STATISTICAL INFERENCE FOR HIGH-DIMENSIONAL LINEAR REGRESSION WITH BLOCKWISE MISSING DATA

Fei Xue[1], Rong Ma[2], and Hongzhe Li[3]

[1]*Purdue University,* [2]*Stanford University and* [3]*University of Pennsylvania*

Supplementary Material

In this Supplementary Material, we provide additional tables and discussion, as well as proofs for all theorems.

S1 Notations

Notation	Definition
$\mathcal{X}$	Random vector of regression covariates
$\mathcal{Y}$	Response variable
ϵ	Error term
$\boldsymbol{\beta}$	Regression coefficient vector
p	Number of all covariates
s	Number of relevant covariates
σ	Standard deviation of the error term

Notation	Definition
S	Number of data sources
$\mathcal{D}$	Index set of all samples
$\mathcal{D}_1$	Index set of samples without response values
$\mathcal{D}_2$	Index set of samples with response values
y	Vector consisting of all samples of the response variable
$\boldsymbol{X}$	Design matrix
N	Number of samples in $\mathcal{D}_1$
n	Number of samples in $\mathcal{D}_2$
R	Number of missing groups
ξ_i	Random group label of the i-th sample
$\mathcal{S}(r)$	Index set of the samples in Group r
$\mathcal{G}(r)$	Index set of the groups where missing variables of Group r and variables in at least one of the other data sources are observed
$a(r)$	Index set of the observed variables in Group r
$a(r)^c$	Index set of the missing variables in Group r
$a(r,k)$	Index set of covariates which are observed in Groups r and k
$\boldsymbol{X}_i$	The i-th sample, that is, the i-th row of X
X_{ij}	The i-th sample of the j-th covariate
$\boldsymbol{X}_{ia(r,k)}$	Vector consisting of the i-th sample of covariates indexed by $a(r,k)$
$X_{ij}^{(k)}$	$E(X_{ij} \mid \boldsymbol{X}_{ia(r,k)})$ if X_{ij} is missing; otherwise X_{ij}
	Imputed vector for the i-th sample based on Group k, that is,

Notation	Definition
$\boldsymbol{X}_i^{(k)}$	$(X_{i1}^{(k)}, \ldots, X_{ip}^{(k)})^T$
$\widehat{\boldsymbol{\gamma}}_{j,a(r,k)}$	Estimate of the coefficient vector for the relationship between X_{ij} and $\boldsymbol{X}_{ia(r,k)}$ for $i \in \mathcal{S}(k)$
$\boldsymbol{e}_j$	A p-dimensional vector with 1 as the j-th element and 0 otherwise
$\widehat{X}_{ij}^{(k)}$	$\widehat{\boldsymbol{\gamma}}_{j,a(r,k)}^T \boldsymbol{X}_{ia(r,k)}$ if X_{ij} is missing; otherwise X_{ij}
$\widehat{\boldsymbol{X}}_i^{(k)}$	Actual imputed vector for the i-th sample based on Group k, that is, $(\widehat{X}_{i1}^{(k)}, \ldots, \widehat{X}_{ip}^{(k)})^T$
$\boldsymbol{X}_{ia(k)}^{(k)}$	Sub-vector of $\boldsymbol{X}_i^{(k)}$ corresponding to all the covariates observed in Group k
$\widehat{\boldsymbol{X}}_{ia(k)}^{(k)}$	Sub-vector of $\widehat{\boldsymbol{X}}_i^{(k)}$ corresponding to all the covariates observed in Group k
$\hat{\theta}_r$	Estimate of observed rate for the r-th group among $\mathcal{D}_2$, that is, $\lvert\mathcal{D}_2 \cap \mathcal{S}(r)\rvert / \lvert\mathcal{D}_2\rvert$
$\boldsymbol{g}(\boldsymbol{\beta})$	Vector of all estimating equations with conditional expectations
$\boldsymbol{g}_n(\boldsymbol{\beta})$	Vector of all estimating equations with actual imputed values
$\widehat{\boldsymbol{\beta}}$	Proposed estimator of $\boldsymbol{\beta}$
$\boldsymbol{g}_n^*(\boldsymbol{\beta})$	Sub-vector of $\boldsymbol{g}_n(\boldsymbol{\beta})$ consisting of estimating equations with fewer imputed values
$\boldsymbol{g}^*(\boldsymbol{\beta})$	Population counterpart of $\boldsymbol{g}_n^*(\boldsymbol{\beta})$
$\boldsymbol{G}_n(\boldsymbol{\beta})$	Matrix defined by $\frac{d}{d\boldsymbol{\beta}} \boldsymbol{g}_n^*(\boldsymbol{\beta})$
$\boldsymbol{G}(\boldsymbol{\beta})$	Matrix defined by $\frac{d}{d\boldsymbol{\beta}} \boldsymbol{g}^*(\boldsymbol{\beta})$
$\widehat{\boldsymbol{v}}_j$	Projection vector for the j-th coefficient

Notation	Definition
$\hat{S}_j(\boldsymbol{\beta})$	Projected estimating function for the j-th coefficient
$\hat{\boldsymbol{v}}_{j,rk}$	Subvector of the projection vector $\hat{\boldsymbol{v}}_j$ corresponding to the estimating functions in $\boldsymbol{g}_n^*(\boldsymbol{\beta})$ associated to Group $k \in \mathcal{G}(r)$
$\tilde{\beta}_j$	Bias-corrected estimator of β_j
$z_{\alpha/2}$	Upper $\alpha/2$-quantile of the standard normal distribution
T_j	Test statistic for null hypothesis H_0: $\beta_j = b_j$
$\boldsymbol{\Sigma}^{(r,k)}$	Matrix defined as $E[I\{\xi_i = r\}\boldsymbol{X}_i^{(k)}(\boldsymbol{X}_i^{(k)})^T]$
N_r	Number of samples in $\mathcal{D}_1 \cap \mathcal{S}(r)$
n_r	Number of samples in $\mathcal{D}_2 \cap \mathcal{S}(r)$
$\boldsymbol{\Sigma}$	Covariance matrix of $\boldsymbol{X}_i$
$\lambda_{\min}(\cdot)$	The smallest singular value of a matrix
$\lambda_{\max}(\cdot)$	The largest singular value of a matrix
ω_{ij}	The (i,j) element of Σ^{-1}
$\widehat{X}$	Set of all the imputed observations
β_s	Signal strength in simulations
p_i	Number of total covariates in the i-th data source
s_i	Number of relevant covariates in the i-th data source in simulations
T	Number of testing responses in the real data application

Table 3: Notations.

S2 Additional simulation results

To estimate the asymptotic variance s_j, we recommend using the moment estimator

$$\widehat{\sigma}^2 = \frac{\sum_{i\in\mathcal{D}_2}\sum_{k\in\mathcal{G}(r),1\le r\le R}\left[I(\xi_i=r)\{y_i-(\widehat{\boldsymbol{X}}_i^{(k)})^T\widehat{\boldsymbol{\beta}}\}\right]^2}{\sum_{k\in\mathcal{G}(r),1\le r\le R}|\mathcal{D}_2\cap\mathcal{S}(r)|} + \max_{\substack{1\le r\le R\\ k\in\mathcal{G}(r)}}\widehat{\boldsymbol{\beta}}_{a(r)^c}^\top\widehat{\boldsymbol{\Omega}}_{r,k}\widehat{\boldsymbol{\beta}}_{a(r)^c}, \quad \text{(S2.1)}$$

for the parameter $\sigma_{r,k}^2$, where $\widehat{\boldsymbol{\Omega}}_{r,k}$ is the sample covariance matrix of the fitted residuals $\{\widehat{\boldsymbol{\epsilon}}_{ia(r)^c}^{(k)} = \boldsymbol{X}_{ia(r)^c} - \widehat{\boldsymbol{\Gamma}}_{r,k}^\top\boldsymbol{X}_{ia(r,k)} : i\in\mathcal{S}(k), k\in\mathcal{G}(r)\}$ from the imputation step.

We implement the multivariate imputation by chained equations (MICE) method using the R package `mice`[1], and apply the debiased Lasso method and the Lasso projection method to each MICE-imputed dataset under simulation Setting 3. We then pool the estimates based on all MICE-imputed datasets by taking averages according to Rubin's rules (Rubin, 1987) for the debiased Lasso method and the Lasso projection method, respectively.

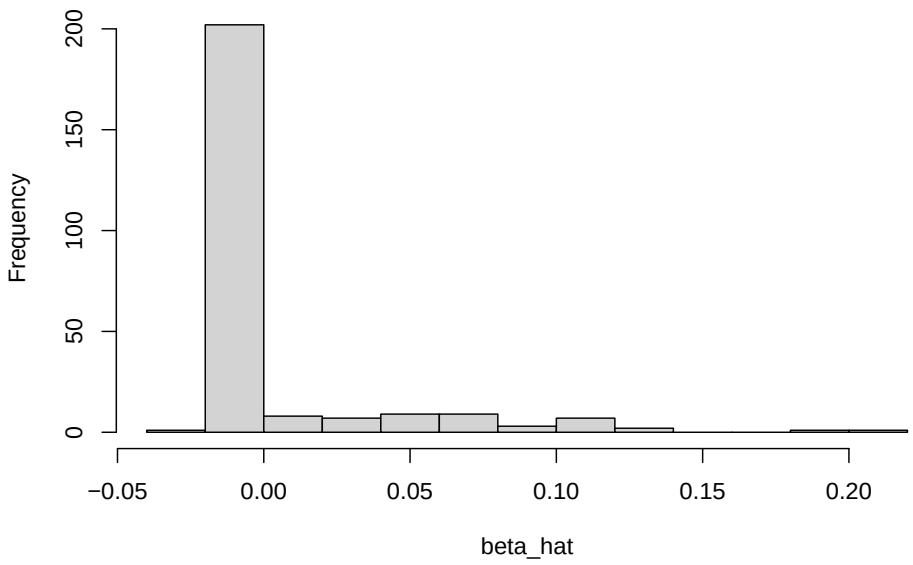
The bias and variance of the corresponding estimators are provided in the last two rows of Table 4. The results show that MICE-based methods produce much larger biases than the proposed method. This is possibly due to that the imputations by the proposed method are more accurate than those by MICE, since the MICE imputation does not fully use the block-wise missing structure, but the blockwise imputation (BI) in the proposed method does. When correlations between covariates are larger ($\rho = 0.3$), empirical standard deviations of these MICE-based methods increases. In addition, we observe that when correlations between covariates are larger ($\rho = 0.3$), empirical standard deviations of these MICE-based methods increases. Similarly, the empirical standard deviations of most other methods also increase as ρ increases.

[1] `https://cran.r-Bproject.org/web/packages/glmnet/index.html`

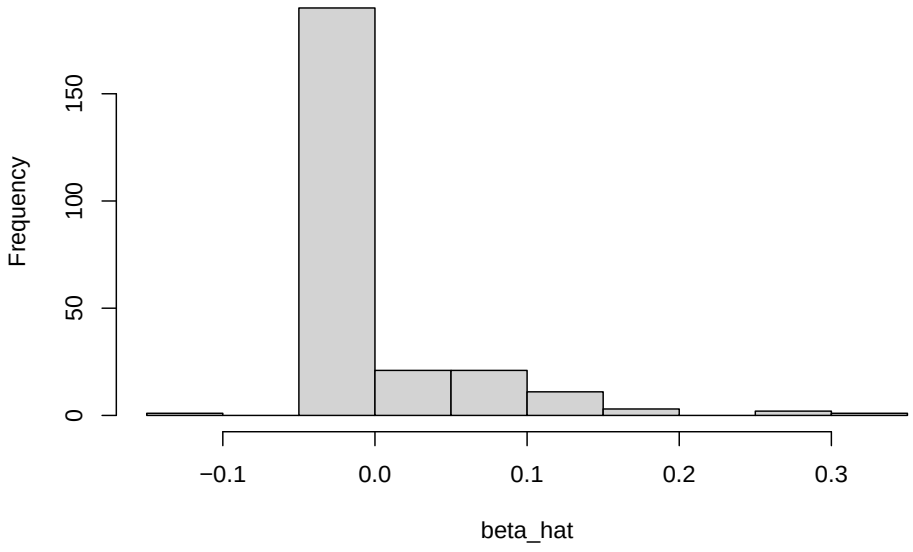
Moreover, as shown in Table 4, methods using single imputation (SI) or MICE (that is, DL-SI, LP-SI, DL-MICE, and LP-MICE) produce larger biases than the proposed method under both $\rho = 0.1$ and $\rho = 0.3$. This is possibly due to that the imputations by the proposed method are more accurate than those by SI or MICE. Compared to the proposed method, the bias of the Lasso projection method with complete cases (LP-CC) is smaller, which is probably because there is no imputation bias in LP-CC. However, since the number of complete cases is limited, LP-CC produces larger empirical standard deviation than the proposed method. In contrast, the debiased Lasso method with complete cases (DL-CC) has large bias and small standard deviation, suggesting that it fails to remove the leading bias of the initial estimators. This is likely due to the limited sample size compared to the sparsity level and the dimensionality, which violates its underlying assumption $n \gg \frac{s^2}{\log^2 p}$ (Javanmard and Montanari, 2014) required for successful bias correction.

Table 4: Averages of the absolute values of empirical bias and empirical standard deviation under Setting 3 based on 250 replications. Proposed: the proposed method. DL-CC: the debiased Lasso method with complete cases. LP-CC: the Lasso projection method with complete cases. DL-SI: the debiased Lasso method with single regression imputation. LP-SI: the Lasso projection method with single regression imputation.

	$\rho = 0.1$		$\rho = 0.3$	
Method	Bias	Standard Deviation	Bias	Standard Deviation
Proposed	0.064	0.158	0.084	0.160
DL-CC	0.174	0.086	0.160	0.106
LP-CC	0.018	0.342	0.005	0.347
DL-SI	0.181	0.034	0.182	0.033
LP-SI	0.141	0.066	0.136	0.068
DL-MICE	0.159	0.125	0.160	0.144
LP-MICE	0.159	0.122	0.160	0.140



(a)



(b)

Figure 1: Histograms of $\widehat{\boldsymbol{\beta}}_j$ under Setting 3 based on 250 replications for different correlations ρ's, where the j-th covariate is a relevant covariate with true signal $\boldsymbol{\beta}_j = 0.2$. (a) $\rho = 0.1$. (b) $\rho = 0.3$.

In addition, we calculate absolute values of empirical biases of $\widehat{\boldsymbol{\beta}}_j$ and $\widetilde{\boldsymbol{\beta}}_j$ for the j-th covariate under Setting 3, and provide the results in Table 5, which shows that the empirical bias of $\widehat{\boldsymbol{\beta}}_j$ is much larger than that of $\widetilde{\boldsymbol{\beta}}_j$. This implies that the bias-correction procedure improves the finite sample performance, which is the main advantage of the approach in Section 2.3 over the one in Section 2.2. Moreover, we provide histograms of $\widehat{\boldsymbol{\beta}}_j$ based on 250 replications in Figure 1, indicating that the empirical distribution of $\widehat{\boldsymbol{\beta}}_j$ is right-skewed and contains a large portion of point-mass at 0. Thus, the asymptotic distribution of $\widehat{\boldsymbol{\beta}}_j$ is likely to be an asymmetric, non-Gaussian (or, in general non-standard) distribution.

Table 5: Averages of the absolute values of empirical biases of $\widehat{\boldsymbol{\beta}}_j$ and $\widetilde{\boldsymbol{\beta}}_j$ for the j-th covariate under Setting 3 based on 250 replications.

Estimator	$\rho = 0.1$	$\rho = 0.3$
$\widetilde{\boldsymbol{\beta}}_j$	0.064	0.084
$\widehat{\boldsymbol{\beta}}_j$	0.188	0.183

S3 Additional results for ADNI Data Set

Table 6: Biomarkers identified by all the methods.

Method	Biomarkers selected
Proposed	ST101SV, ST104TS, ST107TS, ST110TS, ST121TS, ST124SV, ST26TA, ST30SV, ST35TA, ST57TS, ST58TA, ST60TA, ST68SV, ST6SV, ST83CV, ST84CV, ST84SA, ST84TA, ST94CV, CTX_RH_TEMPORALPOLE, LEFT_CHOROID_PLEXUS, PJA2, SFRP1, P4HA3, ABCG1, IKZF5, NLRP10, ACOXL, DNAH2, EIF4ENIF1, TEKT4, PPIL2, TRPM8, SIGMAR1, ANKRD13C, MAGI2
DL-CC	ST129TS, ST15CV, ST15SA, ST147SV, ST24TA, ST26TA, ST29SV, ST32CV, ST34CV, ST40TA, ST45CV, ST48TA, ST60TA, ST60TS, ST72TA, ST73CV, ST76SV, ST77SV, ST84SA, ST85CV, ST93CV, ST97TA, ST97TS, CC_ANTERIOR, CTX_LH_CUNEUS, CTX_LH_LINGUAL, CTX_LH_PERICALCARINE, CTX_RH_CUNEUS, SUMMARYSUVR_WHOLECEREBNORM_1.11CUTOFF, SUMMARYSUVR_COMPOSITE_REFNORM, PHF1, SFRP1, GOLGA8A || GOLGA8B, IKZF5, MECR, NLRP10, PROKR2, TAS2R4, DAZAP2, SERPINH1, LDLR, TTYH1, PSMB2, MAGI2, PCDH9, POSTN
LP-CC	ST147SV, ST32CV, SFRP1
DL-SI	ST129TA, ST130TS, ST18SV, ST24TA, ST26TA, ST29SV, ST30SV, ST31CV, ST31TA, ST32CV, ST40CV, ST40TA, ST46TA, ST48TA, ST58CV, ST58TA, ST59CV, ST80SV, ST83CV, ST85CV, ST89SV, ST90CV, ST90TA, CC_POSTERIOR, CTX_LH_FUSIFORM, LEFT_LATERAL_VENTRICLE, SUMMARYSUVR_WHOLECEREBNORM_1.11CUTOFF, SUMMARYSUVR_COMPOSITE_REFNORM, CTX_LH_FRONTALPOLE, COL4A1, PHF1, SFRP1, SCFD2, ABCG1, CDH2, NAALAD2, TRIM6-TRIM34 || TRIM34, IKZF5, MECR, OXT, NLRP10, PROKR2, ACOXL, TAS2R4, BRDT, CACYBP, PCSK6, DNAH2, SERPINH1, EIF4ENIF1, TEKT4, STXBP1, PPIL2, ABHD14B, LDLR, TRPM8, PRMT6, SIGMAR1, ZNF195, PSMB2, CALD1, MAGI2, PYCR2
LP-SI	ST107TS, ST30SV, ST39SA, ST46TA, CTX_RH_PARAHIPPOCAMPAL, SUMMARYSUVR_COMPOSITE_REFNORM, SFRP1, DNAH2, EIF4ENIF1, TEKT4, TRPM8, SIGMAR1, MAGI2

Table 7: Biomarkers identified by the proposed method and one of other methods.

