max_{k\leq M}|\widehat{\boldsymbol{v}}^{(k)}|_1 \vee 1)\mathcal{R}_\delta + \frac{\mathcal{R}_\delta^2 n_{\mathcal{A}_h}\lambda_X^0}{\lambda_\beta^2}. \end{aligned}$$

And, define $h_0 = h\max_{k\leq M}(\lambda_k/n_k)^{-1} = h\tilde{\lambda}_{\min}^{-1}$, we have

$$\mathcal{R}_\delta \leq h_0^{1/2}\sum_{1\leq k\leq M}\lambda_k|\widehat{\boldsymbol{\delta}}^{(k)} - \boldsymbol{\delta}^{(k)}|_2 \leq Q_{\boldsymbol{\delta}}^{1/2}\sqrt{\sum_{1\leq k\leq M}\frac{\lambda_k^2 h_0}{n_k}}.$$

Then we have

$$n_{\mathcal{A}_h}|\widehat{\boldsymbol{u}}|_2^2 + Q_{\boldsymbol{\delta}} \lesssim (|\widehat{\boldsymbol{u}}|_1 \vee \max_{k\leq M}|\widehat{\boldsymbol{v}}^{(k)}|_1 \vee 1)Q_\delta^{1/2}\sqrt{\sum_{1\leq k\leq M}\frac{\lambda_k^2 h_0}{n_k}} + Q_\delta\sum_{1\leq k\leq M}\frac{\lambda_k^2 h_0 n_{\mathcal{A}_h}\lambda_X^0}{n_k\lambda_\beta^2}$$

Combining with the assumption that $\tilde{\lambda}_{\max}^2\lambda_X^0 h/(\tilde{\lambda}_\beta^2\tilde{\lambda}_{\min}) = o(1)$,

$$\sum_{1\leq k\leq M}\frac{\lambda_k^2 h_0 n_{\mathcal{A}_h}\lambda_X^0}{n_k\lambda_\beta^2} = \sum_{1\leq k\leq M}\frac{n_k^2\tilde{\lambda}_k^2 h_0 n_{\mathcal{A}_h}\lambda_X^0}{n_k n_{\mathcal{A}_h}^2\tilde{\lambda}_\beta^2} \leq \sum_{1\leq k\leq M}\frac{n_k\tilde{\lambda}_k^2 h_0\lambda_X^0}{n_{\mathcal{A}_h}\tilde{\lambda}_\beta^2} \leq \tilde{\lambda}_{\max}^2\lambda_X^0 h/(\tilde{\lambda}_\beta^2\tilde{\lambda}_{\min}),$$

We have that

$$Q_\delta \leq n_{\mathcal{A}_h}(1 \vee \max_{k\leq M}|\widehat{\boldsymbol{w}}^{(k)} - \boldsymbol{w}^{(k)}|_1 \vee \max_{k\leq M}|\widehat{\boldsymbol{v}}^{(k)}|_1)^2\tilde{\lambda}_{\max}^2 h_0.$$

And

$$|\widehat{\boldsymbol{u}}|_2^2 \leq (1 \vee \max_{k\leq M}|\widehat{\boldsymbol{w}}^{(k)} - \boldsymbol{w}^{(k)}|_1 \vee \max_{k\leq M}|\widehat{\boldsymbol{v}}^{(k)}|_1)^2\tilde{\lambda}_{\max}^2 h_0.$$

$$|\widehat{\boldsymbol{u}}|_1 \leq \frac{\mathcal{R}_\delta}{\lambda_\beta} \leq \frac{Q_\delta^{1/2}\sqrt{\sum_{1\leq k\leq M}\frac{\lambda_k^2 h_0}{n_k}}}{\lambda_\beta} \leq \frac{n_{\mathcal{A}_h}(1 \vee \max_{k\leq M}|\widehat{\boldsymbol{w}}^{(k)} - \boldsymbol{w}^{(k)}|_1 \vee \max_{k\leq M}|\widehat{\boldsymbol{v}}^{(k)}|_1)\tilde{\lambda}_{\max}^2 h_0}{\lambda_\beta}.$$

$$\begin{aligned} \max_{k\leq M}|\widehat{\boldsymbol{v}}^{(k)}|_1 \leq& \frac{\mathcal{R}_\delta}{n_{\min}\tilde{\lambda}_{\max}} \leq \frac{Q_\delta^{1/2}\sqrt{\sum_{1\leq k\leq M}\frac{\lambda_k^2 h_0}{n_k}}}{n_{\min}\tilde{\lambda}_{\max}} \\ \leq& \frac{n_{\mathcal{A}_h}(1 \vee \max_{k\leq M}|\widehat{\boldsymbol{w}}^{(k)} - \boldsymbol{w}^{(k)}|_1 \vee \max_{k\leq M}|\widehat{\boldsymbol{v}}^{(k)}|_1)\tilde{\lambda}_{\max} h_0}{n_{\min}}. \end{aligned}$$

