Statistica Sinica: Supplement

Jian Huang, Yuling Jiao, Jin Liu and Can Yang

REMI: REGRESSION WITH MARGINAL INFORMATION AND ITS APPLICATION IN GENOME-WIDE ASSOCIATION STUDIES

University of Iowa, Zhongnan University of Economics and Law,

Duke-NUS Medical School and Hong Kong University of Science and Technology

S1 Supplementary figures and table

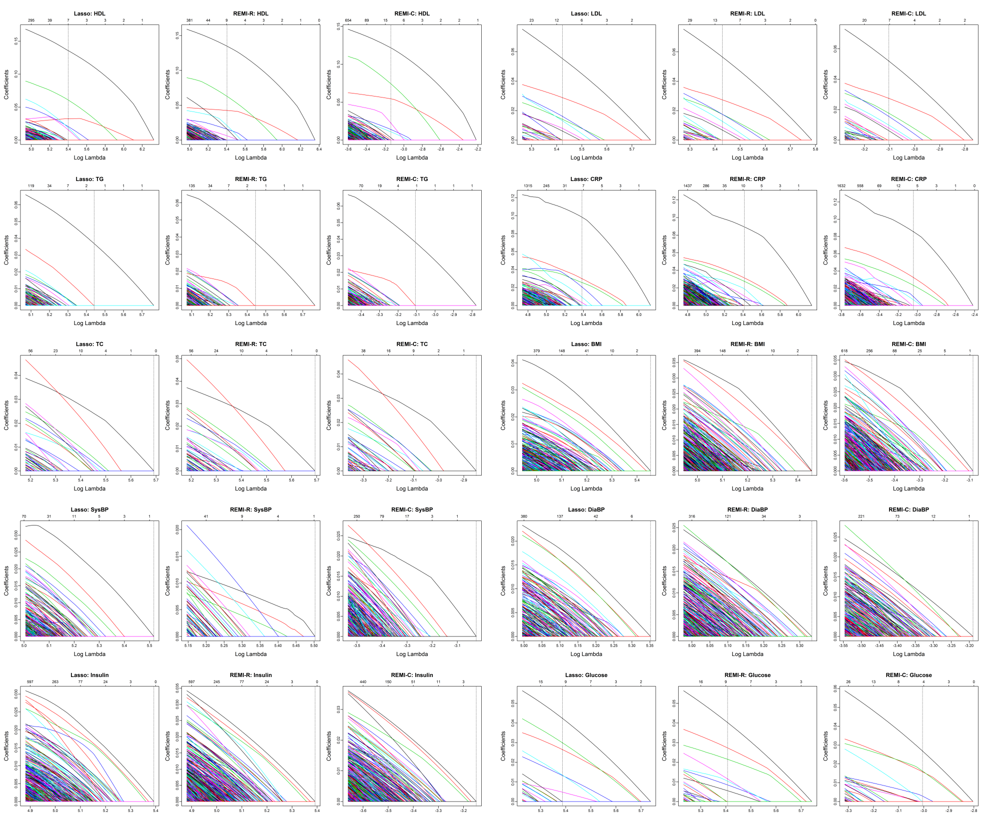

Figure S1: Solution paths of Lasso, REMI-R, and REMI-C for HDL, LDL, TG, CRP TC, BMI, SysBP, DiaBP, Insulin and Glucose using the NFBC1966 data sets.