Method	**Overlapped biomarkers**
DL-CC	ST26TA, ST60TA, ST84SA, SFRP1, IKZF5, NLRP10, MAGI2
LP-CC	SFRP1
DL-SI	ST26TA, ST30SV, ST58TA, ST83CV, SFRP1, ABCG1, IKZF5, NLRP10, ACOXL, DNAH2, EIF4ENIF1, TEKT4, PPIL2, TRPM8, SIGMAR1, MAGI2
LP-SI	ST107TS, ST30SV, SFRP1, DNAH2, EIF4ENIF1, TEKT4, TRPM8, SIGMAR1, MAGI2

Table 8: Averages of absolute mean and standard deviation of differences between true responses and predicted values based on 150 replications. Proposed ($\widehat{\boldsymbol{\beta}}$): the proposed method with the estimator $\widehat{\boldsymbol{\beta}}$. MRI Lasso, PET Lasso, and Gene Lasso: Lasso method using only MRI, PET, and gene expression variables, respectively. CC Lasso: the Lasso method using only complete cases. Naive mean: using the sample mean of the response variable in the training sets for prediction.

Method	Absolute Mean	Standard Deviation
Proposed ($\widehat{\boldsymbol{\beta}}$)	0.715	3.662
MRI Lasso	0.837	3.814
PET Lasso	0.715	4.018
Gene Lasso	0.791	4.369
CC Lasso	0.857	4.321
Naive mean	0.787	4.401

The results in Table 9 show that the proposed method produces the smallest squared bias among all the methods. In addition, the proposed method performs the best in terms of sum of the squared bias and variance.

Table 9: Squared bias and variance of predicted values by each method based on 150 replications. Proposed ($\widehat{\boldsymbol{\beta}}$): the proposed method with the estimator $\widehat{\boldsymbol{\beta}}$. MRI Lasso, PET Lasso, and Gene Lasso: Lasso method using only MRI, PET, and gene expression variables, respectively. CC Lasso: the Lasso method using only complete cases. Naive mean: using the sample mean of the response variable in the training sets for prediction.

Method	Squared bias	Variance
Proposed ($\widehat{\boldsymbol{\beta}}$)	11.910	2.191
MRI Lasso	14.930	0.583
PET Lasso	16.472	0.479
Gene Lasso	15.929	3.999
CC Lasso	17.079	3.063
Naive mean	20.970	0.009

S4 Theoretical results for only supervised samples

When there are only n supervised samples, our analysis suggests that we can randomly select m samples from all the supervised samples preserving the blockwise missing pattern of the entire dataset, perform the imputation step in (2.2) in the manuscript using the selected samples, and construct $\widehat{\boldsymbol{\beta}}$ and $\widetilde{\boldsymbol{\beta}}$ using the remaining samples. Let $\mathcal{A}_1$ denote the index set of the selected samples, and $\mathcal{A}_2$ denote the index set of the remaining samples. Similar to Theorem 2 in the manuscript, under regularity conditions (A1)–(A5), $m \gtrsim (n-m)\log p$, $\log p \ll n-m$, $\tau \asymp \sqrt{\log p/m}$, $\lambda \asymp \sqrt{\log p/(n-m)} + s\sqrt{\log p/m}$, $\lambda' \asymp \sqrt{\log p/(n-m)}$, and $s \ll \min\big\{\frac{\sqrt{n-m}}{\log p}, \sqrt{\frac{m}{(n-m)\log p}}\big\}$, for sufficiently large n, m, p and each $j \in [1:p]$, we have

$$(n-m)(\tilde{\beta}_j - \beta_j)/s_j = AB + D,$$

where

$$s_j^2 = \sum_{i\in\mathcal{A}_2} \sum_{k\in\mathcal{G}(r),1\le r\le R} \frac{|\mathcal{A}_2|^2\sigma_{r,k}^2}{|\mathcal{A}_2\cap\mathcal{S}(r)|^2} I\{\xi_i = r\}(\hat{\boldsymbol{v}}_{j,rk}^\top \boldsymbol{X}_{ia(r,k)})^2, \tag{S4.2}$$

$A \to 1$ and $D \to 0$ in probability, and $B|\widehat{X} \to N(0,1)$ in distribution, in which $\widehat{X} = \{\widehat{\boldsymbol{X}}_i^{(k)}\}_{i\in\mathcal{A}_2}$ is the set of all the imputed observations in $\mathcal{A}_2$. The $\sigma_{r,k}^2$ and $\hat{\boldsymbol{v}}_{j,rk}$ are defined in the paragraph following equation (4.18) in the manuscript. Note that s_j here is of order $\sqrt{n-m}$ by the above equation (S4.2).

When there are both supervised and unsupervised samples of sizes n and N, respectively, by Theorem 2 in the manuscript, the convergence rate (i.e., the order of standard error) of $\tilde{\beta}_j$ is $1/\sqrt{n}$. In contrast, when there are only n supervised samples, the convergence rate of $\tilde{\beta}_j$ is $1/\sqrt{n-m}$ by the above paragraph, which is slower than $1/\sqrt{n}$ for diverging p since $n/(n-m) \gtrsim \log p + 1$ by $m \gtrsim (n-m)\log p$. The asymptotic variances of $\tilde{\beta}_j$ under the two situations are similar averages by equation (4.18) in the manuscript and above equation (S4.2).

S5 Proof of Theorem 1

The proof of the following theorem can be separated into two parts. In the first part, we show that the constraint $\|\boldsymbol{g}_n(\boldsymbol{\beta})\|_\infty \le \lambda$ is feasible with high probability. In the second part, we obtain the rate of convergence of the proposed estimator $\widehat{\boldsymbol{\beta}}$. For simplicity, and without loss of generality, we assume that there are $2n$ samples in $\mathcal{D}_2$, where half of the samples along with the N samples in $\mathcal{D}_1$ are used for imputation, whereas the other half of the samples are used for constructing the estimating function.

Throughout, we adopt the following notations. Recall that $\widehat{\boldsymbol{\gamma}}_{j,a(r,k)}$ is the Lasso estimator

defined in the main paper, we define the matrix $\hat{\boldsymbol{\Gamma}}_{a(r,k),a(r)^c} \in \mathbb{R}^{|a(r,k)|\times|a(r)^c|}$ whose columns are $\widehat{\boldsymbol{\gamma}}_{j,a(r,k)}$ with $j \in a(r)^c$. In addition, we define $\boldsymbol{\gamma}_j = \arg\min_{\boldsymbol{\gamma}\in\mathbb{R}^{p-1}} E(X_{ij} - \boldsymbol{\gamma}^\top \boldsymbol{X}_{i,-j})^2$, and similarly define $\boldsymbol{\Gamma}_{a(r,k),a(r)^c} \in \mathbb{R}^{|a(r,k)|\times|a(r)^c|}$ as a matrix whose columns are $\boldsymbol{\gamma}_{j,a(r,k)}$ with $j \in a(r)^c$. As a consequence, by (A1) - (A4) and the standard estimation bound for the Lasso estimator (Bickel et al., 2009; Negahban et al., 2010), we have

$$\max_{j\in a(r)^c} \|\widehat{\boldsymbol{\gamma}}_{j,a(r,k)} - \boldsymbol{\gamma}_{j,a(r,k)}\|_2 \lesssim \sqrt{\frac{s\log p}{N+n}}, \tag{S5.3}$$

with probability at least $1-p^{-c}$. In addition, for simplicity, we may also write $\boldsymbol{\Gamma}_{a(r,k),a(r)^c}$ as $\boldsymbol{\Gamma}_{r,k}$ and write $\hat{\boldsymbol{\Gamma}}_{a(r,k),a(r)^c}$ as $\hat{\boldsymbol{\Gamma}}_{r,k}$.

Part I. Feasibility. We start with the following lemmas.

Lemma 1. *Under conditions (A1) to (A4), with probability at least* $1-p^{-c}$*, it holds that*

$$\|\boldsymbol{g}(\boldsymbol{\beta})\|_\infty \lesssim \sqrt{\frac{\log p}{n}}. \tag{S5.4}$$

Lemma 2. *Under conditions (A1) to (A4), if* $s \lesssim \frac{N+n}{\log p}$*, then with probability at least* $1-p^{-c}$*, it holds that*

$$\|\boldsymbol{g}_n(\boldsymbol{\beta}) - \boldsymbol{g}(\boldsymbol{\beta})\|_\infty \lesssim s\sqrt{\frac{\log p}{N+n}}\left(1 + s\sqrt{\frac{\log p}{n}}\right) \tag{S5.5}$$

Combining the above two lemmas, we have

$$\|\boldsymbol{g}_n(\boldsymbol{\beta})\|_\infty \le \|\boldsymbol{g}_n(\boldsymbol{\beta}) - \boldsymbol{g}(\boldsymbol{\beta})\|_\infty + \|\boldsymbol{g}(\boldsymbol{\beta})\|_\infty \lesssim \sqrt{\frac{\log p}{n}} + s\sqrt{\frac{\log p}{N+n}}\left(1 + s\sqrt{\frac{\log p}{n}}\right) \tag{S5.6}$$

with probability at least $1-p^{-c}$. Whenever $s \ll \sqrt{n/\log p}$, the RHS of (S5.6) can be bounded by $\lambda \asymp \sqrt{\log p/n} + s\sqrt{\log p/(n+N)}$, which shows that the constraint $\|\boldsymbol{g}_n(\boldsymbol{\beta})\|_\infty \le \lambda$ is feasible with high probability.

Part II. Rate of Convergence. Firstly, if we denote $\boldsymbol{g}_n(\boldsymbol{\beta}) = \frac{1}{n}\sum_{i=1}^n \hat{\mathbf{h}}_i(\boldsymbol{\beta})$, where $\hat{\mathbf{h}}_i(\boldsymbol{\beta})$ has its components $\frac{n}{n_r} I\{\xi_i = r\}[y_i - (\widehat{\boldsymbol{X}}_i^{(k)})^\top \boldsymbol{\beta}]\widehat{\boldsymbol{X}}_{ia(k)}^{(k)}$, it then follows that $\boldsymbol{g}_n(\widehat{\boldsymbol{\beta}}) = \frac{1}{n}\sum_{i=1}^n \hat{\mathbf{h}}_i(\widehat{\boldsymbol{\beta}})$ where $\hat{\mathbf{h}}_i(\widehat{\boldsymbol{\beta}})$ has its components

$$\begin{aligned}
&\frac{n}{n_r} I\{\xi_i = r\}[y_i - (\widehat{\boldsymbol{X}}_i^{(k)})^\top \widehat{\boldsymbol{\beta}})]\widehat{\boldsymbol{X}}_{ia(k)}^{(k)} \\
&= \frac{n}{n_r} I\{\xi_i = r\}[\boldsymbol{X}_i^\top \boldsymbol{\beta} - (\widehat{\boldsymbol{X}}_i^{(k)})^\top \widehat{\boldsymbol{\beta}} + \epsilon_i]\widehat{\boldsymbol{X}}_{ia(k)}^{(k)} \\
&= \frac{n}{n_r} I\{\xi_i = r\}[(\widehat{\boldsymbol{X}}_i^{(k)})^\top (\boldsymbol{\beta} - \widehat{\boldsymbol{\beta}}) + (\boldsymbol{X}_i - \widehat{\boldsymbol{X}}_i^{(k)})^\top \boldsymbol{\beta} + \epsilon_i]\widehat{\boldsymbol{X}}_{ia(k)}^{(k)}. \qquad \text{(S5.7)}
\end{aligned}$$

Now, we will show that

$$\left\| \frac{1}{n}\sum_{i=1}^n \frac{n}{n_r} I\{\xi_i = r\}\widehat{\boldsymbol{X}}_{ia(k)}^{(k)} (\widehat{\boldsymbol{X}}_i^{(k)})^\top (\boldsymbol{\beta} - \widehat{\boldsymbol{\beta}}) \right\|_\infty \lesssim \lambda. \qquad \text{(S5.8)}$$

To see this, by the definition of $\widehat{\boldsymbol{\beta}}$ and Part I, we have

$$\left\| \frac{1}{n}\sum_{i=1}^n \hat{\mathbf{h}}_{irk}(\widehat{\boldsymbol{\beta}}) \right\|_\infty \leq \|\boldsymbol{g}_n(\widehat{\boldsymbol{\beta}})\|_\infty \lesssim \lambda, \qquad \text{(S5.9)}$$

with probability at least $1 - p^{-c}$. Then, in light of the decomposition (S5.7), it suffices to show that, with probability at least $1 - p^{-c}$,

$$\left\| \frac{1}{n}\sum_{i=1}^n I\{\xi_i = r\}\epsilon_i \widehat{\boldsymbol{X}}_{ia(k)}^{(k)} \right\|_\infty \leq \lambda, \qquad \text{(S5.10)}$$

and

$$\left\| \frac{1}{n}\sum_{i=1}^n I\{\xi_i = r\}(\boldsymbol{X}_i - \widehat{\boldsymbol{X}}_i^{(k)})^\top \boldsymbol{\beta} \widehat{\boldsymbol{X}}_{ia(k)}^{(k)} \right\|_\infty \leq \lambda. \qquad \text{(S5.11)}$$

This is done by the following lemma. Thus (S5.8) holds.

Lemma 3. *Under the conditions of Theorem 1, (S5.10) and (S5.11) hold with probability at least* $1 - p^{-c}$.

By (S5.8), we have

$$\left|\frac{1}{n}\sum_{i=1}^{n}\frac{n}{n_r}I\{\xi_i=r\}[\boldsymbol{\beta}-\widehat{\boldsymbol{\beta}}]_{a(k)}^{\top}\widehat{\boldsymbol{X}}_{ia(k)}^{(k)}(\widehat{\boldsymbol{X}}_{i}^{(k)})^{\top}(\boldsymbol{\beta}-\widehat{\boldsymbol{\beta}})\right| \lesssim \lambda\|[\boldsymbol{\beta}-\widehat{\boldsymbol{\beta}}]_{a(k)}\|_1 \leq \lambda s^{1/2}\|[\boldsymbol{\beta}-\widehat{\boldsymbol{\beta}}]_{a(k)}\|_2. \tag{S5.12}$$

Next, we would like to prove the lower bound

$$\left|\frac{1}{n}\sum_{i=1}^{n}\frac{n}{n_r}I\{\xi_i=r\}[\boldsymbol{\beta}-\widehat{\boldsymbol{\beta}}]_{a(k)}^{\top}\widehat{\boldsymbol{X}}_{ia(k)}^{(k)}(\widehat{\boldsymbol{X}}_{i}^{(k)})^{\top}(\boldsymbol{\beta}-\widehat{\boldsymbol{\beta}})\right| \gtrsim \|[\boldsymbol{\beta}-\widehat{\boldsymbol{\beta}}]_{a(k)}\|_2\|\boldsymbol{\beta}-\widehat{\boldsymbol{\beta}}\|_2, \tag{S5.13}$$

which along with the upper bound (S5.12) implies the final results

$$\|\boldsymbol{\beta}-\widehat{\boldsymbol{\beta}}\|_2 \leq s^{1/2}\lambda, \qquad \text{and} \qquad \|\boldsymbol{\beta}-\widehat{\boldsymbol{\beta}}\|_1 \leq s\lambda.$$

To show (S5.13), we need the following key proposition concerning a restricted singular value condition.