Hence,

$$|\widehat{\boldsymbol{u}}|_1 \le (1 \vee |\widehat{\boldsymbol{u}}|_1 \vee \max_{k\le M} |\widehat{\boldsymbol{v}}^{(k)}|_1)\tilde{\lambda}_{\max}^2 h/(\tilde{\lambda}_\beta \tilde{\lambda}_{\min}),$$

$$\max_{k\le M} |\widehat{\boldsymbol{v}}^{(k)}|_1 \le (1 \vee |\widehat{\boldsymbol{u}}|_1 \vee \max_{k\le M} |\widehat{\boldsymbol{v}}^{(k)}|_1)\tilde{\lambda}_{\max} h n_{\mathcal{A}_h}/(n_{\min}\tilde{\lambda}_{\min}),$$

which combining with $\tilde{\lambda}_{\max}^2 h/(\tilde{\lambda}_\beta \tilde{\lambda}_{\min}) = o(1)$ and $\tilde{\lambda}_{\max} h n_{\mathcal{A}_h}/(n_{\min}\tilde{\lambda}_{\min}) = o(1)$ implies that

$$|\widehat{\boldsymbol{u}}|_1 \le \tilde{\lambda}_{\max}^2 h/(\tilde{\lambda}_\beta \tilde{\lambda}_{\min}), \quad |\widehat{\boldsymbol{u}}|_2^2 \le \tilde{\lambda}_{\max}^2 h/\tilde{\lambda}_{\min}.$$

Case (ii). If $R_\delta + 2h\sum_{k=1}^M \lambda_k \le \lambda_\beta |(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*)_S|_1$, then

$$\text{RHS of (S7.9)} \le 3\lambda_\beta |(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*)_S|_1.$$

Moreover, $|(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*)_{S^c}| \le 2|(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*)_S|_1$ and $\sum_{1\le k\le M} \lambda_k |\widehat{\boldsymbol{\delta}}^{(k)} - \boldsymbol{\delta}^{(k)}|_1 \le 3\lambda_\beta |(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*)_S|_1$. Similar to case (i), we have

$$n_{\mathcal{A}_h}|\widehat{\boldsymbol{u}}|_2^2 + \sum_{k=1}^M n_k (\widehat{\boldsymbol{v}}^{(k)})^T \widehat{\boldsymbol{\Sigma}}_{\widehat{\boldsymbol{\beta}}}^{(0)} \widehat{\boldsymbol{v}}^{(k)} \lesssim (1 \vee |\widehat{\boldsymbol{u}}|_1 \vee \max_{k\le M} |\widehat{\boldsymbol{v}}^{(k)}|_1)\lambda_\beta |\widehat{\boldsymbol{u}}_S|_1 + n_{\mathcal{A}_h}\lambda_X^0 |\widehat{\boldsymbol{u}}|_1^2.$$

Combining with the assumption that $s_0\lambda_X^0 \le c$ for some sufficiently small constant c, we have

$$n_{\mathcal{A}_h}|\widehat{\boldsymbol{u}}|_2^2 + \sum_{k=1}^M n_k (\widehat{\boldsymbol{v}}^{(k)})^T \widehat{\boldsymbol{\Sigma}}_{\widehat{\boldsymbol{\beta}}}^{(0)} \widehat{\boldsymbol{v}}^{(k)} \lesssim (1 \vee |\widehat{\boldsymbol{u}}|_1 \vee \max_{k\le M} |\widehat{\boldsymbol{v}}^{(k)}|_1)\lambda_\beta |\widehat{\boldsymbol{u}}_S|_1.$$