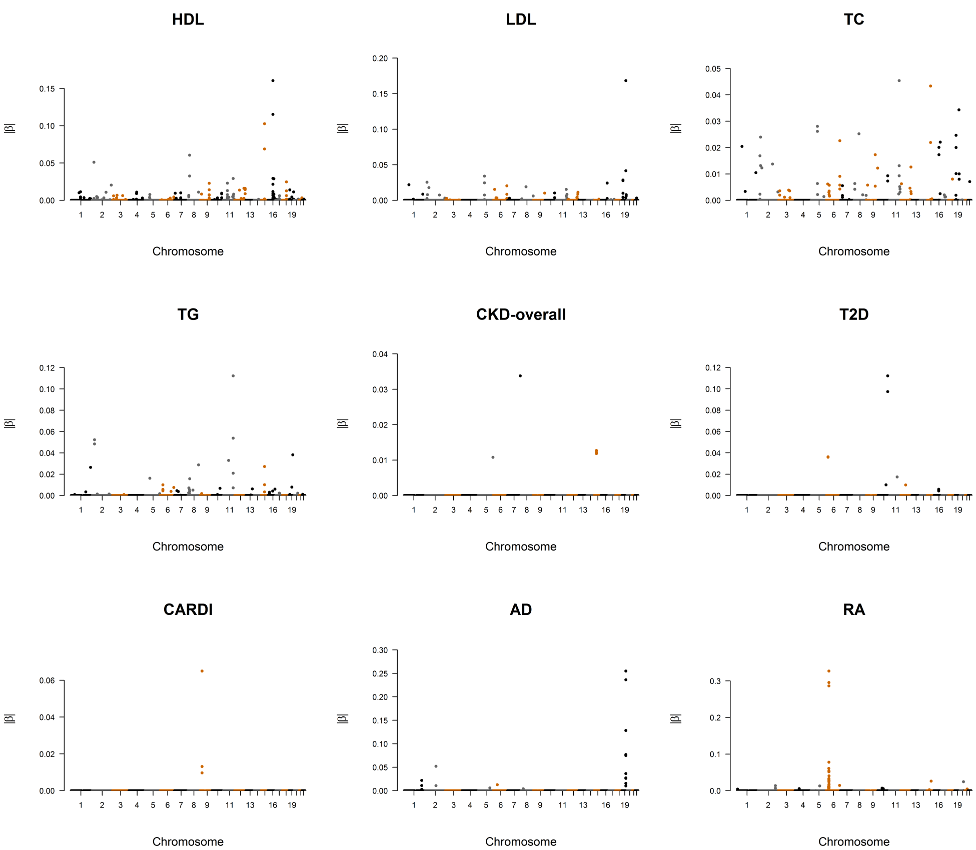

Figure S2: Manhattan plots of $|\widehat{\boldsymbol{\beta}}^{\mathrm{r}}|$ from REMI-R for HDL, LDL, TC, TG, CKD-overall, T2D, CARDI, AD and RA.

Table S1: GWAS data sets in our experiment

ID	YEAR	Traits	Sample Size	SNPs	Link
AD	2013	Alzheimer Disorder	54162	1149751	http://www.pasteur-lille.fr/en/recherche/u744/igap/igap_download.php
CARDI	2015	Coronary Artery Disease	817857	1197724	http://www.cardiogramplusc4d.org/data-downloads/
CKD-overall	2015	eGFRcrea in overall population	133715	984086	https://www.nhlbi.nih.gov/research/intramural/researchers/ckdgen
HDL	2013	High-Density-Lipid cholesterol	94272	992986	http://csg.sph.umich.edu//abecasis/public/lipids2013/
Ht	2014	Height	252778	827344	http://portals.broadinstitute.org/collaboration/giant/index.php/GIANT_consortium_data_files
LDL	2013	Low-Density-Lipid cholesterol	89851	990583	http://csg.sph.umich.edu//abecasis/public/lipids2013/
TC	2013	Total Cholesterol	94556	992889	http://csg.sph.umich.edu//abecasis/public/lipids2013/
TG	2013	Triglycerides	90974	990915	http://csg.sph.umich.edu//abecasis/public/lipids2013/
RA	2010	Rheumatoid Arthritis	25708	989551	http://www.broadinstitute.org/ftp/pub/rheumatoid_arthritis/Stahl_etal_2010NG/
T2D	2008	Type 2 Diabetes	63390	1061515	http://diagram-consortium.org/downloads.html

S2 Technical details

Lemma 1. *(Lemma 2.7.7 of Vershynin (2018) and Remark 5.18 of Vershynin (2010).) Let ξ_1, ξ_2 be sub-Gaussian random variables with noise level $\|\xi_1\|_{\psi_2} \leq \sigma_{\xi_1}$ and $\|\xi_2\|_{\psi_2} \leq \sigma_{\xi_2}$, respectively. Then both $\xi_1\xi_2$ and $\xi_1\xi_2 - \mathbb{E}[\xi_1\xi_2]$ are sub-exponential random variables, and there exist an absolute constant $C > 0$ such that $\|\xi_1\xi_2 - \mathbb{E}[\xi_1\xi_2]\|_{\psi_1} \leq C\sigma_{\xi_1}\sigma_{\xi_2}$. Here, for a random variable z we define, $\|z\|_{\psi_i} = \inf\left\{t > 0 : \mathbb{E}[\exp\left(|z|^i/t^i\right)] \leq 2\right\}, i \in \{1, 2\}$.*