Proposition 1. *Under the conditions of Theorem 1, it holds that*

$$\inf_{\substack{\|\mathbf{u}\|_2=1,\mathbf{u}\in E_s\\ \|\mathbf{u}_{a(k)}\|_2\geq 1/2}}\left|\frac{\mathbf{u}_{a(k)}^{\top}}{\|\mathbf{u}_{a(k)}\|_2}\frac{1}{n}\sum_{i=1}^{n}\frac{n}{n_r}I\{\xi_i=r\}\widehat{\boldsymbol{X}}_{ia(k)}^{(k)}(\widehat{\boldsymbol{X}}_{i}^{(k)})^{\top}\mathbf{u}\right| \geq c_0 \tag{S5.14}$$

with probability at least $1-p^{-c}$*, where*

$$E_s(p)=\{\boldsymbol{\delta}\in\mathbb{R}^p : \|\boldsymbol{\delta}\|_2=1, \|\boldsymbol{\delta}_{S^c}\|_1 \leq \|\boldsymbol{\delta}_S\|_1, |S|\leq s\}.$$

Now by the definition of $\widehat{\boldsymbol{\beta}}$, it follows from the same argument of Candes and Tao (2007) that

$$\|[\widehat{\boldsymbol{\beta}}-\boldsymbol{\beta}]_{S^c}\|_1 \leq \|[\hat{\boldsymbol{\beta}}-\boldsymbol{\beta}]_S\|_1, \tag{S5.15}$$

where $S=\mathrm{supp}(\boldsymbol{\beta})$. If in addition

$$\|[\boldsymbol{\beta}-\widehat{\boldsymbol{\beta}}]_{a(k)}\|_2 \geq \|[\boldsymbol{\beta}-\widehat{\boldsymbol{\beta}}]_{a(k)^c}\|_2, \tag{S5.16}$$

we have $\frac{\|[\boldsymbol{\beta}-\widehat{\boldsymbol{\beta}}]_{a(k)}\|_2}{\|\boldsymbol{\beta}-\widehat{\boldsymbol{\beta}}\|_2} \geq 1/2$ so that by (S5.14), the lower bound (S5.13) hold. The rest of the proof is devoted to (S5.16).

Proof of (S5.16). The existence of r and $k \in \mathcal{G}(r)$ such that (S5.16) holds follows directly from the assumption that, there exists an $r \in \{1, ..., R\}$ and $k, k' \in \mathcal{G}(r)$ such that $a(k)^c \subset a(k')$ and $a(k')^c \subset a(k)$. In this case, if

$$\|[\boldsymbol{\beta} - \widehat{\boldsymbol{\beta}}]_{a(k)}\|_2 < \|[\boldsymbol{\beta} - \widehat{\boldsymbol{\beta}}]_{a(k)^c}\|_2,$$

we have

$$\|[\boldsymbol{\beta} - \widehat{\boldsymbol{\beta}}]_{a(k')^c}\|_2 < \|[\boldsymbol{\beta} - \widehat{\boldsymbol{\beta}}]_{a(k')}\|_2.$$

Thus, there exists some $k \in \mathcal{G}(r)$ such that (S5.16) holds.

S6 Proof of Theorem 2

Without loss of generality, we assume there are $2n$ supervised samples in $\mathcal{D}_2$, split into two halves. We need the following lemmas.

Lemma 4. *Suppose the smallest singular value of* $\mathbb{E}\boldsymbol{G} \in \mathbb{R}^{M_g' \times p}$, *where* $M_g' = \sum_{r=1}^{R} \sum_{k \in \mathcal{G}(r)} |a(r,k)|$, *is bounded away from zero. Then there exists infinite numbers of* $\boldsymbol{v} \in \mathbb{R}^{M_g}$ *such that* $\mathbb{E}\boldsymbol{G}^\top \boldsymbol{v} = \boldsymbol{e}_j$ *and* $\|\boldsymbol{v}\|_2 \leq C$ *for some constant* $C > 0$.

Lemma 5. *Let* $\boldsymbol{v}_j^* \in \mathbb{R}^{M_g'}$ *be any vector such that* $\mathbb{E}\boldsymbol{G}^\top \boldsymbol{v}_j^* = \boldsymbol{e}_j$. *Suppose the smallest singular value of* $\mathbb{E}\boldsymbol{G}$ *is bounded away from zero, and* $R \asymp 1$. *Then with probability at least* $1 - p^{-c}$, *it holds that*

$$\|(\boldsymbol{v}_j^*)^\top \boldsymbol{G}_n - \boldsymbol{e}_j^\top\|_\infty \lesssim \lambda'.$$

Lemma 6. *Let* $\hat{S}_j(\boldsymbol{\beta}) = \hat{\boldsymbol{v}}_j^\top \boldsymbol{g}_n^*(\boldsymbol{\beta})$. *If* $s\lambda\lambda'\sqrt{n} = o(1)$, *it holds that*

$$\sqrt{n}\hat{S}_j(\hat{\boldsymbol{\beta}}_j^*) = \sqrt{n}\hat{S}_j(\boldsymbol{\beta}) + o_P(1). \tag{S6.17}$$

Lemma 7. *Conditional on* $\widehat{X} = \{\widehat{\boldsymbol{X}}_i^{(k)}\}_{i\in\mathcal{D}_2}$, *it holds that*

$$\frac{n\hat{S}_j(\boldsymbol{\beta}) - \mu_j}{s_j}\bigg|\widehat{X} \to_d N(0,1), \tag{S6.18}$$

where

$$\mu_j = \sum_{i=1}^{n}\sum_{k\in\mathcal{G}(r),1\le r\le R}\frac{n}{n_r}I\{\xi_i = r\}\boldsymbol{\beta}_{a(r)^c}^\top(\boldsymbol{\Gamma}_{r,k} - \hat{\boldsymbol{\Gamma}}_{r,k})^\top\boldsymbol{X}_{ia(r,k)}\hat{\boldsymbol{v}}_{rk}^\top\boldsymbol{X}_{ia(r,k)}$$

and

$$s_j^2 = \sum_{i=1}^{n}\sum_{k\in\mathcal{G}(r),1\le r\le R}\frac{n^2\sigma_{r,k}^2}{n_r^2}I\{\xi_i = r\}[\hat{\boldsymbol{v}}_{rk}^\top\boldsymbol{X}_{ia(r,k)}]^2,$$

where $\sigma_{r,k}^2 = \sigma^2 + \boldsymbol{\beta}_{a(r)^c}^\top\mathbb{E}[\boldsymbol{\epsilon}_{ia(r)^c}^{(k)}(\boldsymbol{\epsilon}_{ia(r)^c}^{(k)})^\top]\boldsymbol{\beta}_{a(r)^c}$, *and* $\boldsymbol{\epsilon}_{ia(r)^c}^{(k)} = \boldsymbol{X}_{ia(r)^c} - \boldsymbol{\Gamma}_{r,k}^\top\boldsymbol{X}_{ia(r,k)} \in \mathbb{R}^{|a(r)^c|}$ *is the residual term of the* i*-th sample in the regression model of* $\boldsymbol{X}_{ia(r)^c}$ *with* $\boldsymbol{X}_{ia(r,k)}$ *as covariates.*

Let $\tilde{\boldsymbol{\beta}}_j^*$ be such that $\tilde{\boldsymbol{\beta}}_j^*$ is $\hat{\boldsymbol{\beta}}_j^*$ by replacing its j-th component by $\tilde{\beta}_j$. By the mean value theorem, we have

$$0 = \hat{S}_j(\tilde{\boldsymbol{\beta}}_j^*) = \hat{S}_j(\hat{\boldsymbol{\beta}}_j^*) + \hat{\boldsymbol{v}}_j^\top[\boldsymbol{G}_n]_{\cdot j}(\tilde{\beta}_j - \beta_j). \tag{S6.19}$$

By Lemma 6, we have

$$0 = \sqrt{n}\hat{S}_j(\boldsymbol{\beta}) + \sqrt{n}\hat{\boldsymbol{v}}_j^\top[\boldsymbol{G}_n]_{\cdot j}(\tilde{\beta}_j - \beta_j) + o_P(1),$$

or

$$\frac{\sqrt{n}(\tilde{\beta}_j - \beta_j)}{s_j/\sqrt{n}} = -\frac{n\hat{S}_j(\boldsymbol{\beta})}{s_j\hat{\boldsymbol{v}}_j^\top[\boldsymbol{G}_n]_{\cdot j}} + o_P(1) = -\frac{n\hat{S}_j(\boldsymbol{\beta}) - \mu_j}{s_j\hat{\boldsymbol{v}}_j^\top[\boldsymbol{G}_n]_{\cdot j}} - \frac{\mu_j}{s_j\hat{\boldsymbol{v}}_j^\top[\boldsymbol{G}_n]_{\cdot j}} + o_P(1).$$

In the above equation, by the definition of $\hat{\boldsymbol{v}}_j$ and Lemma 5, it holds that, with probability at least $1 - p^{-c}$

$$|\hat{\boldsymbol{v}}_j^\top [\boldsymbol{G}_n(\hat{\boldsymbol{\beta}}_j^*)]_{.j} - 1| \lesssim \sqrt{\frac{s \log p}{n}}.$$

or

$$\hat{\boldsymbol{v}}_j^\top [\boldsymbol{G}_n(\hat{\boldsymbol{\beta}}_j^*)]_{.j} \to_P 1.$$

Hence, it suffices to show

$$\mu_j / \sqrt{n} \to_P 0 \tag{S6.20}$$

and

$$s_j^2 / n \geq c > 0 \tag{S6.21}$$

with probability at least $1 - p^{-c}$. With these, it follows that

$$\frac{n(\tilde{\beta}_j - \beta_j)}{s_j} = AB + D,$$

where $A \to 1$, $D \to 0$ in probability and $B|\widehat{X} \to_d N(0,1)$. The rest of the proof is devoted to (S6.20) and (S6.21).

Proof of (S6.20) and (S6.21). The proof of (S6.20) can be established as soon as we prove

$$\left| \sum_{k \in \mathcal{G}(r), 1 \leq r \leq R} \frac{n}{n_r} I\{\xi_i = r\} \boldsymbol{\beta}_{a(r)^c}^\top (\boldsymbol{\Gamma}_{r,k} - \hat{\boldsymbol{\Gamma}}_{r,k})^\top \boldsymbol{X}_{ia(r,k)} \hat{\boldsymbol{v}}_{rk}^\top \boldsymbol{X}_{ia(r,k)} \right| = o_P(n^{-1/2}).$$

Now since

$$\begin{aligned}
&\left| \sum_{k \in \mathcal{G}(r), 1 \leq r \leq R} \frac{1}{n} \sum_{i=1}^n \frac{n}{n_r} I\{\xi_i = r\} \boldsymbol{\beta}_{a(r)^c}^\top (\hat{\boldsymbol{\Gamma}}_{r,k} - \boldsymbol{\Gamma}_{r,k})^\top \boldsymbol{X}_{ia(r,k)} (\boldsymbol{X}_{ia(r,k)})^\top \hat{\boldsymbol{v}}_{j,rk} \right| \\
&\lesssim \sum_{k \in \mathcal{G}(r), 1 \leq r \leq R} \sqrt{\frac{1}{n} \sum_{i=1}^n [\boldsymbol{\beta}_{a(r)^c}^\top (\hat{\boldsymbol{\Gamma}}_{r,k} - \boldsymbol{\Gamma}_{r,k})^\top \boldsymbol{X}_{ia(r,k)}]^2} \sqrt{\frac{1}{n} \sum_{i=1}^n \frac{n^2}{n_r^2} I\{\xi_i = r\} [(\boldsymbol{X}_{ia(r,k)})^\top \hat{\boldsymbol{v}}_{j,rk}]^2}.
\end{aligned} \tag{S6.22}$$

By (S7.28) we have

$$\max_{k,r} \|(\hat{\mathbf{\Gamma}}_{r,k} - \mathbf{\Gamma}_{r,k})\boldsymbol{\beta}_{a(r)^c}\|_2 \leq \|\boldsymbol{\beta}_{a(r)^c}\|_1 \cdot \max_{1\leq j\leq a(r)^c} \|\widehat{\boldsymbol{\gamma}}_{j,a(r,k)} - \boldsymbol{\gamma}_{j,a(r,k)}\|_2 \lesssim s\sqrt{\frac{\log p}{N+n}} \quad \text{(S6.23)}$$

with probability at least $1 - p^{-c}$. By the definition of $\hat{\boldsymbol{v}}_j$, we have

$$\sqrt{\frac{1}{n}\sum_{i=1}^{n} \frac{n^2}{n_r^2} I\{\xi_i = r\}[(\boldsymbol{X}_{ia(r,k)})^\top \hat{\boldsymbol{v}}_{j,rk}]^2} \leq \sqrt{\frac{1}{n}\sum_{i=1}^{n} \frac{n^2}{n_r^2} I\{\xi_i = r\}[(\boldsymbol{X}_{ia(r,k)})^\top \boldsymbol{v}^*_{j,rk}]^2}, \quad \text{(S6.24)}$$

where by concentration inequality of sub-exponential random variables, the right hand side of the above inequality is bounded by a constant with probability at least $1 - p^{-c}$. Thus, we have

$$\left| \sum_{k\in\mathcal{G}(r),1\leq r\leq R} \frac{n}{n_r} I\{\xi_i = r\} \boldsymbol{\beta}^\top_{a(r)^c} (\mathbf{\Gamma}_{r,k} - \hat{\mathbf{\Gamma}}_{r,k})^\top \boldsymbol{X}_{ia(r,k)} \hat{\boldsymbol{v}}^\top_{rk} \boldsymbol{X}_{ia(r,k)} \right| \lesssim s\sqrt{\frac{\log p}{N+n}} = o(n^{-1/2})$$

with high probability under the condition $s \ll (\frac{N}{n\log p})^{1/2}$ and $N \gg n$. Now to show (S6.21), since $\sigma^2_{r,k} \geq \sigma^2$, it suffices to show

$$\frac{\sigma^2}{n}\sum_{i=1}^{n} \sum_{k\in\mathcal{G}(r),1\leq r\leq R} \frac{n^2}{n_r^2} I\{\xi_i = r\}[\hat{\boldsymbol{v}}^\top_{j,rk} \boldsymbol{X}_{ia(r,k)}]^2 \geq c > 0. \quad \text{(S6.25)}$$

To see this, note that by definition of $\hat{\boldsymbol{v}}_j$, we have

$$\left| \sum_{k\in\mathcal{G}(r),1\leq r\leq R} \frac{1}{n}\sum_{i=1}^{n} \frac{n}{n_r} I\{\xi_i = r\} \widehat{X}^{(k)}_{ij} (\boldsymbol{X}_{ia(r,k)})^\top \hat{\boldsymbol{v}}_{j,rk} - 1 \right| \leq \lambda',$$

or

$$\sum_{k\in\mathcal{G}(r),1\leq r\leq R} \frac{1}{n}\sum_{i=1}^{n} \frac{n}{n_r} I\{\xi_i = r\} \widehat{X}^{(k)}_{ij} (\boldsymbol{X}_{ia(r,k)})^\top \hat{\boldsymbol{v}}_{j,rk} \geq 1 - \lambda'.$$

By Cauchy-Schwartz inequality, we have

$$\sum_{k\in\mathcal{G}(r),1\leq r\leq R} \sqrt{\frac{1}{n}\sum_{i=1}^{n} [\widehat{X}^{(k)}_{ij}]^2} \sqrt{\frac{1}{n}\sum_{i=1}^{n} \frac{n^2}{n_r^2} I\{\xi_i = r\}[(\boldsymbol{X}_{ia(r,k)})^\top \hat{\boldsymbol{v}}_{j,rk}]^2} \geq 1 - \lambda',$$

or

$$\sqrt{\sum_{k\in\mathcal{G}(r),1\leq r\leq R} \frac{1}{n}\sum_{i=1}^{n} \frac{n^2}{n_r^2} I\{\xi_i = r\}[(\boldsymbol{X}_{ia(r,k)})^\top \hat{\boldsymbol{v}}_{j,rk}]^2} \gtrsim \frac{1-\lambda'}{\max_{k\in\mathcal{G}(r),1\leq r\leq R} \sqrt{\frac{1}{n}\sum_{i=1}^{n} [\widehat{X}^{(k)}_{ij}]^2}}.$$

Now since by concentration inequality of subexponential random variables with probability at least $1 - p^{-c}$,

$$\max_{k \in \mathcal{G}(r), 1 \le r \le R} \sqrt{\frac{1}{n} \sum_{i=1}^{n} [\widehat{X}_{ij}^{(k)}]^2} \ge c > 0,$$

it follows that under the same event (S6.25) holds. □

S7 Proof of Auxiliary Lemmas

S7.1 Proof of Lemma 1

By definition, $g(\boldsymbol{\beta})$ has its subvectors of the form

$$\begin{aligned} \frac{1}{n_r} \sum_{i=1}^{n} I\{\xi_i = r\}[y_i - (\boldsymbol{X}_i^{(k)})^\top \boldsymbol{\beta})]\boldsymbol{X}_{ia(k)}^{(k)} &= \frac{1}{n_r} \sum_{i=1}^{n} I\{\xi_i = r\}[\boldsymbol{X}_i^\top \boldsymbol{\beta} + \epsilon_i - (\boldsymbol{X}_i^{(k)})^\top \boldsymbol{\beta})]\boldsymbol{X}_{ia(k)}^{(k)} \\ &= \frac{1}{n_r} \sum_{i=1}^{n} I\{\xi_i = r\}[(\boldsymbol{X}_i - \boldsymbol{X}_i^{(k)})^\top \boldsymbol{\beta} + \epsilon_i]\boldsymbol{X}_{ia(k)}^{(k)}. \end{aligned}$$

By assumptions (A4) and $\|\boldsymbol{\beta}\|_2 \le C$, each of $(\boldsymbol{X}_i - \boldsymbol{X}_i^{(k)})^\top \boldsymbol{\beta}$, ϵ_i and $\boldsymbol{X}_{ia(k)}^{(k)}$ are sub-Gaussian random variables and are mutually independent across different samples $i \in \{1, ..., n\}$. Since the product of two sub-Gaussian random variables are sub-exponential, and by the properties of conditional expectations

$$\mathbb{E}[(\boldsymbol{X}_i - \boldsymbol{X}_i^{(k)})^\top \boldsymbol{\beta} + \epsilon_i]\boldsymbol{X}_{ia(k)}^{(k)} = 0, \qquad i = 1, ..., n,$$

then it follows that, as long as $0 < c_1 < n/n_r < c_2 < 1$ for each $1 \le r \le R$ (A2), each component of $\boldsymbol{g}(\boldsymbol{\beta})$ is a centred sub-exponential variable. Applying the concentration inequality for independent sub-exponential random variables (see, for example, Proposition 5.16 of Vershynin (2010)), we have

$$P\left(\|\boldsymbol{g}(\boldsymbol{\beta})\|_\infty \le C\sqrt{\frac{\log M_g}{n}} \right) \ge 1 - M_g^{-c}$$

for some constants $C, c > 0$. The final results follows from (A2) so that $p \lesssim M_g \leq R^2 p \lesssim p$. □