Rearranging the terms, we arrive at

$$|\widehat{\boldsymbol{u}}|_2^2 \lesssim s\tilde{\lambda}_\beta^2 (1 \vee |\widehat{\boldsymbol{u}}|_1 \vee \max_{k\le M} |\widehat{\boldsymbol{v}}^{(k)}|_1)^2.$$

Using the constraint in this case, we have

$$|\widehat{\boldsymbol{u}}|_1 \lesssim |\widehat{\boldsymbol{u}}_S|_1 \lesssim s\tilde{\lambda}_\beta (1 \vee |\widehat{\boldsymbol{u}}|_1 \vee \max_{k\le M} |\widehat{\boldsymbol{v}}^{(k)}|_1), \quad \max_{k\le M} |\widehat{\boldsymbol{v}}^{(k)}|_1 \le \frac{s\tilde{\lambda}_\beta^2 n_{\mathcal{A}_h}}{\tilde{\lambda}_{\max} n_{\min}} (1 \vee |\widehat{\boldsymbol{u}}|_1 \vee \max_{k\le M} |\widehat{\boldsymbol{v}}^{(k)}|_1),$$

which combining with $s\tilde{\lambda}_\beta = o(1)$ and $s\tilde{\lambda}_\beta^2 n_{\mathcal{A}_h}/(\tilde{\lambda}_{\max} n_{\min}) = o(1)$, implies that

$$|\widehat{\boldsymbol{u}}|_2^2 \lesssim s\tilde{\lambda}_\beta^2, \quad |\widehat{\boldsymbol{u}}|_1 \le s\tilde{\lambda}_\beta, \quad \max_{k\le M} |\widehat{\boldsymbol{v}}^{(k)}|_1 \le \frac{s\tilde{\lambda}_\beta^2 n_{\mathcal{A}_h}}{\tilde{\lambda}_{\max} n_{\min}}.$$

Case (iii) If $2h\sum_{k=1}^{M}\lambda_k \geq \mathcal{R}_\delta + \lambda_\beta|(\widehat{\boldsymbol{\beta}}-\boldsymbol{\beta}^*)_S|_1$, then

$$\text{RHS of (A.3)} \leq 6h\sum_{k=1}^{M}\lambda_k, \quad \mathcal{R}_\delta \leq 2h\sum_{k=1}^{M}\lambda_k.$$

In this case, we have

$$n_{\mathcal{A}_h}|\widehat{\boldsymbol{u}}|_2^2+\sum_{k=1}^{M} n_k(\widehat{\boldsymbol{v}}^{(k)})^\top \widehat{\boldsymbol{\Sigma}}^{(0)}_{\widehat{\boldsymbol{\beta}}}\widehat{\boldsymbol{v}}^{(k)} \lesssim h\sum_{k=1}^{M}\lambda_k+(1\vee\max_k|\widehat{\boldsymbol{v}}^{(k)}|_1\vee|\widehat{\boldsymbol{u}}|_1)(\lambda_\beta|\widehat{\boldsymbol{u}}_S|_1+\mathcal{R}_\delta)+n_{\mathcal{A}_h}\lambda_X^0|\widehat{\boldsymbol{u}}|_1^2.$$

Under the constraint in this case, we have

$$\begin{aligned}&n_{\mathcal{A}_h}|\widehat{\boldsymbol{u}}|_2^2+\sum_{k=1}^{M} n_k(\widehat{\boldsymbol{v}}^{(k)})^T \widehat{\boldsymbol{\Sigma}}^{(0)}_{\widehat{\boldsymbol{\beta}}}\widehat{\boldsymbol{v}}^{(k)}\\ \lesssim& h\sum_{k=1}^{M}\lambda_k(1\vee\max_k|\widehat{\boldsymbol{v}}^{(k)}|_1\vee|\widehat{\boldsymbol{u}}|_1)+n_{\mathcal{A}_h}\lambda_X^0|\widehat{\boldsymbol{u}}|_1^2\end{aligned}$$

Recall that $h\sum_{k=1}^{M}\lambda_k \leq c_\lambda n_{\mathcal{A}_h}h\tilde{\lambda}_{\max}$, we arrive at

$$|\widehat{\boldsymbol{u}}|_2^2 \lesssim h\tilde{\lambda}_{\max}(1\vee\max_k|\widehat{\boldsymbol{v}}^{(k)}|_1\vee|\widehat{\boldsymbol{u}}|_1)+\lambda_X^0|\widehat{\boldsymbol{u}}|_1^2$$

$$|\widehat{\boldsymbol{u}}|_1 \lesssim h\sum_k \lambda_k/\lambda_\beta \lesssim h\tilde{\lambda}_{\max}/\tilde{\lambda}_\beta, \quad \max_{k\leq M}|\widehat{\boldsymbol{v}}^{(k)}|_1 \leq hn_{\mathcal{A}_h}/n_{\min}.$$