Lemma 2. *(Corollary 5.17 of Vershynin (2010)) Let $\xi_1, ..., \xi_m$ be independent centered sub-exponential random variables. Then for every $t > 0$ one has*

$$\mathbb{P}[|\sum_{i=1}^{m} \xi_i|/m \geq t] \leq 2\exp(-C\min\{\frac{t^2}{K^2}, \frac{t}{K}\}m),$$

where, C is a absolute constant and $K = \max_{i=1,..m}\{\|\xi_i\|_{\psi_1}\}$.

Lemma 3. *Suppose the rows of $\mathbf{X}$ and $\mathbf{X}_{\mathrm{r}}$ are i.i.d sub-Gaussian samples drawn from population with mean $\mathbf{0}$ and covariance matrix $\mathbf{\Sigma}$. Then, with probability at least $1 - 1/p^2$, we have*

$$\|\widehat{\mathbf{\Sigma}} - \mathbf{\Sigma}\|_\infty \leq \frac{2C_1}{\sqrt{C}}\sqrt{\frac{\log p}{n}},$$

and

$$\|\widehat{\mathbf{\Sigma}}_{\mathrm{r}} - \mathbf{\Sigma}\|_\infty \leq \frac{2C_1}{\sqrt{C}}\sqrt{\frac{\log p}{n_{\mathrm{r}}}},$$

as long as $n > \frac{4}{C}\log p$ *and* $n_{\mathrm{r}} > \frac{4}{C}\log p$.

Proof of Lemma 3. Since the proof of these two results are similar, we give one of them. Let $\mathbf{x}_i$ be the $i-$th row of $\mathbf{X}$, $i = 1, ...n$, and $(\mathbf{x}_i)_j$ denote the $j-$th entry of $\mathbf{x}_i$. Define $G^i_{j,k} := (\mathbf{x}_i)_j(\mathbf{x}_i)_k - \mathbb{E}[(\mathbf{x}_i)_j(\mathbf{x}_i)_k] \in \mathcal{R}^1, i = 1, ..., n, j = 1, ..., p, k = 1, .., p$, which is sub-exponential with $\|G^i_{j,k}\|_{\psi_1} \le C_1$ by Lemma 1. Therefore,

$$\begin{aligned}
\mathbb{P}[\|\widehat{\mathbf{\Sigma}} - \mathbf{\Sigma}\|_\infty \ge t] &= \mathbb{P}[\|\sum_{i=1}^n(\mathbf{x}_i^T\mathbf{x}_i - \mathbb{E}[\mathbf{x}_i^T\mathbf{x}_i])/n\|_\infty \ge t] \\
&= \mathbb{P}[\bigcup_{j=1,k=1}^{p,p}|\sum_{i=1}^n G^i_{j,k}/n| \ge t] \\
&\le \sum_{j=1,k=1}^{p,p}\mathbb{P}[|\sum_{i=1}^n G^i_{j,k}|/n \ge t] \\
&\le p^2\exp(-C\min\{\frac{t^2}{C_1^2}, \frac{t}{C_1}\}n) \qquad \text{(S2.1)} \\
&\le p^2\exp(-C\frac{t^2}{C_1^2}n)
\end{aligned}$$

where the first inequality is due to union bound, and the second one follows from Lemma 2 and the last inequality is because of restricting $t \le C_1$. Then Lemma 3 follows from setting $t = \frac{2C_1}{\sqrt{C}}\sqrt{\frac{\log p}{n}}$ and the assumption that $n > \frac{4}{C}\log p$. □

Lemma 4. *Under the same assumption as Lemma 3, we have*

$$\|(\widehat{\mathbf{\Sigma}} - \widehat{\mathbf{\Sigma}}_{\mathrm{r}})\boldsymbol{\beta}^*_{\mathcal{A}}\|_\infty \le \frac{2C_1C_3}{\sqrt{C}}(\sqrt{\frac{\log p}{n}} + \sqrt{\frac{\log p}{n_{\mathrm{r}}}}),$$

holds with probability at least $1 - 2/p^2$.