S7.2 Proof of Lemma 3

To prove (S5.10), we note that $\widehat{\boldsymbol{X}}^{(k)}_{ia(k)}$ can be partitioned into two subvectors, namely, $X_{ia(r,k)}$ and $\widehat{\boldsymbol{X}}_{ia(r)^c}$. On the one hand, by the concentration inequality for independent centred sub-exponential random random variables, we have

$$P\left(\left\|\frac{1}{n}\sum_{i=1}^{n} I\{\xi_i = r\}\epsilon_i \boldsymbol{X}_{ia(r,k)}\right\|_\infty \leq C\sqrt{\frac{\log p}{n}}\right) \geq 1 - p^{-c}. \tag{S7.26}$$

On the other hand, for any $j \in a(r)^c$,

$$[\widehat{\boldsymbol{X}}_{ia(r)^c}\epsilon_i]_j = \widehat{\boldsymbol{\gamma}}^\top_{j,a(r,k)}\boldsymbol{X}_{ia(r,k)}\epsilon_i = (\widehat{\boldsymbol{\gamma}}_{j,a(r,k)} - \boldsymbol{\gamma}_{j,a(r,k)})^\top \boldsymbol{X}_{ia(r,k)}\epsilon_i + \boldsymbol{\gamma}^\top_{j,a(r,k)}\boldsymbol{X}_{ia(r,k)}\epsilon_i.$$

By the concentration inequality for independent centred sub-exponential random random variables, we have

$$P\left(\max_{j\in a(r)^c}\left|\frac{1}{n}\sum_{i=1}^{n} I\{\xi_i = r\}\boldsymbol{\gamma}^\top_{j,a(r,k)}\boldsymbol{X}_{ia(r,k)}\epsilon_i\right| \leq C\sqrt{\frac{\log p}{n}}\right) \geq 1 - p^{-c}.$$

In addition, by (S5.3), with probability at least $1 - p^{-c}$, we have

$$\begin{aligned}
&\max_{j\in a(r)^c}\left|\frac{1}{n}\sum_{i=1}^{n} I\{\xi_i = r\}(\widehat{\boldsymbol{\gamma}}_{j,a(r,k)} - \boldsymbol{\gamma}_{j,a(r,k)})^\top \boldsymbol{X}_{ia(r,k)}\epsilon_i\right|_\infty \\
&\leq \max_{j\in a(r)^c}\|\widehat{\boldsymbol{\gamma}}_{j,a(r,k)} - \boldsymbol{\gamma}_{j,a(r,k)}\|_1\left\|\frac{1}{n}\sum_{i=1}^{n} I\{\xi_i = r\}\boldsymbol{X}_{ia(r,k)}\epsilon_i\right\|_\infty \\
&\lesssim s\sqrt{\frac{\log p}{N+n}}\sqrt{\frac{\log p}{n}}.
\end{aligned}$$

Therefore,

$$P\left(\max_{j\in a(r)^c}\left|\frac{1}{n}\sum_{i=1}^{n} I\{\xi_i = r\}\widehat{\boldsymbol{\gamma}}^\top_{j,a(r,k)}\boldsymbol{X}_{ia(r,k)}\epsilon_i\right| \leq C\left(1 + s\sqrt{\frac{\log p}{N+n}}\right)\sqrt{\frac{\log p}{n}}\right) \geq 1 - p^{-c}. \tag{S7.27}$$

Inequality (S5.10) then follows from (S7.26) and (S7.27).

To prove (S5.11), note that

$$(\boldsymbol{X}_i - \widehat{\boldsymbol{X}}_i^{(k)})^\top \boldsymbol{\beta} \widehat{\boldsymbol{X}}_{ia(k)}^{(k)} = [\boldsymbol{X}_{ia(r,k)}^\top (\boldsymbol{\Gamma}_{r,k} - \widehat{\boldsymbol{\Gamma}}_{r,k}) + \boldsymbol{\epsilon}_{ia(r)^c}^{(k)}] \boldsymbol{\beta}_{a(r)^c} \widehat{\boldsymbol{X}}_{ia(k)}^{(k)}.$$

Again, the above vector consists of two parts, corresponding to the above partition of $\widehat{\boldsymbol{X}}_{ia(k)}^{(k)}$. For the first part, it follows that

$$\begin{aligned}
&\left\| \frac{1}{n} \sum_{i=1}^n I\{\xi_i = r\} [\boldsymbol{X}_{ia(r,k)}^\top (\boldsymbol{\Gamma}_{r,k} - \widehat{\boldsymbol{\Gamma}}_{r,k}) + \boldsymbol{\epsilon}_{ia(r)^c}^{(k)}] \boldsymbol{\beta}_{a(r)^c} \boldsymbol{X}_{ia(r,k)} \right\|_\infty \\
&= \max_{k \in a(r,k)} \left| \frac{1}{n} \sum_{i=1}^n I\{\xi_i = r\} \boldsymbol{\beta}_{a(r)^c}^\top [\boldsymbol{X}_{ia(r,k)}^\top (\boldsymbol{\Gamma}_{r,k} - \widehat{\boldsymbol{\Gamma}}_{r,k}) + \boldsymbol{\epsilon}_{ia(r)^c}^{(k)}] X_{ik} \right| \\
&\le \|\boldsymbol{\beta}_{a(r)^c}^\top (\widehat{\boldsymbol{\Gamma}}_{r,k} - \boldsymbol{\Gamma}_{r,k})^\top\|_2 \cdot \max_{k \in a(r,k)} \left| \frac{1}{n} \sum_{i=1}^n I\{\xi_i = r\} \mathbf{u}_{r,k}^\top \boldsymbol{X}_{ia(r,k)} X_{ik} \right| \\
&\quad + \max_{k \in a(r,k)} \left| \frac{1}{n} \sum_{i=1}^n I\{\xi_i = r\} \boldsymbol{\beta}_{a(r)^c}^\top \boldsymbol{\epsilon}_{ia(r)^c}^{(k)} X_{ik} \right|,
\end{aligned}$$

where $\mathbf{u}_{r,k}^\top = \boldsymbol{\beta}_{a(r)^c}^\top (\widehat{\boldsymbol{\Gamma}}_{r,k} - \boldsymbol{\Gamma}_{r,k})^\top / \|\boldsymbol{\beta}_{a(r)^c}^\top (\widehat{\boldsymbol{\Gamma}}_{r,k} - \boldsymbol{\Gamma}_{r,k})^\top\|_2$. Note that $\|\boldsymbol{\beta}_{a(r)^c}\|_1 \le \sqrt{s}\|\boldsymbol{\beta}\|_2 \lesssim \sqrt{s}$. Then we have

$$\begin{aligned}
\|(\widehat{\boldsymbol{\Gamma}}_{r,k} - \boldsymbol{\Gamma}_{r,k}) \boldsymbol{\beta}_{a(r)^c}\|_2 &\le \sum_{j \in a(r)^c} |\beta_j| \cdot \|\widehat{\boldsymbol{\gamma}}_{j,a(r,k)} - \boldsymbol{\gamma}_{j,a(r,k)}\|_2 \\
&\le \|\boldsymbol{\beta}_{a(r)^c}\|_1 \cdot \max_{1 \le j \le a(r)^c} \|\widehat{\boldsymbol{\gamma}}_{j,a(r,k)} - \boldsymbol{\gamma}_{j,a(r,k)}\|_2 \lesssim s \sqrt{\frac{\log p}{N+n}}, \qquad \text{(S7.28)}
\end{aligned}$$

and, by the concentration inequality of sub-exponential random variables and the independence between $\widehat{\boldsymbol{\Gamma}}_{r,k}$ and $\boldsymbol{X}_i$,

$$\max_{k \in a(r,k)} \left| \frac{1}{n} \sum_{i=1}^n I\{\xi_i = r\} \mathbf{u}_{r,k}^\top \boldsymbol{X}_{ia(r,k)} X_{ik} \right| \le C, \qquad \text{(S7.29)}$$

with probability at least $1 - p^{-c}$. Moreover, by the fact that $\mathbb{E}[\boldsymbol{\beta}_{a(r)^c}^\top \boldsymbol{\epsilon}_{ia(r)^c}^{(k)} X_{ik}] = 0$ and the concentration inequality of sub-exponential random variables, we also have

$$\max_{k \in a(r,k)} \left| \frac{1}{n} \sum_{i=1}^n I\{\xi_i = r\} \boldsymbol{\beta}_{a(r)^c}^\top \boldsymbol{\epsilon}_{ia(r)^c}^{(k)} X_{ik} \right| \lesssim \sqrt{\frac{\log p}{n}},$$

with probability at least $1-p^{-c}$.

For the second part, it follows that

$$\begin{aligned}
&\left\|\frac{1}{n}\sum_{i=1}^{n} I\{\xi_i=r\}[\boldsymbol{X}_{ia(r,k)}^{\top}(\boldsymbol{\Gamma}_{r,k}-\widehat{\boldsymbol{\Gamma}}_{r,k})+\boldsymbol{\epsilon}_{ia(r)^c}^{(k)}]\boldsymbol{\beta}_{a(r)^c}\boldsymbol{X}_{ia(r,k)}^{\top}\widehat{\boldsymbol{\Gamma}}_{r,k}\right\|_{\infty}\\
&=\max_{k\in a(r)^c}\left|\frac{1}{n}\sum_{i=1}^{n} I\{\xi_i=r\}\boldsymbol{\beta}_{a(r)^c}^{\top}[(\boldsymbol{\Gamma}_{r,k}-\widehat{\boldsymbol{\Gamma}}_{r,k})^{\top}\boldsymbol{X}_{ia(r,k)}+\boldsymbol{\epsilon}_{ia(r)^c}^{(k)}]\boldsymbol{X}_{ia(r,k)}^{\top}\widehat{\boldsymbol{\gamma}}_{j,a(r,k)}\right|\\
&\le\|\boldsymbol{\beta}_{a(r)^c}^{\top}(\widehat{\boldsymbol{\Gamma}}_{r,k}-\boldsymbol{\Gamma}_{r,k})^{\top}\|_2\cdot\max_{k\in a(r)^c}\left|\frac{1}{n}\sum_{i=1}^{n} I\{\xi_i=r\}\mathbf{u}_{r,k}^{\top}\boldsymbol{X}_{ia(r,k)}\boldsymbol{X}_{ia(r,k)}^{\top}\boldsymbol{\gamma}_{j,a(r,k)}\right|\\
&\quad+\|\boldsymbol{\beta}_{a(r)^c}^{\top}(\widehat{\boldsymbol{\Gamma}}_{r,k}-\boldsymbol{\Gamma}_{r,k})^{\top}\|_2\cdot\max_{k\in a(r)^c}\left|\frac{1}{n}\sum_{i=1}^{n} I\{\xi_i=r\}\mathbf{u}_{r,k}^{\top}\boldsymbol{X}_{ia(r,k)}\boldsymbol{X}_{ia(r,k)}^{\top}(\widehat{\boldsymbol{\gamma}}_{j,a(r,k)}-\boldsymbol{\gamma}_{j,a(r,k)})\right|\\
&\quad+\max_{k\in a(r)^c}\left|\frac{1}{n}\sum_{i=1}^{n} I\{\xi_i=r\}\boldsymbol{\beta}_{a(r)^c}^{\top}\boldsymbol{\epsilon}_{ia(r)^c}^{(k)}\boldsymbol{X}_{ia(r,k)}^{\top}\widehat{\boldsymbol{\gamma}}_{j,a(r,k)}\right|.
\end{aligned}$$

By the concentration inequality for sub-exponential random variables and the fact that $\sup_{j,r,k}\|[\boldsymbol{\Gamma}_{r,k}]_{\cdot j}\|_2\le C$, which is implied by the assumption that $\boldsymbol{\Sigma}$ has bounded eigenvalues from both above and below, we have with probability at least $1-p^{-c}$,

$$\max_{k\in a(r)^c}\left|\frac{1}{n}\sum_{i=1}^{n} I\{\xi_i=r\}\mathbf{u}_{r,k}^{\top}\boldsymbol{X}_{ia(r,k)}\boldsymbol{X}_{ia(r,k)}^{\top}\boldsymbol{\gamma}_{j,a(r,k)}\right|\le C. \tag{S7.30}$$

Similarly, since $\mathbb{E}\boldsymbol{\epsilon}_{ia(r)^c}^{(k)}\boldsymbol{X}_{ia(r,k)}^{\top}\widehat{\boldsymbol{\gamma}}_{j,a(r,k)}=\mathbb{E}[\mathbb{E}[\boldsymbol{\epsilon}_{ia(r)^c}^{(k)}\boldsymbol{X}_{ia(r,k)}^{\top}\widehat{\boldsymbol{\gamma}}_{j,a(r,k)}|\widehat{X}]]=0$, we also have

$$\max_{k\in a(r)^c}\left|\frac{1}{n}\sum_{i=1}^{n} I\{\xi_i=r\}\boldsymbol{\beta}_{a(r)^c}^{\top}\boldsymbol{\epsilon}_{ia(r)^c}^{(k)}\boldsymbol{X}_{ia(r,k)}^{\top}\widehat{\boldsymbol{\gamma}}_{j,a(r,k)}\right|\lesssim\sqrt{\frac{\log p}{n}},$$

with probability at least $1-p^{-c}$. In addition,

$$\begin{aligned}
&\max_{k\in a(r)^c}\left|\frac{1}{n}\sum_{i=1}^{n} I\{\xi_i=r\}\mathbf{u}_{r,k}^{\top}\boldsymbol{X}_{ia(r,k)}\boldsymbol{X}_{ia(r,k)}^{\top}(\widehat{\boldsymbol{\gamma}}_{j,a(r,k)}-\boldsymbol{\gamma}_{j,a(r,k)})\right|\\
&\le\max_{k\in a(r)^c}\|\widehat{\boldsymbol{\gamma}}_{j,a(r,k)}-\boldsymbol{\gamma}_{j,a(r,k)}\|_2\cdot\left\|\frac{1}{n}\sum_{i=1}^{n} I\{\xi_i=r\}\mathbf{u}_{r,k}^{\top}\boldsymbol{X}_{ia(r,k)}\boldsymbol{X}_{ia(r,k)}^{\top}\mathbf{v}_j\right\|_2\lesssim\sqrt{\frac{s\log p}{N+n}},
\end{aligned}$$

where in the second last inequality $\mathbf{v}_j=(\widehat{\boldsymbol{\gamma}}_{j,a(r,k)}-\boldsymbol{\gamma}_{j,a(r,k)})/\|\widehat{\boldsymbol{\gamma}}_{j,a(r,k)}-\boldsymbol{\gamma}_{j,a(r,k)}\|_2$ and the last inequality follows again from the concentration inequality for the sub-exponential random variables. The above inequalities imply (S5.11).