Hence,

$$|\widehat{\boldsymbol{u}}|_2^2 \leq h\tilde{\lambda}_{\max}+\frac{h^2\tilde{\lambda}_{\max}^2\lambda_X^0}{\tilde{\lambda}_\beta^2} \lesssim h\tilde{\lambda}_{\max},$$

where the last inequality is due to that $h\tilde{\lambda}_{\max}^2\lambda_X^0/\tilde{\lambda}_\beta^2 = o(\tilde{\lambda}_{\min})$.

To summarize, we have obtained that

$$|X(\widehat{\boldsymbol{\beta}}-\boldsymbol{\beta}^*)|_2^2/n \vee |\widehat{\boldsymbol{u}}|_2^2 \lesssim h\tilde{\lambda}_{\max}^2/\tilde{\lambda}_{\min}+s_0\tilde{\lambda}_\beta^2, \quad |\widehat{\boldsymbol{u}}|_1 \lesssim s_0\tilde{\lambda}_\beta + h\tilde{\lambda}_{\max}^2/(\tilde{\lambda}_\beta\tilde{\lambda}_{\min}).$$

Under the assumption that $n_k \asymp n_j$ for all $1 \leq k \leq j$ and that M is fixed, it follows that $n_{\min} \asymp n_{\max} \asymp N \gg n_0$. Consequently,

$$\tilde{\lambda}_{\max} \asymp \tilde{\lambda}_{\min} \asymp \tilde{\lambda}_0,$$

yielding the desired results.

Bibliography

Basu, S. and Michailidis, G. (2015). Regularized estimation in sparse high-dimensional time series models. *The Annals of Statistics*, 43:1535–1567.

Bühlmann, P. and van de Geer, S. (2011). *Statistics for High-Dimensional Data: Methods, Theory and Applications*. Springer, New York.

Grenander, U. and Szegö, G. (1958). *Toeplitz Forms and Their Applications*. Cambridge University Press, London.

Javanmard, A. and Montanari, A. (2014). Confidence intervals and hypothesis testing for high-dimensional regression. *Journal of Machine Learning Research*, 15:2869–2909.

Li, S. (2020). Debiasing the debiased lasso with bootstrap. *Electronic Journal of Statistics*, 14:2298–2337.

Li, S., Cai, T. T., and Li, H. (2022). Transfer learning for high-dimensional linear regression: prediction, estimation and minimax optimality. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 84:149–173.

Li, S., Zhang, L., Cai, T. T., and Li, H. (2024). Estimation and inference for high-dimensional generalized linear models with knowledge transfer. *Journal of the American Statistical Association*, 119:1274–1285.

Liu, W. and Wu, W. B. (2010). Asymptotics of spectral density estimates. *Econometric Theory*, 26:1218–1245.

Loh, P.-L. and Wainwright, M. J. (2012). High-dimensional regression with noisy and missing data: Provable guarantees with nonconvexity. *The Annals of Statistics*, 40:1637–1664.

Politis, D. N., Romano, J. P., and Wolf, M. (1999). *Subsampling*. Springer, New York.

Rudelson, M. and Vershynin, R. (2013). Hanson-Wright inequality and sub-gaussian concentration. *Electronic Communications in Probability*, 18:1–9.

Tian, Y. and Feng, Y. (2023). Transfer learning under high-dimensional generalized linear models. *Journal of the American Statistical Association*, 118:2684–2697.

Wu, W. B. and Pourahmadi, M. (2009). Banding sample autocovariance matrices of stationary processes. *Statistica Sinica*, 19:1755–1768.

Wu, W. B. and Wu, Y. (2016). Performance bounds for parameter estimates of high-dimensional linear models with correlated errors. *Electronic Journal of Statistics*, 10:352–379.

Yuan, P. and Guo, X. (2022). High-dimensional inference for linear model with correlated errors. *Metrika*, 85:21–52.