Proof of Lemma 4.

$$\begin{aligned}
\|(\widehat{\mathbf{\Sigma}} - \widehat{\mathbf{\Sigma}}_{\mathrm{r}})\boldsymbol{\beta}^*_{\mathcal{A}}\|_\infty &\leq \|(\widehat{\mathbf{\Sigma}} - \mathbf{\Sigma})\boldsymbol{\beta}^*_{\mathcal{A}}\|_\infty + \|(\mathbf{\Sigma} - \widehat{\mathbf{\Sigma}}_{\mathrm{r}})\boldsymbol{\beta}^*_{\mathcal{A}}\|_\infty \\
&\leq \|\widehat{\mathbf{\Sigma}} - \mathbf{\Sigma}\|_\infty \|\boldsymbol{\beta}^*_{\mathcal{A}}\|_1 + \|\mathbf{\Sigma} - \widehat{\mathbf{\Sigma}}_{\mathrm{r}}\|_\infty \|\boldsymbol{\beta}^*_{\mathcal{A}}\|_1 \\
&\leq \frac{2C_1C_3}{\sqrt{C}}\sqrt{\frac{\log p}{n}} + \frac{2C_1C_3}{\sqrt{C}}\sqrt{\frac{\log p}{n_{\mathrm{r}}}},
\end{aligned}$$

where the first inequality is due to triangle inequality, and the second inequality follows from Cauchy-Schwartz inequality, and the last one holds with probability larger than $1 - 2/p^2$ due to Lemma 3. This finishes the proof of Lemma 4. $\square$

Lemma 5. *Suppose the rows of* $\mathbf{X}$ *are i.i.d sub-Gaussian samples drawn from population with mean* $\mathbf{0}$ *and covariance matrix* $\mathbf{\Sigma}$*, and the entries of noise* $\boldsymbol{\epsilon}$ *are i.i.d centered sub-Gaussian with noise level* σ_ϵ*. With probability at least* $1 - 1/p^3$*, we have*

$$\|\widetilde{\boldsymbol{\epsilon}}\|_\infty < 2\sigma_\epsilon \frac{C_1}{\sqrt{C}}\sqrt{\frac{\log p}{n}},$$

provided that $n \geq \frac{4\log p}{C}$.

Proof of Lemma 5. We have,

$$\begin{aligned}
&\mathbb{P}[\|\widetilde{\boldsymbol{\epsilon}}\|_\infty < t] = \mathbb{P}[\|\mathbf{X}^T\boldsymbol{\epsilon}/n\|_\infty < t] \\
&= 1 - \mathbb{P}[\|\mathbf{X}^T\boldsymbol{\epsilon}/n\|_\infty \geq t] \\
&= 1 - \mathbb{P}[\bigcup_{j=1}^{p} |\mathbf{X}_j^T\boldsymbol{\epsilon}/n| \geq t] \\
&\geq 1 - \sum_{j=1}^{p} \mathbb{P}[|\sum_{i=1}^{n}(\mathbf{X}_j)_i\epsilon_i|/n \geq t] \\
&\geq 1 - p\exp(-C\min\{\frac{t^2}{C_1^2\sigma_{\boldsymbol{\epsilon}}^2}, \frac{t}{C_1\sigma}\}n) \\
&\geq 1 - p\exp(-C\frac{t^2}{C_1^2\sigma_\epsilon^2}n) \\
&\geq 1 - 1/p^3, \qquad \text{(S2.2)}
\end{aligned}$$

the first inequality is due to union bound, and the second one follows from Lemma 1 and Lemma 2, where we use $\|(\mathbf{X}_j)_i\epsilon_i - \mathbb{E}[(\mathbf{X}_j)_i\epsilon_i]\|_{\psi_1} \leq C\sigma_\epsilon C_1$, and the last two inequality follows from by setting $t = 2\sigma_\epsilon C_1\sqrt{\frac{\log p}{Cn}}$ and the assumption that $n > \frac{4}{C}\log p$, i.e., with probability at least $1-1/p^3$, we have

$$\|\widetilde{\boldsymbol{\epsilon}}\|_\infty \leq 2\sigma_\epsilon C_1\sqrt{\frac{\log p}{Cn}}.$$