S7.3 Proof of Lemma 2

For each $r \in [1:R]$, $i \in \mathcal{S}(r)$ and $k \in \mathcal{G}(r)$, we have

$$\begin{aligned}\widehat{\boldsymbol{h}}_{irk}(\boldsymbol{\beta}) &= I(\xi_i = r)[\{\boldsymbol{X}_i - \widehat{\boldsymbol{X}}_i^{(k)}\}^\top \boldsymbol{\beta} + \epsilon_i] \cdot \widehat{\boldsymbol{X}}_{ia(k)}^{(k)} \\ &= I(\xi_i = r)\left[\{\boldsymbol{X}_{ia(r)^c} - \widehat{\boldsymbol{X}}_{ia(r)^c}^{(k)}\}^\top \boldsymbol{\beta}_{a(r)^c} + \epsilon_i\right] \cdot \widehat{\boldsymbol{X}}_{ia(k)}^{(k)},\end{aligned}$$

and therefore

$$\boldsymbol{g}(\boldsymbol{\beta}) - g_n(\boldsymbol{\beta}) = \frac{1}{n}\sum_{i=1}^{n}\{\mathbf{h}_i(\boldsymbol{\beta}) - \hat{\mathbf{h}}_i(\boldsymbol{\beta})\}, \tag{S7.31}$$

where $\hat{\mathbf{h}}_i(\boldsymbol{\beta}) = (\hat{\theta}_1^{-1}\widehat{\boldsymbol{h}}_{i1}(\boldsymbol{\beta})^\top, \ldots, \hat{\theta}_R^{-1}\widehat{\boldsymbol{h}}_{iR}(\boldsymbol{\beta})^\top)^\top$ and $\mathbf{h}_i(\boldsymbol{\beta}) = (\hat{\theta}_1^{-1}\boldsymbol{h}_{i1}(\boldsymbol{\beta})^\top, \ldots, \hat{\theta}_R^{-1}\boldsymbol{h}_{iR}(\boldsymbol{\beta})^\top)^\top$ which consist of all the (imputed) estimating functions for the i-th sample. In particular, for $k \in \mathcal{G}(r)$, we have

$$\begin{aligned}&\frac{1}{n}\sum_{i=1}^{n}\hat{\theta}_r^{-1}\left\{\boldsymbol{h}_{irk}(\boldsymbol{\beta}) - \hat{\boldsymbol{h}}_{irk}(\boldsymbol{\beta})\right\} \\ =\ &\frac{1}{n}\sum_{i=1}^{n}\hat{\theta}_r^{-1} I(\xi_i = r)\left([y_i - \{\boldsymbol{X}_i^{(k)}\}^\top\boldsymbol{\beta}]\cdot\boldsymbol{X}_{ia(k)}^{(k)} - [y_i - \{\widehat{\boldsymbol{X}}_i^{(k)}\}^\top\boldsymbol{\beta}]\widehat{\boldsymbol{X}}_{ia(k)}^{(k)}\right) \\ =\ &\frac{1}{n\hat{\theta}_r}\sum_{i=1}^{n} I(\xi_i = r)\left(\left[\{\boldsymbol{X}_{ia(r)^c} - \boldsymbol{X}_{ia(r)^c}^{(k)}\}^\top\boldsymbol{\beta}_{a(r)^c} + \epsilon_i\right]\boldsymbol{X}_{ia(k)}^{(k)}\right. \\ &\left. - \left[\{\boldsymbol{X}_{ia(r)^c} - \widehat{\boldsymbol{X}}_{ia(r)^c}^{(k)}\}^\top\boldsymbol{\beta}_{a(r)^c} + \epsilon_i\right]\widehat{\boldsymbol{X}}_{ia(k)}^{(k)}\right) \\ =\ &\frac{1}{n\hat{\theta}_r}\sum_{i=1}^{n} I(\xi_i = r)\left(\boldsymbol{X}_{ia(k)}^{(k)}\{\boldsymbol{X}_{ia(r)^c} - \boldsymbol{X}_{ia(r)^c}^{(k)}\}^\top\boldsymbol{\beta}_{a(r)^c} - \widehat{\boldsymbol{X}}_{ia(k)}^{(k)}\{\boldsymbol{X}_{ia(r)^c} - \widehat{\boldsymbol{X}}_{ia(r)^c}^{(k)}\}^\top\boldsymbol{\beta}_{a(r)^c}\right. \\ &\left. + \epsilon_i\left\{\boldsymbol{X}_{ia(k)}^{(k)} - \widehat{\boldsymbol{X}}_{ia(k)}^{(k)}\right\}\right).\end{aligned}$$

Then,

$$
\begin{aligned}
&\left\| \frac{1}{n}\sum_{i=1}^{n} \hat{\theta}_r^{-1} \left\{ \boldsymbol{h}_{irk}(\boldsymbol{\beta}) - \hat{\boldsymbol{h}}_{irk}(\boldsymbol{\beta}) \right\} \right\|_\infty \\
= \;& \hat{\theta}_r^{-1} \max_{1\le r\le R, k\in\mathcal{G}(r), j\in a(k)} \left| \frac{1}{n}\sum_{i=1}^{n} I(\xi_i = r) \left(X_{ij}^{(k)} \{ \boldsymbol{X}_{ia(r)^c} - \boldsymbol{X}_{ia(r)^c}^{(k)} \}^\top \boldsymbol{\beta}_{a(r)^c} \right.\right. \\
& \left.\left. - \widehat{X}_{ij}^{(k)} \{ \boldsymbol{X}_{ia(r)^c} - \widehat{\boldsymbol{X}}_{ia(r)^c}^{(k)} \}^\top \boldsymbol{\beta}_{a(r)^c} + \epsilon_i \cdot \left\{ X_{ij}^{(k)} - \widehat{X}_{ij}^{(k)} \right\} \right) \right| \\
= \;& \hat{\theta}_r^{-1} \max_{1\le r\le R, k\in\mathcal{G}(r), j\in a(k)} \left| \frac{1}{n}\sum_{i=1}^{n} I(\xi_i = r) \left(X_{ij}^{(k)} \boldsymbol{X}_{ia(r,k)}^\top \left\{ \hat{\boldsymbol{\Gamma}}_{a(r,k),a(r)^c} - \boldsymbol{\Gamma}_{a(r,k),a(r)^c} \right\} \boldsymbol{\beta}_{a(r)^c} \right.\right. \\
& + \left\{ \widehat{X}_{ij}^{(k)} - X_{ij}^{(k)} \right\} \left[\boldsymbol{X}_{ia(r,k)}^\top \left\{ \hat{\boldsymbol{\Gamma}}_{a(r,k),a(r)^c} - \boldsymbol{\Gamma}_{a(r,k),a(r)^c} \right\} - \left\{ \boldsymbol{\epsilon}_{ia(r)^c}^{(k)} \right\}^\top \right] \boldsymbol{\beta}_{a(r)^c} \\
& \left.\left. + \epsilon_i \cdot \left\{ X_{ij}^{(k)} - \widehat{X}_{ij}^{(k)} \right\} \right) \right|, \qquad \text{(S7.32)}
\end{aligned}
$$

where $\boldsymbol{\epsilon}_{ia(r)^c}^{(k)}$ is the residual term of the i-th sample in the regression model of $\boldsymbol{X}_{ia(r)^c}$ with $\boldsymbol{X}_{ia(r,k)}$ as covariates. In particular, $\boldsymbol{\epsilon}_{ia(r)^c}^{(k)}$ is centered and uncorrelated with $\boldsymbol{X}_{ia(r,k)}$.

If $j \in a(r,k)$, then

$$
\begin{aligned}
& \left| \frac{1}{n_r}\sum_{i=1}^{n} I(\xi_i = r) \left(X_{ij}^{(k)} \boldsymbol{X}_{ia(r,k)}^\top \left\{ \hat{\boldsymbol{\Gamma}}_{a(r,k),a(r)^c} - \boldsymbol{\Gamma}_{a(r,k),a(r)^c} \right\} \boldsymbol{\beta}_{a(r)^c} \right.\right. \\
& + \left\{ \hat{X}_{ij}^{(k)} - X_{ij}^{(k)} \right\} \left[\boldsymbol{X}_{ia(r,k)}^\top \left\{ \widehat{\boldsymbol{\Gamma}}_{a(r,k),a(r)^c} - \boldsymbol{\Gamma}_{a(r,k),a(r)^c} \right\} - \left\{ \boldsymbol{\epsilon}_{ia(r)^c}^{(k)} \right\}^\top \right] \boldsymbol{\beta}_{a(r)^c} \\
& \left.\left. + \epsilon_i \cdot \left\{ X_{ij}^{(k)} - \widehat{X}_{ij}^{(k)} \right\} \right) \right| \\
= \;& \left| \frac{1}{n_r}\sum_{i=1}^{n} I(\xi_i = r) X_{ij} \boldsymbol{X}_{ia(r,k)}^\top \mathbf{u}_{r,k} \right| \cdot \| (\hat{\boldsymbol{\Gamma}}_{a(r,k),a(r)^c} - \boldsymbol{\Gamma}_{a(r,k),a(r)^c}) \boldsymbol{\beta}_{a(r)^c} \|_2 \qquad \text{(S7.33)} \\
\lesssim \;& s\sqrt{\frac{\log p}{N+n}}, \qquad \text{(S7.34)}
\end{aligned}
$$

where the last inequality follows from (S7.28) and (S7.29).

If $j \in a(k) \setminus a(r,k)$, then

$$
\begin{aligned}
& \frac{1}{n_r}\sum_{i=1}^{n} I(\xi_i = r) \left(X_{ij}^{(k)} \mathbf{X}_{ia(r,k)}^\top \left\{ \hat{\boldsymbol{\Gamma}}_{a(r,k),a(r)^c} - \boldsymbol{\Gamma}_{a(r,k),a(r)^c} \right\} \boldsymbol{\beta}_{a(r)^c} \right. \\
& \left. + \left\{ \widehat{X}_{ij}^{(k)} - X_{ij}^{(k)} \right\} \left[\boldsymbol{X}_{ia(r,k)}^\top \left\{ \hat{\boldsymbol{\Gamma}}_{a(r,k),a(r)^c} - \boldsymbol{\Gamma}_{a(r,k),a(r)^c} \right\} - \left\{ \boldsymbol{\epsilon}_{ia(r)^c}^{(k)} \right\}^\top \right] \boldsymbol{\beta}_{a(r)^c} \right.
\end{aligned}
$$

$$
\begin{aligned}
&+\epsilon_i \cdot \left\{ X_{ij}^{(k)} - \widehat{X}_{ij}^{(k)} \right\} \Big) \\
= \; & \left\{ \frac{1}{n_r} \sum_{i=1}^{n} I(\xi_i = r) X_{ij}^{(k)} \mathbf{X}_{ia(r,k)}^{\top} \right\} \left\{ \hat{\mathbf{\Gamma}}_{a(r,k),a(r)^c} - \mathbf{\Gamma}_{a(r,k),a(r)^c} \right\} \boldsymbol{\beta}_{a(r)^c} \\
& + \left\{ \boldsymbol{\gamma}_{j,a(r,k)} - \widehat{\boldsymbol{\gamma}}_{j,a(r,k)} \right\}^{\top} \left\{ \frac{1}{n_r} \sum_{i=1}^{n} I(\xi_i = r) \boldsymbol{X}_{ia(r,k)} \boldsymbol{X}_{ia(r,k)}^{\top} \right\} \left\{ \hat{\mathbf{\Gamma}}_{a(r,k),a(r)^c} - \mathbf{\Gamma}_{a(r,k),a(r)^c} \right\} \boldsymbol{\beta}_{a(r)^c} \\
& + \left\{ \boldsymbol{\gamma}_{j,a(r,k)} - \widehat{\boldsymbol{\gamma}}_{j,a(r,k)} \right\}^{\top} \left\{ \frac{1}{n_r} \sum_{i=1}^{n} I(\xi_i = r) \boldsymbol{X}_{ia(r,k)} \left\{ \boldsymbol{\epsilon}_{ia(r)^c}^{(k)} \right\}^{\top} \right\} \boldsymbol{\beta}_{a(r)^c} \\
& + \left\{ \boldsymbol{\gamma}_{j,a(r,k)} - \widehat{\boldsymbol{\gamma}}_{j,a(r,k)} \right\}^{\top} \frac{1}{n_r} \sum_{i=1}^{n} I(\xi_i = r) \epsilon_i \boldsymbol{X}_{ia(r,k)} \\
= \; & \boldsymbol{\gamma}_{j,a(r,k)}^{\top} \left\{ \frac{1}{n_r} \sum_{i=1}^{n} I(\xi_i = r) \mathbf{X}_{ia(r,k)} \boldsymbol{X}_{ia(r,k)}^{\top} \right\} \left\{ \hat{\mathbf{\Gamma}}_{a(r,k),a(r)^c} - \mathbf{\Gamma}_{a(r,k),a(r)^c} \right\} \boldsymbol{\beta}_{a(r)^c} \\
& + \left\{ \boldsymbol{\gamma}_{j,a(r,k)} - \widehat{\boldsymbol{\gamma}}_{j,a(r,k)} \right\}^{\top} \left\{ \frac{1}{n_r} \sum_{i=1}^{n} I(\xi_i = r) \boldsymbol{X}_{ia(r,k)} \boldsymbol{X}_{ia(r,k)}^{\top} \right\} \left\{ \hat{\mathbf{\Gamma}}_{a(r,k),a(r)^c} - \mathbf{\Gamma}_{a(r,k),a(r)^c} \right\} \boldsymbol{\beta}_{a(r)^c} \\
& + \left\{ \boldsymbol{\gamma}_{j,a(r,k)} - \widehat{\boldsymbol{\gamma}}_{j,a(r,k)} \right\}^{\top} \left\{ \frac{1}{n_r} \sum_{i=1}^{n} I(\xi_i = r) \boldsymbol{X}_{ia(r,k)} \left\{ \boldsymbol{\epsilon}_{ia(r)^c}^{(k)} \right\}^{\top} \right\} \boldsymbol{\beta}_{a(r)^c} \\
& + \left\{ \boldsymbol{\gamma}_{j,a(r,k)} - \widehat{\boldsymbol{\gamma}}_{j,a(r,k)} \right\}^{\top} \frac{1}{n_r} \sum_{i=1}^{n} I(\xi_i = r) \epsilon_i \boldsymbol{X}_{ia(r,k)} \\
= \; & \hat{\theta}_r^{-1} \boldsymbol{\gamma}_{j,a(r,k)}^{\top} \left\{ \frac{1}{n} \sum_{i=1}^{n} I(\xi_i = r) \boldsymbol{X}_{ia(r,k)} \boldsymbol{X}_{ia(r,k)}^{\top} \right\} \left\{ \hat{\mathbf{\Gamma}}_{a(r,k),a(r)^c} - \mathbf{\Gamma}_{a(r,k),a(r)^c} \right\} \boldsymbol{\beta}_{a(r)^c} \\
& + \hat{\theta}_r^{-1} \left\{ \boldsymbol{\gamma}_{j,a(r,k)} - \widehat{\boldsymbol{\gamma}}_{j,a(r,k)} \right\}^{\top} \left\{ \frac{1}{n} \sum_{i=1}^{n} I(\xi_i = r) \boldsymbol{X}_{ia(r,k)} \boldsymbol{X}_{ia(r,k)}^{\top} \right\} \left\{ \hat{\mathbf{\Gamma}}_{a(r,k),a(r)^c} - \mathbf{\Gamma}_{a(r,k),a(r)^c} \right\} \boldsymbol{\beta}_{a(r)^c} \\
& + \hat{\theta}_r^{-1} \left\{ \boldsymbol{\gamma}_{j,a(r,k)} - \widehat{\boldsymbol{\gamma}}_{j,a(r,k)} \right\}^{\top} \left\{ \frac{1}{n} \sum_{i=1}^{n} I(\xi_i = r) \boldsymbol{X}_{ia(r,k)} \left\{ \boldsymbol{\epsilon}_{ia(r)^c}^{(k)} \right\}^{\top} \right\} \boldsymbol{\beta}_{a(r)^c} \\
& + \hat{\theta}_r^{-1} \left\{ \boldsymbol{\gamma}_{j,a(r,k)} - \widehat{\boldsymbol{\gamma}}_{j,a(r,k)} \right\}^{\top} \frac{1}{n} \sum_{i=1}^{n} I(\xi_i = r) \epsilon_i \boldsymbol{X}_{ia(r,k)} \\
= \; & \hat{\theta}_r^{-1} \boldsymbol{\gamma}_{j,a(r,k)}^{\top} \mathbf{S}_n \left\{ \hat{\mathbf{\Gamma}}_{a(r,k),a(r)^c} - \mathbf{\Gamma}_{a(r,k),a(r)^c} \right\} \boldsymbol{\beta}_{a(r)^c} \\
& + \hat{\theta}_r^{-1} \boldsymbol{\gamma}_{j,a(r,k)}^{\top} \mathbf{C}^*_{a(r,k),a(r,k)} \left\{ \hat{\mathbf{\Gamma}}_{a(r,k),a(r)^c} - \mathbf{\Gamma}_{a(r,k),a(r)^c} \right\} \boldsymbol{\beta}_{a(r)^c} \\
& + \hat{\theta}_r^{-1} \left\{ \boldsymbol{\gamma}_{j,a(r,k)} - \widehat{\boldsymbol{\gamma}}_{j,a(r,k)} \right\}^{\top} \mathbf{S}_n \left\{ \hat{\mathbf{\Gamma}}_{a(r,k),a(r)^c} - \mathbf{\Gamma}_{a(r,k),a(r)^c} \right\} \boldsymbol{\beta}_{a(r)^c} \\
& + \hat{\theta}_r^{-1} \left\{ \boldsymbol{\gamma}_{j,a(r,k)} - \widehat{\boldsymbol{\gamma}}_{j,a(r,k)} \right\}^{\top} \mathbf{C}^*_{a(r,k),a(r,k)} \left\{ \hat{\mathbf{\Gamma}}_{a(r,k),a(r)^c} - \mathbf{\Gamma}_{a(r,k),a(r)^c} \right\} \boldsymbol{\beta}_{a(r)^c}
\end{aligned}
$$

$$+\hat{\theta}_r^{-1}\left\{\boldsymbol{\gamma}_{j,a(r,k)}-\widehat{\boldsymbol{\gamma}}_{j,a(r,k)}\right\}^\top\left\{\frac{1}{n}\sum_{i=1}^n I(\xi_i=r)\boldsymbol{X}_{ia(r,k)}\left\{\boldsymbol{\epsilon}_{ia(r)^c}^{(k)}\right\}^\top\right\}\boldsymbol{\beta}_{a(r)^c}$$

$$+\hat{\theta}_r^{-1}\left\{\boldsymbol{\gamma}_{j,a(r,k)}-\widehat{\boldsymbol{\gamma}}_{j,a(r,k)}\right\}^\top\frac{1}{n}\sum_{i=1}^n I(\xi_i=r)\epsilon_i\boldsymbol{X}_{ia(r,k)}, \tag{S7.35}$$

where $\mathbf{C}^*_{a(r,k),a(r,k)}=\mathbb{E}\left\{I(\xi_i=r)\boldsymbol{X}_{ia(r,k)}\boldsymbol{X}_{ia(r,k)}^\top\right\}$ and

$$\mathbf{S}_n=\frac{1}{n}\sum_{i=1}^n I(\xi_i=r)\boldsymbol{X}_{ia(r,k)}\boldsymbol{X}_{ia(r,k)}^\top-\mathbf{C}^*_{a(r,k),a(r,k)}.$$

For the first term in (S7.35), we have, with probability at least $1-p^{-c}$

$$\begin{aligned}
&\left|\boldsymbol{\gamma}_{j,a(r,k)}^\top\mathbf{S}_n\left\{\hat{\boldsymbol{\Gamma}}_{a(r,k),a(r)^c}-\boldsymbol{\Gamma}_{a(r,k),a(r)^c}\right\}\boldsymbol{\beta}_{a(r)^c}\right|\\
\leq\ &\left\|\boldsymbol{\gamma}_{j,a(r,k)}\right\|_1\left\|\mathbf{S}_n\right\|_\infty\left\|(\hat{\boldsymbol{\Gamma}}_{a(r,k),a(r)^c}-\boldsymbol{\Gamma}_{a(r,k),a(r)^c})\boldsymbol{\beta}_{a(r)^c}\right\|_1\\
\leq\ &C\sqrt{s}\sqrt{\frac{\log p}{n}}s^{3/2}\sqrt{\frac{\log p}{N+n}}\\
\lesssim\ &\frac{s^2\log p}{\sqrt{(N+n)n}}
\end{aligned}\tag{S7.36}$$

where the second last inequality follows from

$$\left\|(\hat{\boldsymbol{\Gamma}}_{a(r,k),a(r)^c}-\boldsymbol{\Gamma}_{a(r,k),a(r)^c})\boldsymbol{\beta}_{a(r)^c}\right\|_1\leq\max_{j\in a(r)^c}\left\|\widehat{\boldsymbol{\gamma}}_{j,a(r,k)}-\boldsymbol{\gamma}_{j,a(r,k)}\right\|_1\left\|\boldsymbol{\beta}_{a(r)^c}\right\|_1\lesssim s^{3/2}\sqrt{\frac{\log p}{N+n}}.$$

and

$$P\left(\left\|\mathbf{S}_n\right\|_\infty\geq C\sqrt{\frac{\log p}{n}}\right)\leq p^{-c}, \tag{S7.37}$$

which is a direct consequence of the concentration inequality of sub-exponential random variables. For the second term in (S7.35), we have with probability at least $1-p^{-c}$

$$\begin{aligned}
&\left|\boldsymbol{\gamma}_{j,a(r,k)}^\top\mathbf{C}^*_{a(r,k),a(r,k)}\left\{\hat{\boldsymbol{\Gamma}}_{a(r,k),a(r)^c}-\boldsymbol{\Gamma}_{a(r,k),a(r)^c}\right\}\boldsymbol{\beta}_{a(r)^c}\right|\\
\leq\ &\left\|\boldsymbol{\gamma}_{j,a(r,k)}\right\|_2\left\|\mathbf{C}^*_{a(r,k),a(r,k)}\right\|_2\left\|\left\{\hat{\boldsymbol{\Gamma}}_{a(r,k),a(r)^c}-\boldsymbol{\Gamma}_{a(r,k),a(r)^c}\right\}\boldsymbol{\beta}_{a(r)^c}\right\|_2\\
\lesssim\ &s\sqrt{\frac{\log p}{N+n}},
\end{aligned}\tag{S7.38}$$

where the last inequality follows from (S7.28). For the third term, we have, with probability at least $1 - p^{-c}$,

$$\begin{aligned}
&\left|\left\{\boldsymbol{\gamma}_{j,a(r,k)} - \widehat{\boldsymbol{\gamma}}_{j,a(r,k)}\right\}^\top \mathbf{S}_n \left\{\hat{\boldsymbol{\Gamma}}_{a(r,k),a(r)^c} - \boldsymbol{\Gamma}_{a(r,k),a(r)^c}\right\} \boldsymbol{\beta}_{a(r)^c}\right| \\
\leq\ & \left\|\boldsymbol{\gamma}_{j,a(r,k)} - \widehat{\boldsymbol{\gamma}}_{j,a(r,k)}\right\|_1 \left\|\mathbf{S}_n\right\|_\infty \left\|\hat{\boldsymbol{\Gamma}}_{a(r,k),a(r)^c} - \boldsymbol{\Gamma}_{a(r,k),a(r)^c}\right\|_1 \left\|\boldsymbol{\beta}_{a(r)^c}\right\|_1 \\
\leq\ & Cs\sqrt{\frac{\log p}{N+n}}\sqrt{\frac{\log p}{n}} s\sqrt{\frac{\log p}{N+n}} \left\|\boldsymbol{\beta}_{a(r)^c}\right\|_1 \\
\lesssim\ & \frac{s^2 \log^{3/2} p}{(N+n)\sqrt{n}}.
\end{aligned} \tag{S7.39}$$

The fourth term satisfies

$$\begin{aligned}
&\left|\left\{\boldsymbol{\gamma}_{j,a(r,k)} - \widehat{\boldsymbol{\gamma}}_{j,a(r,k)}\right\}^\top \mathbf{C}^*_{a(r,k),a(r,k)} \left\{\hat{\boldsymbol{\Gamma}}_{a(r,k),a(r)^c} - \boldsymbol{\Gamma}_{a(r,k),a(r)^c}\right\} \boldsymbol{\beta}_{a(r)^c}\right| \\
\leq\ & \left\|\boldsymbol{\gamma}_{j,a(r,k)} - \widehat{\boldsymbol{\gamma}}_{j,a(r,k)}\right\|_2 \left\|\mathbf{C}^*_{a(r,k),a(r,k)}\right\|_2 \left\|\left\{\hat{\boldsymbol{\Gamma}}_{a(r,k),a(r)^c} - \boldsymbol{\Gamma}_{a(r,k),a(r)^c}\right\} \boldsymbol{\beta}_{a(r)^c}\right\|_2 \\
\lesssim\ & \sqrt{\frac{s\log p}{N+n}} \cdot s\sqrt{\frac{\log p}{N+n}} \\
\lesssim\ & \frac{s^{3/2}\log p}{N+n},
\end{aligned} \tag{S7.40}$$

with probability at least $1 - p^{-c}$. The fifth term satisfies

$$\begin{aligned}
&\left|\left\{\boldsymbol{\gamma}_{j,a(r,k)} - \widehat{\boldsymbol{\gamma}}_{j,a(r,k)}\right\}^\top \left\{\frac{1}{n}\sum_{i=1}^n I(\xi_i = r) \boldsymbol{X}_{ia(r,k)} \left\{\boldsymbol{\epsilon}^{(k)}_{ia(r)^c}\right\}^\top\right\} \boldsymbol{\beta}_{a(r)^c}\right| \\
\leq\ & \left\|\boldsymbol{\gamma}_{j,a(r,k)} - \widehat{\boldsymbol{\gamma}}_{j,a(r,k)}\right\|_1 \left\|\frac{1}{n}\sum_{i=1}^n I(\xi_i = r) \boldsymbol{X}_{ia(r,k)} \left\{\boldsymbol{\epsilon}^{(k)}_{ia(r)^c}\right\}^\top\right\|_\infty \left\|\boldsymbol{\beta}_{a(r)^c}\right\|_1 \\
\lesssim\ & s\sqrt{\frac{\log p}{N+n}}\sqrt{\frac{\log p}{n}} \left\|\boldsymbol{\beta}_{a(r)^c}\right\|_1 \\
\lesssim\ & \frac{s^{3/2}\log p}{\sqrt{(N+n)n}},
\end{aligned} \tag{S7.41}$$

with probability at least $1 - p^{-c}$. Lastly, the sixth term satisfies

$$\left| \left\{ \boldsymbol{\gamma}_{j,a(r,k)} - \widehat{\boldsymbol{\gamma}}_{j,a(r,k)} \right\}^\top \frac{1}{n} \sum_{i=1}^n I(\xi_i = r) \epsilon_i \boldsymbol{X}_{ia(r,k)} \right| \tag{S7.42}$$

$$\leq \left\| \boldsymbol{\gamma}_{j,a(r,k)} - \widehat{\boldsymbol{\gamma}}_{j,a(r,k)} \right\|_1 \left\| \frac{1}{n} \sum_{i=1}^n I(\xi_i = r) \epsilon_i \boldsymbol{X}_{ia(r,k)} \right\|_\infty$$

$$\lesssim s \sqrt{\frac{\log p}{N+n}} \sqrt{\frac{\log p}{n}} \tag{S7.43}$$

with probability at least $1 - p^{-c}$. Combining all the pieces together, it follows that with probability at least $1 - p^{-c}$,

$$\|\boldsymbol{g}_n(\boldsymbol{\beta}) - \boldsymbol{g}(\boldsymbol{\beta})\|_\infty \lesssim s \sqrt{\frac{\log p}{N+n}} + s \sqrt{\frac{\log p}{N+n}} \cdot s \sqrt{\frac{\log p}{n}} \asymp s \sqrt{\frac{\log p}{N+n}} \left(1 + s \sqrt{\frac{\log p}{n}} \right).$$

where the last inequality follow from $s \lesssim (n+N)/\log p$.

S7.4 Proof of Proposition 1

The proof is separated into two parts. For simplicity, we write $E_s(p)$ as E_s when there is no risk of confusion. In the first part, we show that with probability at least $1 - p^{-c}$,

$$\sup_{\substack{\|\mathbf{u}\|_2 = 1, \mathbf{u} \in E_s \\ \|\mathbf{u}_{a(k)}\|_2 \geq 1/2}} \left| \frac{\mathbf{u}_{a(k)}^\top}{\|\mathbf{u}_{a(k)}\|_2} \frac{1}{n} \sum_{i=1}^n \frac{n}{n_r} I\{\xi_i = r\} [\widehat{\boldsymbol{X}}_{ia(k)}^{(k)} (\widehat{\boldsymbol{X}}_i^{(k)})^\top - \boldsymbol{X}_{ia(k)}^{(k)} (\boldsymbol{X}_i^{(k)})^\top] \mathbf{u} \right| \leq c/4. \tag{S7.44}$$

In the second part, we show that with probability at least $1 - p^{-c}$,

$$\inf_{\substack{\|\mathbf{u}\|_2 = 1, \mathbf{u} \in E_s \\ \|\mathbf{u}_{a(k)}\|_2 \geq 1/2}} \left| \frac{\mathbf{u}_{a(k)}^\top}{\|\mathbf{u}_{a(k)}\|_2} \frac{1}{n} \sum_{i=1}^n \frac{n}{n_r} I\{\xi_i = r\} \boldsymbol{X}_{ia(k)}^{(k)} (\boldsymbol{X}_i^{(k)})^\top \mathbf{u} \right| \geq c. \tag{S7.45}$$

The above inequalities implies (S5.14).

Part I. Note that by Hölder's inequality

$$\sup_{\substack{\|\mathbf{u}\|_2 = 1, \mathbf{u} \in E_s \\ \|\mathbf{u}_{a(k)}\|_2 \geq 1/2}} \left| \frac{\mathbf{u}_{a(k)}^\top}{\|\mathbf{u}_{a(k)}\|_2} \frac{1}{n} \sum_{i=1}^n \frac{n}{n_r} I\{\xi_i = r\} [\widehat{\boldsymbol{X}}_{ia(k)}^{(k)} (\widehat{\boldsymbol{X}}_i^{(k)})^\top - \boldsymbol{X}_{ia(k)}^{(k)} (\boldsymbol{X}_i^{(k)})^\top] \mathbf{u} \right|$$

$$\lesssim s \left\| \frac{1}{n} \sum_{i=1}^n \frac{n}{n_r} I\{\xi_i = r\} [\widehat{\boldsymbol{X}}_{ia(k)}^{(k)} (\widehat{\boldsymbol{X}}_i^{(k)})^\top - \boldsymbol{X}_{ia(k)}^{(k)} (\boldsymbol{X}_i^{(k)})^\top] \right\|_\infty .$$

By definition, it holds that, after proper permutation of coordinates

$$\widehat{\boldsymbol{X}}_i^{(k)} = ((\boldsymbol{X}_{ia(r)}^{(k)})^\top, (\widehat{\boldsymbol{X}}_{ia(r)^c}^{(k)})^\top)^\top = (\boldsymbol{X}_{ia(r)}^\top, \boldsymbol{X}_{ia(r,k)}^\top \hat{\boldsymbol{\Gamma}}_{r,k})^\top = \begin{bmatrix} \mathbf{I}_{|a(r)|} \\ [\hat{\boldsymbol{\Gamma}}_{r,k}^\top, \mathbf{0}] \end{bmatrix} \boldsymbol{X}_{ia(r)},$$

and

$$\widehat{\boldsymbol{X}}_{ia(k)}^{(k)} = ((\boldsymbol{X}_{i,a(k)\cap a(r)}^{(k)})^\top, (\widehat{\boldsymbol{X}}_{ia(r)^c}^{(k)})^\top)^\top = (\boldsymbol{X}_{ia(r,k)}^\top, \boldsymbol{X}_{ia(r,k)}^\top \hat{\boldsymbol{\Gamma}}_{r,k})^\top = \begin{bmatrix} \mathbf{I}_{|a(r,k)|} \\ \hat{\boldsymbol{\Gamma}}_{r,k}^\top \end{bmatrix} \boldsymbol{X}_{ia(r,k)}.$$

Note that

$$\begin{aligned}
&\left\| \frac{1}{n} \sum_{i=1}^n \frac{n}{n_r} I\{\xi_i = r\} [\widehat{\boldsymbol{X}}_{ia(k)}^{(k)} (\widehat{\boldsymbol{X}}_i^{(k)})^\top - \boldsymbol{X}_{ia(k)}^{(k)} (\boldsymbol{X}_i^{(k)})^\top] \right\|_\infty \\
&\le \left\| \frac{1}{n} \sum_{i=1}^n \frac{n}{n_r} I\{\xi_i = r\} [\widehat{\boldsymbol{X}}_{ia(k)}^{(k)} - \boldsymbol{X}_{ia(k)}^{(k)}] (\boldsymbol{X}_i^{(k)})^\top \right\|_\infty + \left\| \frac{1}{n} \sum_{i=1}^n \frac{n}{n_r} I\{\xi_i = r\} \boldsymbol{X}_{ia(k)}^{(k)} (\boldsymbol{X}_i^{(k)} - \widehat{\boldsymbol{X}}_i^{(k)})^\top \right\|_\infty \\
&\quad + \left\| \frac{1}{n} \sum_{i=1}^n \frac{n}{n_r} I\{\xi_i = r\} (\widehat{\boldsymbol{X}}_{ia(k)}^{(k)} - \boldsymbol{X}_{ia(k)}^{(k)}) (\boldsymbol{X}_i^{(k)} - \widehat{\boldsymbol{X}}_i^{(k)})^\top \right\|_\infty
\end{aligned}$$

To control the first term, we have

$$\begin{aligned}
&\left\| \frac{1}{n} \sum_{i=1}^n \frac{n}{n_r} I\{\xi_i = r\} [\widehat{\boldsymbol{X}}_{ia(k)}^{(k)} - \boldsymbol{X}_{ia(k)}^{(k)}] (X_i^{(k)})^\top \right\|_\infty \\
&= \left\| \frac{1}{n} \sum_{i=1}^n \frac{n}{n_r} I\{\xi_i = r\} \begin{bmatrix} \mathbf{0} \\ \hat{\boldsymbol{\Gamma}}_{r,k}^\top - \boldsymbol{\Gamma}_{r,k}^\top \end{bmatrix} \boldsymbol{X}_{ia(r,k)} \boldsymbol{X}_{ia(r)}^\top \begin{bmatrix} \mathbf{I}_{|a(r)|} & [\boldsymbol{\Gamma}_{r,k}, \mathbf{0}]^\top \end{bmatrix} \right\|_\infty \\
&= \max_{\substack{1 \le j \le |a(k)| \\ 1 \le k \le p}} \left| \frac{1}{n} \sum_{i=1}^n \frac{n}{n_r} I\{\xi_i = r\} \mathbf{u}_j^\top \boldsymbol{X}_{ia(r,k)} \boldsymbol{X}_{ia(r)}^\top \mathbf{v}_k \right|
\end{aligned}$$

where $\mathbf{u}_j$ is the j-th row of $\begin{bmatrix} \mathbf{0} \\ \hat{\boldsymbol{\Gamma}}_{r,k}^\top - \boldsymbol{\Gamma}_{r,k}^\top \end{bmatrix}$ and $\mathbf{v}_k$ is the k-th row of $\begin{bmatrix} \mathbf{I}_{|a(r)|} \\ [\boldsymbol{\Gamma}_{r,k}^\top, \mathbf{0}] \end{bmatrix}$. Now since for any $j \in a(r)^c$, $\|[\boldsymbol{\Gamma}_{r,k}]_{.j}\|_2 \le C$, and for all $1 \le k \le p$, $\|\mathbf{v}_k\|_2 \le C$, then $\mathbf{X}_{ia(r)}^\top \mathbf{v}_k$ is sub-Gaussian random variable with parameter bounded by some absolute constant. On the other hand, by the Lasso bound (S5.3) for the columns of $\hat{\boldsymbol{\Gamma}}_{r,k}$, we have $\|\mathbf{u}_j\|_2 \le \sqrt{s \log p/(n+N)}$,

so that conditional on $\hat{\mathbf{\Gamma}}_{r,k}$, $\mathbf{u}_j^\top \mathbf{X}_{ia(r,k)}$ is a sub-Gaussian random variable with parameter asymptotically bounded by $\sqrt{s\log p/(n+N)}$. In addition, with probability at least $1-p^{-c}$, the mean value satisfies

$$\left|\frac{n}{n_r}\mathbf{u}_j^\top \mathbb{E} I\{\xi_i = r\}\boldsymbol{X}_{ia(r,k)}\boldsymbol{X}_{ia(r)}^\top \mathbf{v}_k\right| \lesssim \sqrt{\frac{s\log p}{N+n}}\lambda_{\max}(\Sigma)$$

where the last inequality follows from $\lambda_{\max}(\mathbb{E} I\{\xi_i = r\}\boldsymbol{X}_i\boldsymbol{X}_i^\top) = \sup_{\|u\|_2=1}\mathbb{E} I\{\xi_i = r\}(\boldsymbol{X}_i^\top \mathbf{u})^2 \le \lambda_{\max}(\Sigma)$. Thus, by concentration inequality for sub-exponential random variables,

$$P\bigg(\max_{\substack{1\le j\le |a(k)|\\ 1\le k\le p}}\left|\frac{1}{n}\sum_{i=1}^n \frac{n}{n_r} I\{\xi_i = r\}\mathbf{u}_j^\top \boldsymbol{X}_{ia(r,k)}\boldsymbol{X}_{ia(r)}^\top \mathbf{v}_k\right| \le C_1\sqrt{\frac{s\log p}{N+n}} + C_2\frac{\sqrt{s}\log p}{\sqrt{(n+N)n}}\bigg) \ge 1-p^{-c}.$$

Similarly, one can also show that

$$P\bigg(\bigg\|\frac{1}{n}\sum_{i=1}^n \frac{n}{n_r} I\{\xi_i = r\}\boldsymbol{X}_{ia(k)}^{(k)}(\boldsymbol{X}_i^{(k)} - \widehat{\boldsymbol{X}}_i^{(k)})^\top]\bigg\|_\infty \le C\sqrt{\frac{s\log p}{N+n}}\bigg) \ge 1-p^{-c},$$

and

$$P\bigg(\bigg\|\frac{1}{n}\sum_{i=1}^n \frac{n}{n_r} I\{\xi_i = r\}[\widehat{\boldsymbol{X}}_{ia(k)}^{(k)} - \boldsymbol{X}_{ia(k)}^{(k)}](\boldsymbol{X}_i^{(k)} - \widehat{\boldsymbol{X}}_i^{(k)})^\top\bigg\|_\infty \le C\sqrt{\frac{s\log p}{N+n}}\bigg) \ge 1-p^{-c}.$$

Since $s \lesssim \frac{N+n}{\log p}$, we have shown (S7.44) holds with high probability.