□

Lemma 6. *Under the same assumption as Lemma 5, we have,*

$$\|\widehat{\boldsymbol{\Sigma}}\boldsymbol{\beta}_{\mathcal{I}}^*\|_\infty \leq \frac{2C_1C_4}{\sqrt{C}}\sqrt{\frac{\log p}{n}} + 2C_2\sigma_\epsilon\sqrt{\frac{\log p}{n}}.$$

with probability larger than $1 - 1/p^2$.

Proof of Lemma 6.

$$\begin{aligned}
&\|\widehat{\mathbf{\Sigma}}\boldsymbol{\beta}^*_{\mathcal{I}}\|_\infty \leq \|\mathbf{\Sigma}\boldsymbol{\beta}^*_{\mathcal{I}}\|_\infty + \|(\widehat{\mathbf{\Sigma}} - \mathbf{\Sigma})\boldsymbol{\beta}^*_{\mathcal{I}}\|_\infty \\
&\leq \max_{j=1,\dots p}\{\|\mathbf{\Sigma}_j\|_1\}\|\boldsymbol{\beta}^*_{\mathcal{I}}\|_\infty + \|\widehat{\mathbf{\Sigma}} - \mathbf{\Sigma}\|_\infty\|\boldsymbol{\beta}^*_{\mathcal{I}}\|_1 \\
&\leq 2C_2\sigma_\epsilon\sqrt{\frac{\log p}{n}} + \frac{2C_1C_4}{\sqrt{C}}\sqrt{\frac{\log p}{n}},
\end{aligned}$$

where first inequality is due to triangle inequality, and the second one follows from Cauchy-Schwartz inequality, the third inequality holds with probability larger than $1 - 1/p^2$ by using (4.7) and Lemma 3. This completes the proof of Lemma 6. □

Now we are ready to prove Theorem 1.

Proof of Theorem 1. (i). Let $\Delta = \widehat{\boldsymbol{\beta}}^{\mathrm{c}} - \boldsymbol{\beta}^*_{\mathcal{A}}$. Define the event

$$\mathcal{E} = \{2\|(\widehat{\mathbf{\Sigma}} - \widehat{\mathbf{\Sigma}_r})\boldsymbol{\beta}^*_{\mathcal{A}} + \widehat{\mathbf{\Sigma}}\boldsymbol{\beta}^*_{\mathcal{I}} + \widetilde{\boldsymbol{\epsilon}}\|_\infty \leq \lambda_0\}.$$

The optimality of $\widehat{\boldsymbol{\beta}}^{\mathrm{c}}$ implies that

$$\langle\widehat{\boldsymbol{\beta}}^{\mathrm{c}}, \widehat{\boldsymbol{\Sigma}}_{\mathrm{r}}\widehat{\boldsymbol{\beta}}^{\mathrm{c}}\rangle - 2\langle\widetilde{\mathbf{y}}, \widehat{\boldsymbol{\beta}}^{\mathrm{c}}\rangle + \lambda\|\widehat{\boldsymbol{\beta}}^{\mathrm{c}}\|_1 \leq \langle\boldsymbol{\beta}^*_{\mathcal{A}}, \widehat{\boldsymbol{\beta}}^{\mathrm{c}}\boldsymbol{\beta}^*_{\mathcal{A}}\rangle - 2\langle\widetilde{\mathbf{y}}, \boldsymbol{\beta}^*_{\mathcal{A}}\rangle + \lambda\|\boldsymbol{\beta}^*_{\mathcal{A}}\|_1,$$