Part II. If we denote

$$\mathbf{H}_n = \frac{1}{n}\sum_{i=1}^n \frac{n}{n_r} I\{\xi_i = r\}\boldsymbol{X}_{ia(k)}^{(k)}(\boldsymbol{X}_i^{(k)})^\top,$$

we have

$$\begin{aligned}\inf_{\substack{\mathbf{u}\in E_s\\ \|\mathbf{u}_{a(k)}\|_2\ge 1/2}}\left|\frac{\mathbf{u}_{a(k)}^\top}{\|\mathbf{u}_{a(k)}\|_2}\mathbf{H}_n\mathbf{u}\right| &\ge \inf_{\substack{\mathbf{u}\in E_s\\ \|\mathbf{u}_{a(k)}\|_2\ge 1/2}}\frac{\mathbf{u}_{a(k)}^\top[\mathbf{H}_n]_{.a(k)}\mathbf{u}_{a(k)}}{\|\mathbf{u}_{a(k)}\|_2} - \sup_{\substack{\mathbf{u}\in E_s\\ \|\mathbf{u}_{a(k)}\|_2\ge 1/2}}\frac{|\mathbf{u}_{a(k)}^\top[\mathbf{H}_n]_{.m(k)}\mathbf{u}_{m(k)}|}{\|\mathbf{u}_{a(k)}\|_2}\\ &\ge \frac{1}{2}\inf_{\|\mathbf{v}\|_2=1,\|\mathbf{v}\|_1\le 4\sqrt{s}}|\mathbf{v}^\top[\mathbf{H}_n]_{.a(k)}\mathbf{v}| - \sup_{\substack{\|\mathbf{v}\|_1\le 4\sqrt{s},\|\mathbf{u}\|_1\le 2\sqrt{s}\\ \|\mathbf{v}\|_2=1,\|\mathbf{u}\|_2=1}}|\mathbf{v}^\top[\mathbf{H}_n]_{.m(k)}\mathbf{u}|,\end{aligned}$$

where the last inequality follows that, for any $\mathbf{u}\in E_s$ such that $\|\mathbf{u}_{a(k)}\|_2 \ge 1/2$,

$$\left\|\frac{\mathbf{u}_{a(k)}}{\|\mathbf{u}_{a(k)}\|_2}\right\|_1 \le 2\|\mathbf{u}_{a(k)}\|_1 \le 2\|\mathbf{u}\|_1 \le 4\sqrt{s},$$

$$\|\mathbf{u}_{m(k)}\|_1 \le \|\mathbf{u}\|_1 \le 2\sqrt{s}.$$

If we can show that, with probability at least $1 - p^{-c}$,

$$\frac{1}{2} \inf_{\|\mathbf{v}\|_2=1, \|\mathbf{v}\|_1 \le 4\sqrt{s}} |\mathbf{v}^\top [\mathbf{H}_n]_{.a(k)} \mathbf{v}| > 3c, \tag{S7.46}$$

and

$$\sup_{\substack{\|\mathbf{v}\|_1 \le 4\sqrt{s}, \|\mathbf{u}\|_1 \le 2\sqrt{s} \\ \|\mathbf{v}\|_2=1, \|\mathbf{u}\|_2=1}} |\mathbf{v}^\top [\mathbf{H}_n]_{.m(k)} \mathbf{u}| < 2c, \tag{S7.47}$$

it follows that under the same event (S5.14) holds. Suppose

$$\mathbf{\Sigma}^{(k)} = \mathbb{E} \frac{n}{n_r} I\{\xi_i = r\} \boldsymbol{X}_i^{(k)} (\boldsymbol{X}_i^{(k)}) \in \mathbb{R}^{p \times p}.$$

In the following, we will show that, under the condition

$$\lambda_{\min}(\mathbf{\Sigma}^{(k)}_{a(k),a(k)}) \ge 7c > c \ge \lambda_{\max}(\mathbf{\Sigma}^{(k)}_{a(k),m(k)}), \tag{S7.48}$$

inequalities (S7.46) and (S7.47) holds with the stated probability.

In fact, under condition (S7.48), it can be shown through Theorem 2.5 of Zhou (2009) (see also Corollary 2.7 of Mendelson et al. (2007) or Theorem 2.1 of Mendelson et al. (2008)) that, for $s \ll \frac{n}{\log p}$ and $n \gg \log p$, with probability at least $1 - e^{-cn}$, (S7.46) holds. Specifically, by Theorem 2.5 of Zhou (2009), the proof of (S7.46) reduces to show that

$$\mathbb{E} \sup_{\|\mathbf{v}\|_2=1, \|\mathbf{v}\|_1 \le 4\sqrt{s}} \left| \langle \mathbf{g}, [\mathbf{\Sigma}^{(k)}_{a(k),a(k)}]^{1/2} \mathbf{v} \rangle \right| \lesssim \sqrt{s \log p} \ll \sqrt{n}, \tag{S7.49}$$

where $\mathbf{g} \sim N(0, \mathbf{I}_p)$. To see this, simply note that

$$\mathbb{E} \sup_{\|\mathbf{v}\|_2=1, \|\mathbf{v}\|_1 \le 4\sqrt{s}} \left| \langle \mathbf{g}, [\mathbf{\Sigma}^{(k)}_{a(k),a(k)}]^{1/2} \mathbf{v} \rangle \right| \le \sup_{\|\mathbf{v}\|_2=1, \|\mathbf{v}\|_1 \le 4\sqrt{s}} \|\mathbf{v}\|_1 \cdot \mathbb{E} \|[\mathbf{\Sigma}^{(k)}_{a(k),a(k)}]^{1/2} \mathbf{g}\|_\infty \lesssim \sqrt{s \log p},$$

where the last inequality follows from the following property of the maxima of sub-Gaussian random variables:

$$\mathbb{E} \max_{1 \le i \le n} g_i \le C \sqrt{\sigma^2 \log n}$$

where $g_1, ..., g_n$ are centred sub-Gaussian random variables with parameter σ^2, not necessarily independent.

It remains to show (S7.47). Note that

$$\begin{aligned}\sup_{\substack{\|\mathbf{v}\|_1\le 4\sqrt{s},\|\mathbf{u}\|_1\le 2\sqrt{s}\\ \|\mathbf{v}\|_2=1,\|\mathbf{u}\|_2=1}} |\mathbf{v}^\top [\mathbf{H}_n]_{.m(k)}\mathbf{u}| \le & \sup_{\substack{\|\mathbf{v}\|_1\le 4\sqrt{s},\|\mathbf{u}\|_1\le 2\sqrt{s}\\ \|\mathbf{v}\|_2=1,\|\mathbf{u}\|_2=1}} |\mathbf{v}^\top [\mathbf{H}_n]_{.m(k)}\mathbf{u} - \mathbb{E}\mathbf{v}^\top [\mathbf{H}_n]_{.m(k)}\mathbf{u}| \\ & + \sup_{\substack{\|\mathbf{v}\|_1\le 4\sqrt{s},\|\mathbf{u}\|_1\le 2\sqrt{s}\\ \|\mathbf{v}\|_2=1,\|\mathbf{u}\|_2=1}} \mathbb{E}\mathbf{v}^\top [\mathbf{H}_n]_{.m(k)}\mathbf{u} \\ \le & \sup_{\substack{\|\mathbf{v}\|_1\le 4\sqrt{s},\|\mathbf{u}\|_1\le 2\sqrt{s}\\ \|\mathbf{v}\|_2=1,\|\mathbf{u}\|_2=1}} |\mathbf{v}^\top [\mathbf{H}_n]_{.m(k)}\mathbf{u} - \mathbb{E}\mathbf{v}^\top [\mathbf{H}_n]_{.m(k)}\mathbf{u}| \\ & + \lambda_{\max}(\mathbf{\Sigma}^{(k)}_{a(k),m(k)}).\end{aligned}$$

It suffices to show that, with probability at least $1 - p^{-c}$,

$$\sup_{\substack{\|\mathbf{v}\|_1\le 4\sqrt{s},\|\mathbf{u}\|_1\le 2\sqrt{s}\\ \|\mathbf{v}\|_2=1,\|\mathbf{u}\|_2=1}} |\mathbf{v}^\top [\mathbf{H}_n]_{.m(k)}\mathbf{u} - \mathbb{E}\mathbf{v}^\top [\mathbf{H}_n]_{.m(k)}\mathbf{u}| \le c. \tag{S7.50}$$

To see this, note that

$$\begin{aligned}& \sup_{\substack{\|\mathbf{v}\|_1\le 4\sqrt{s},\|\mathbf{u}\|_1\le 2\sqrt{s}\\ \|\mathbf{v}\|_2=1,\|\mathbf{u}\|_2=1}} |\mathbf{v}^\top [\mathbf{H}_n]_{.m(k)}\mathbf{u} - \mathbb{E}\mathbf{v}^\top [\mathbf{H}_n]_{.m(k)}\mathbf{u}| \\ \le & \sup_{\substack{\|\mathbf{v}\|_1\le 4\sqrt{s},\|\mathbf{u}\|_1\le 2\sqrt{s}\\ \|\mathbf{v}\|_2=1,\|\mathbf{u}\|_2=1}} \|\mathbf{v}\|_1\|\mathbf{u}\|_1 \cdot \|[\mathbf{H}_n]_{.m(k)} - \mathbb{E}[\mathbf{H}_n]_{.m(k)}\|_\infty,\end{aligned}$$

and

$$\sup_{\substack{\|\mathbf{v}\|_1\le 4\sqrt{s},\|\mathbf{u}\|_1\le 2\sqrt{s}\\ \|\mathbf{v}\|_2=1,\|\mathbf{u}\|_2=1}} \|\mathbf{u}\|_1\|\mathbf{v}\|_1 \le 8s.$$

Then by the concentration inequality for independent sub-exponential random variables, we have, for any $(i, j) \in \{1, ..., |a(k)|\} \times m(k)$,

$$P\left(|[\mathbf{H}_n]_{ij} - [\mathbb{E}\mathbf{H}_n]_{ij}| \le C\sqrt{\frac{\log p}{n}}\right) \ge 1 - p^{-c},$$

so that

$$P\left(\|\mathbf{H}_n - \mathbb{E}\mathbf{H}_n\|_\infty \le C\sqrt{\frac{\log p}{n}}\right) \ge 1 - p^{-c}.$$

Then as long as $s \lesssim \sqrt{n/\log p}$, we have (S7.50) hold with probability at least $1 - p^{-c}$. This completes the proof of the proposition.

S7.5 Proof of Lemma 4

By the theory of inhomogeneous system of linear equations, $\mathbb{E}\boldsymbol{G}^\top \boldsymbol{v}_j = \boldsymbol{e}_j$ has infinitely numbers of solutions if $\text{rank}(\mathbb{E}\boldsymbol{G}) = \text{rank}([\mathbb{E}\boldsymbol{G}, \boldsymbol{e}_j]) < M_g'$. When the smallest singular value of $\mathbb{E}\boldsymbol{G}$ is nonnegative, we have $\text{rank}(\mathbb{E}\boldsymbol{G}) = p$, and the matrix $[\mathbb{E}\boldsymbol{G}, \boldsymbol{e}_j] \in \mathbb{R}^{p\times(M_g'+1)}$ is also of rank p, which by assumption is smaller than M_g'.

Secondly, note that

$$1 = \|\boldsymbol{e}_j\| = \|\mathbb{E}\boldsymbol{G}^\top \boldsymbol{v}_j\| \ge \lambda_{\min}(\mathbb{E}\boldsymbol{G})\|\boldsymbol{v}_j\|_2.$$

Then as long as $\lambda_{\min}(\mathbb{E}\boldsymbol{G}) \ge c > 0$, we have $\|\boldsymbol{v}_j\|_2 \le C$.

S7.6 Proof of Lemma 5

Since

$$\|(\boldsymbol{v}_j^*)^\top \boldsymbol{G}_n - \boldsymbol{e}_j^\top\|_\infty \le \|(\boldsymbol{v}_j^*)^\top(\boldsymbol{G}_n - \boldsymbol{G})\|_\infty + \|(\boldsymbol{v}_j^*)^\top \boldsymbol{G} - \boldsymbol{e}_j^\top\|_\infty,$$

it suffices to show that

$$\|(\boldsymbol{v}_j^*)^\top \boldsymbol{G} - \boldsymbol{e}_j^\top\|_\infty \lesssim \sqrt{\frac{\log p}{n}}, \tag{S7.51}$$

and

$$\|(\boldsymbol{v}_j^*)^\top(\boldsymbol{G}_n - \boldsymbol{G})\|_\infty = \max_{1\le j\le p} |(\boldsymbol{v}_j^*)^\top[\boldsymbol{G}_n - \boldsymbol{G}]_{\cdot j}| \lesssim R^2\sqrt{\frac{s\log p}{N+n}}, \tag{S7.52}$$

holds with probability at least $1 - p^{-c}$.

To prove (S7.51), we use the similar argument as in the proof of Lemma 1. We notice that each component of $(\boldsymbol{v}_j^*)^\top \boldsymbol{G}$ is an average of n independent sub-exponential random variables and $\mathbb{E}(\boldsymbol{v}_j^*)^\top \boldsymbol{G} = \boldsymbol{e}_j^\top$. Then by concentration inequality for sub-exponential random variables (Vershynin, 2010), inequality (S7.51) holds with probability at least $1 - p^{-c}$.

The rest of the proof is devoted to (S7.52). Let $\boldsymbol{h}_{irk}^*(\boldsymbol{\beta}) = (h_{irk\ell}^*(\boldsymbol{\beta}))_{\ell \in a(r,k)}$. We can write

$$\boldsymbol{G}_{\cdot j} = \left[\frac{1}{n_r} \sum_{i=1}^n \frac{\partial h_{irk\ell}^*(\boldsymbol{\beta})}{\partial \beta_j} \right]_{1 \le r \le R, k \in \mathcal{G}(r), \ell \in a(r,k)},$$

where the M_g elements are ordered first by the index r, and then k, and finally ℓ. Now for each (r, k, ℓ), we have

$$\frac{\partial h_{irk\ell}^*(\boldsymbol{\beta})}{\partial \beta_j} = -I(\xi_i = r) X_{ij}^{(k)} X_{i\ell}.$$

If j-th variable is not observed in ith sample, we have

$$\frac{\partial h_{irk\ell}^*(\boldsymbol{\beta})}{\partial \beta_j} = -I(\xi_i = r) \boldsymbol{\gamma}_{j,a(r,k)}^\top \boldsymbol{X}_{ia(r,k)} X_{i\ell},$$

where we used $X_{ij}^{(k)} = \boldsymbol{\gamma}_{j,a(r,k)}^\top \boldsymbol{X}_{ia(r,k)}$. If j-th variable is observed in ith sample, we have

$$\frac{\partial h_{irk\ell}^*(\boldsymbol{\beta})}{\partial \beta_j} = -I(\xi_i = r) X_{ij} X_{i\ell}.$$

Similarly, if we write

$$[\boldsymbol{G}_n]_{\cdot j} = \left[\frac{1}{n_r} \sum_{i=1}^n \frac{\partial \hat{h}_{irk\ell}^*(\boldsymbol{\beta})}{\partial \beta_j} \right]_{1 \le r \le R, k \in \mathcal{G}(r), \ell \in a(r,k)},$$

where

$$\frac{\partial \hat{h}_{irk\ell}^*(\boldsymbol{\beta})}{\partial \beta_j} = -I(\xi_i = r) \widehat{\boldsymbol{\gamma}}_{j,a(r,k)}^\top \boldsymbol{X}_{ia(r,k)} X_{i\ell},$$

if j-th variable is not observed in ith sample, and

$$\frac{\partial \hat{h}_{irk\ell}^*(\boldsymbol{\beta})}{\partial \beta_j} = -I(\xi_i = r) X_{ij} X_{i\ell},$$

if j-th variable is observed in ith sample. Thus, we have

$$[\boldsymbol{G}-\boldsymbol{G}_n]_{\cdot j}=\left[\tfrac{1}{n_r}\textstyle\sum_{i=1}^{n}(\frac{\partial h^*_{irk\ell}(\boldsymbol{\beta})}{\partial\beta_j}-\frac{\partial\hat{h}^*_{irk\ell}(\boldsymbol{\beta})}{\partial\beta_j})\right]_{1\le r\le R,k\in\mathcal{G}(r),\ell\in a(r,k)}. \tag{S7.53}$$

If the j-th variable is not observed in i-th sample, we have

$$\frac{\partial\hat{h}^*_{irk}(\boldsymbol{\beta})}{\partial\beta_j}-\frac{\partial h^*_{irk}(\boldsymbol{\beta})}{\partial\beta_j}=-I(\xi_i=r)(\widehat{\boldsymbol{\gamma}}_{j,a(r,k)}-\boldsymbol{\gamma}_{j,a(r,k)})^\top\boldsymbol{X}_{ia(r,k)}\boldsymbol{X}_{ia(r,k)} \tag{S7.54}$$

In the following we control the left hand side of (S7.52) for j not observed in Group r. Specifically, let $\boldsymbol{v}^*_{j,rk}$ be the subvector of $\boldsymbol{v}^*_j$ associated to Group $k\in\mathcal{G}(r)$. We have, with probability at least $1-p^{-c}$,

$$\begin{aligned}
&\sum_{r=1}^{R}\sum_{k\in\mathcal{G}(r)}\frac{1}{n_r}\sum_{i=1}^{n}I(\xi_i=r)(\widehat{\boldsymbol{\gamma}}_{j,a(r,k)}-\boldsymbol{\gamma}_{j,a(r,k)})^\top\boldsymbol{X}_{ia(r,k)}\boldsymbol{X}^\top_{ia(r,k)}\boldsymbol{v}^*_{j,rk}\\
&\le\sum_{r=1}^{R}\sum_{k\in\mathcal{G}(r)}\sqrt{\frac{1}{n}\sum_{i=1}^{n}\left[\frac{n}{n_r}I(\xi_i=r)\boldsymbol{X}^\top_{ia(r,k)}(\widehat{\boldsymbol{\gamma}}_{j,a(r,k)}-\boldsymbol{\gamma}_{j,a(r,k)})\right]^2}\\
&\quad\times\sqrt{\frac{1}{n}\sum_{i=1}^{n}I(\xi_i=r)(\boldsymbol{X}^\top_{ia(r,k)}\boldsymbol{v}^*_{j,rk})^2}\\
&\lesssim\sum_{r=1}^{R}\sum_{k\in\mathcal{G}(r)}\sqrt{\frac{1}{n}\sum_{i=1}^{n}\left[\boldsymbol{X}^\top_{ia(r,k)}(\widehat{\boldsymbol{\gamma}}_{j,a(r,k)}-\boldsymbol{\gamma}_{j,a(r,k)})\right]^2}\sqrt{\frac{1}{n}\sum_{i=1}^{n}I(\xi_i=r)(\boldsymbol{X}^\top_{ia(r,k)}\boldsymbol{v}^*_{j,rk})^2}\\
&\lesssim R^2\sqrt{\frac{s\log p}{N+n}}
\end{aligned}$$

where the last inequality follows from the standard Lasso bound

$$P\left(\frac{1}{n}\sum_{i=1}^{n}\left[\boldsymbol{X}^\top_{ia(r,k)}(\widehat{\boldsymbol{\gamma}}_{j,a(r,k)}-\boldsymbol{\gamma}_{j,a(r,k)})\right]^2\le C\frac{s\log p}{N+n}\right)\ge 1-p^{-c}, \tag{S7.55}$$

and the inequality

$$P\left(\max_{1\le r\le R,k\in\mathcal{G}(r)}\frac{1}{n}\sum_{i=1}^{n}I(\xi_i=r)((\boldsymbol{X}^{(k)}_{ia(k)})^\top\boldsymbol{v}^*_{j,rk})^2\le C\right)\ge 1-p^{-c}, \tag{S7.56}$$

which can be established by

$$P\left(\max_{1\le r\le R,k\in\mathcal{G}(r)}\left|\frac{1}{n}\sum_{i=1}^{n}(I(\xi_i=r)\boldsymbol{X}^\top_{ia(r,k)}\boldsymbol{v}^*_{j,rk})^2-\mathbb{E}I(\xi_i=r)(\boldsymbol{X}^\top_{ia(r,k)}\boldsymbol{v}^*_{j,rk})^2\right|\le C\sqrt{\frac{\log p}{n}}\right)\ge 1-p^{-c},$$

as a consequence of Proposition 5.16 of Vershynin (2010), and

$$\mathbb{E}I(\xi_i = r)(\boldsymbol{X}_{ia(r,k)}^\top \boldsymbol{v}_{j,rk}^*)^2 \leq \|\boldsymbol{v}_j^*\|_2^2 \lambda_{\max}(\boldsymbol{\Sigma}) = O(1),$$

where the last inequality follows from Lemma 4 and condition (A3).

If the j-th variable observed in ith sample, we have

$$\frac{\partial h_{irk\ell}^*(\boldsymbol{\beta})}{\partial \beta_j} - \frac{\partial \hat{h}_{irk\ell}^*(\boldsymbol{\beta})}{\partial \beta_j} = 0 \tag{S7.57}$$

Then combining the above two pieces, we have shown that (S7.52) holds.

S7.7 Proof of Lemma 6

By mean value theorem, we have

$$\hat{S}_j(\hat{\boldsymbol{\beta}}_j^*) = \hat{\boldsymbol{v}}_j^\top \boldsymbol{g}_n^*(\hat{\boldsymbol{\beta}}^*) = \hat{\boldsymbol{v}}_j^\top \boldsymbol{g}_n^*(\boldsymbol{\beta}) + \hat{\boldsymbol{v}}_j^\top \boldsymbol{G}_n(\hat{\boldsymbol{\beta}}^* - \boldsymbol{\beta}). \tag{S7.58}$$

To control the second term in (S7.58), we note that, by Theorem 1, with probability at least $1 - p^{-c}$,

$$|\hat{\boldsymbol{v}}_j^\top \boldsymbol{G}_n(\hat{\boldsymbol{\beta}}_j^* - \boldsymbol{\beta})| \leq \|[\hat{\boldsymbol{v}}_j^\top \boldsymbol{G}_n]_{-j}\|_\infty \|\hat{\boldsymbol{\beta}}_j^* - \boldsymbol{\beta}\|_1 \lesssim s\lambda'\lambda.$$

Thus, (S6.17) is proved if $s\lambda'\lambda\sqrt{n} \to 0$, or, under the specific choices of the tuning parameters $\lambda \asymp \sqrt{\log p/n} + s\sqrt{\log p/N} \asymp \sqrt{\log p/n}$ and $\lambda' \asymp \sqrt{\log p/n}$, if $s \ll \frac{\sqrt{n}}{\log p}$.

S7.8 Proof of Lemma 7

By definition

$$\begin{aligned}\hat{S}_j(\boldsymbol{\beta}) &= \sum_{k\in\mathcal{G}(r),1\le r\le R}\frac{1}{n}\sum_{i=1}^{n}\hat{\boldsymbol{v}}_{rk}^{\top}\hat{h}_{irk}^{*}(\boldsymbol{\beta})\\ &= \frac{1}{n}\sum_{i=1}^{n}\sum_{k\in\mathcal{G}(r),1\le r\le R}\frac{n}{n_r}I\{\xi_i=r\}[(\boldsymbol{X}_i-\widehat{\boldsymbol{X}}_i^{(k)})^{\top}\boldsymbol{\beta}+\epsilon_i]\hat{\boldsymbol{v}}_{rk}^{\top}\boldsymbol{X}_{ia(r,k)}\\ &= \frac{1}{n}\sum_{i=1}^{n}\sum_{k\in\mathcal{G}(r),1\le r\le R}\frac{n}{n_r}I\{\xi_i=r\}[\boldsymbol{\beta}_{a(r)^c}^{\top}[(\boldsymbol{\Gamma}_{r,k}-\hat{\boldsymbol{\Gamma}}_{r,k})^{\top}\boldsymbol{X}_{ia(r,k)}+\boldsymbol{\epsilon}_{ia(r)^c}^{(k)}]+\epsilon_i]\hat{\boldsymbol{v}}_{rk}^{\top}\boldsymbol{X}_{ia(r,k)}\\ &= \frac{1}{n}\sum_{i=1}^{n}\sum_{k\in\mathcal{G}(r),1\le r\le R}\frac{n}{n_r}I\{\xi_i=r\}[\boldsymbol{\beta}_{a(r)^c}^{\top}(\boldsymbol{\Gamma}_{r,k}-\hat{\boldsymbol{\Gamma}}_{r,k})^{\top}\boldsymbol{X}_{ia(r,k)}+\boldsymbol{\beta}_{a(r)^c}^{\top}\boldsymbol{\epsilon}_{ia(r)^c}^{(k)}+\epsilon_i]\hat{\boldsymbol{v}}_{rk}^{\top}\boldsymbol{X}_{ia(r,k)}\end{aligned}$$

Let

$$\eta_i = \sum_{k\in\mathcal{G}(r),1\le r\le R}\frac{n}{n_r}I\{\xi_i=r\}[\boldsymbol{\beta}_{a(r)^c}^{\top}(\boldsymbol{\Gamma}_{r,k}-\hat{\boldsymbol{\Gamma}}_{r,k})^{\top}\boldsymbol{X}_{ia(r,k)}+\boldsymbol{\beta}_{a(r)^c}^{\top}\boldsymbol{\epsilon}_{ia(r)^c}^{(k)}+\epsilon_i]\hat{\boldsymbol{v}}_{rk}^{\top}\boldsymbol{X}_{ia(r,k)}.$$

We have

$$\mathbb{E}[\eta_i|\widehat{X}] = \sum_{k\in\mathcal{G}(r),1\le r\le R}\frac{n}{n_r}I\{\xi_i=r\}\boldsymbol{\beta}_{a(r)^c}^{\top}(\boldsymbol{\Gamma}_{r,k}-\hat{\boldsymbol{\Gamma}}_{r,k})^{\top}\boldsymbol{X}_{ia(r,k)}\hat{\boldsymbol{v}}_{rk}^{\top}\boldsymbol{X}_{ia(r,k)}\equiv\mu_{ji},$$

and

$$\mathrm{Var}(\eta_i|\widehat{X}) = \sum_{k\in\mathcal{G}(r),1\le r\le R}\frac{n^2\sigma_{r,k}^2}{n_r^2}I\{\xi_i=r\}[\hat{\boldsymbol{v}}_{rk}^{\top}\boldsymbol{X}_{ia(r,k)}]^2, \tag{S7.59}$$

where $\sigma_{r,k}^2=\sigma^2+\boldsymbol{\beta}_{a(r)^c}^{\top}\mathbb{E}[\boldsymbol{\epsilon}_{ia(r)^c}^{(k)}(\boldsymbol{\epsilon}_{ia(r)^c}^{(k)})^{\top}]\boldsymbol{\beta}_{a(r)^c}$. By the central limit theorem, we have the desired result.

S8 Discussion of the Eigenvalue Condition

Note that

$$\boldsymbol{X}_i^{(k)}=((\boldsymbol{X}_{ia(r)}^{(k)})^{\top},(\boldsymbol{X}_{ia(r)^c}^{(k)})^{\top})^{\top}=(\boldsymbol{X}_{ia(r)}^{\top},\boldsymbol{X}_{ia(r,k)}^{\top}\boldsymbol{\Gamma}_{r,k})^{\top}=\begin{bmatrix}\mathbf{I}_{|a(r)|}\\ [\boldsymbol{\Gamma}_{r,k}^{\top},0]\end{bmatrix}\boldsymbol{X}_{ia(r)}. \tag{S8.60}$$

It follows that

$$\begin{aligned}
\boldsymbol{\Sigma}^{(r,k)} &= \mathbb{E} I\{\xi_i = r\} \boldsymbol{X}_i^{(k)} (\boldsymbol{X}_i^{(k)})^\top \\
&= \mathbb{E} I\{\xi_i = r\} \begin{bmatrix} \mathbf{I}_{|a(r)|} \\ [\boldsymbol{\Gamma}_{r,k}^\top, \mathbf{0}] \end{bmatrix} \boldsymbol{X}_{ia(r)} \boldsymbol{X}_{ia(r)}^\top \begin{bmatrix} \mathbf{I}_{|a(r)|} & [\boldsymbol{\Gamma}_{r,k}^\top, \mathbf{0}]^\top \end{bmatrix} \\
&= \begin{bmatrix} \mathbf{I}_{|a(r)|} \\ [\boldsymbol{\Gamma}_{r,k}^\top, \mathbf{0}] \end{bmatrix} \mathbf{C}^*_{a(r),a(r)} \begin{bmatrix} \mathbf{I}_{|a(r)|} & [\boldsymbol{\Gamma}_{r,k}^\top, \mathbf{0}]^\top \end{bmatrix} \\
&= \begin{bmatrix} \mathbf{C}^*_{a(r),a(r)} & \mathbf{C}^*_{a(r),a(r,k)} \boldsymbol{\Gamma}_{r,k} \\ \boldsymbol{\Gamma}_{r,k}^\top \mathbf{C}^*_{a(r,k),a(r)} & \boldsymbol{\Gamma}_{r,k}^\top \mathbf{C}^*_{a(r,k),a(r,k)} \boldsymbol{\Gamma}_{r,k} \end{bmatrix}
\end{aligned} \tag{S8.61}$$

It can be seen that condition (S7.48) essentially reduces to that on $\mathbf{C}^* = \mathbb{E} 1\{\xi_i = r\} \boldsymbol{X}_i \boldsymbol{X}_i^\top$ and $\boldsymbol{\Gamma}_{r,k}$. To see that there exists situation where the condition (S7.48) holds, it suffices to consider the case where $\|\boldsymbol{\Gamma}_{r,k}\| \to 0$ or $\|[\boldsymbol{\Sigma}^{-1}]_{a(r)^c,a(r,k)}\| \to 0$ as $(n,p) \to \infty$, and the missing is completely at random (MCAR) with missing proportion $\mathbb{E} I\{\xi_i = r\} \geq c_0 > 0$. In this case, there exists some $c > 0$ such that

$$\lambda_{\min}(\boldsymbol{\Sigma}^{(r,k)}_{a(k),a(k)}) \geq 7c > c \geq \lambda_{\max}(\boldsymbol{\Sigma}^{(r,k)}_{a(k),m(k)}).$$

Under the Gaussian design, the above sufficient condition is more interpretable: under MCAR, we require there exists a pair $(r,k) \in [1:R] \times \mathcal{G}(r)$ such that the missing variables in $a(r)^c$ are asymptotically conditionally independent of the observed variables in $a(r,k)$.

S9 The FDR Control Procedure

To test the simultaneous hypotheses

$$H_{0j} : \beta_j = 0 \quad \text{vs} \quad H_{aj} : \beta_j \neq 0, \quad 1 \leq j \leq p,$$

we apply the modified BH procedure of Ma et al. (2020). Specifically, we reject the null hypothesis H_{0j} at significance level α if $|T_{0j}| \geq \hat{t}$, where

$$\hat{t} = \inf \left\{ 0 \leq t \leq b_p : \frac{p\{2 - 2\Phi(t)\}}{\max\{\sum_{j=1}^{p} I(|T_{0j}| \geq t), 1\}} \leq \alpha \right\}, \tag{S9.62}$$

$b_p = \sqrt{2 \log p - 2 \log \log p}$, $T_{0j} = n\tilde{\beta}_j / \hat{s}_j$, and $\Phi(t)$ is the cumulative distribution function of the standard normal distribution.

References

Bickel, P. J., Y. Ritov, and A. B. Tsybakov (2009). Simultaneous analysis of Lasso and Dantzig selector. *The Annals of Statistics 37*(4), 1705–1732.

Candes, E. and T. Tao (2007). The dantzig selector: Statistical estimation when p is much larger than n. *The Annals of Statistics 35*(6), 2313–2351.

Javanmard, A. and A. Montanari (2014). Confidence intervals and hypothesis testing for high-dimensional regression. *Journal of Machine Learning Research 15*(1), 2869–2909.

Ma, R., T. T. Cai, and H. Li (2020). Global and simultaneous hypothesis testing for high-dimensional logistic regression models. *Journal of the American Statistical Association*, 1–15.

Mendelson, S., A. Pajor, and N. Tomczak-Jaegermann (2007). Reconstruction and subgaussian operators in asymptotic geometric analysis. *Geometric and Functional Analysis 17*(4), 1248–1282.

Mendelson, S., A. Pajor, and N. Tomczak-Jaegermann (2008). Uniform uncertainty principle for Bernoulli and subgaussian ensembles. *Constructive Approximation 28*(3), 277–289.

Negahban, S., P. Ravikumar, M. J. Wainwright, and B. Yu (2010). A unified framework for high-dimensional analysis of M-estimators with decomposable regularizers. *Technical Report Number 979*.

Rubin, D. B. (1987). *Multiple imputation for nonresponse in surveys*. John Wiley & Sons, New York.

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

Zhou, S. (2009). Restricted eigenvalue conditions on subgaussian random matrices. *arXiv preprint arXiv:0912.4045*.