$$\Downarrow \text{(eq1)}$$

$$\langle\widehat{\boldsymbol{\beta}}^{\mathrm{c}} - \boldsymbol{\beta}^*_{\mathcal{A}}, \widehat{\boldsymbol{\Sigma}}_{\mathrm{r}}(\widehat{\boldsymbol{\beta}}^{\mathrm{c}} - \boldsymbol{\beta}^*_{\mathcal{A}})\rangle + 2\langle\boldsymbol{\beta}^*_{\mathcal{A}}, \widehat{\boldsymbol{\Sigma}}_{\mathrm{r}}(\widehat{\boldsymbol{\beta}}^{\mathrm{c}} - \boldsymbol{\beta}^*_{\mathcal{A}})\rangle + \lambda(\|\widehat{\boldsymbol{\beta}}^{\mathrm{c}}{}_{\mathcal{A}}\|_1 + \|\widehat{\boldsymbol{\beta}}^{\mathrm{c}}{}_{\mathcal{I}}\|_1) \leq 2\langle\widetilde{\mathbf{y}}, \widehat{\boldsymbol{\beta}}^{\mathrm{c}} - \boldsymbol{\beta}^*_{\mathcal{A}}\rangle + \lambda\|\boldsymbol{\beta}^*_{\mathcal{A}}\|,$$

$$\Downarrow \text{(eq2)}$$

$$\langle\Delta, \widehat{\boldsymbol{\Sigma}}_{\mathrm{r}}\Delta\rangle + \lambda\|\Delta_{\mathcal{I}}\|_1 \leq 2\langle\widetilde{\mathbf{y}}, \Delta\rangle - 2\langle\widehat{\boldsymbol{\Sigma}}_{\mathrm{r}}\boldsymbol{\beta}^*_{\mathcal{A}}, \Delta\rangle + \lambda\|\boldsymbol{\beta}^*_{\mathcal{A}}\| - \lambda\|\widehat{\boldsymbol{\beta}}^{\mathrm{c}}{}_{\mathcal{A}}\|_1,$$

$$\Downarrow \text{(eq3)}$$

$$\langle\Delta, \widehat{\boldsymbol{\Sigma}}_{\mathrm{r}}\Delta\rangle + \lambda\|\Delta_{\mathcal{I}}\|_1 \leq 2\langle(\widehat{\boldsymbol{\Sigma}} - \widehat{\boldsymbol{\Sigma}}_{\mathrm{r}})\boldsymbol{\beta}^*_{\mathcal{A}} + \widehat{\boldsymbol{\Sigma}}\boldsymbol{\beta}^*_{\mathcal{I}} + \widetilde{\boldsymbol{\epsilon}}, \Delta\rangle + \lambda\|\Delta_{\mathcal{A}}\|,$$

$$\Downarrow \text{(eq4)}$$

$$\langle\Delta, \widehat{\boldsymbol{\Sigma}}_{\mathrm{r}}\Delta\rangle + \lambda\|\Delta_{\mathcal{I}}\|_1 \leq 2\|(\widehat{\boldsymbol{\Sigma}} - \widehat{\boldsymbol{\Sigma}}_{\mathrm{r}})\boldsymbol{\beta}^*_{\mathcal{A}} + \widehat{\boldsymbol{\Sigma}}\boldsymbol{\beta}^*_{\mathcal{I}} + \widetilde{\boldsymbol{\epsilon}}\|_\infty\|\Delta\|_1 + \lambda\|\Delta_{\mathcal{A}}\|,$$

$$\Downarrow \text{(eq5)}$$

$$\langle\Delta, \widehat{\boldsymbol{\Sigma}}_{\mathrm{r}}\Delta\rangle + \lambda\|\Delta_{\mathcal{I}}\|_1 \leq \frac{\lambda}{2}(\|\Delta_{\mathcal{A}}\|_1 + \|\Delta_{\mathcal{I}}\|_1) + \lambda\|\Delta_{\mathcal{A}}\|,$$

$$\Downarrow \text{(eq6)}$$

$$\langle\Delta, \widehat{\boldsymbol{\Sigma}}_{\mathrm{r}}\Delta\rangle + \frac{\lambda}{2}\|\Delta_{\mathcal{I}}\|_1 \leq \frac{3}{2}\lambda\|\Delta_{\mathcal{A}}\|_1, \tag{S2.3}$$

where, (eq1) and (eq2) and (eq6) are due to some algebra, and (eq3) follows from (4.8), and (eq4) uses Cauchy-Schwartz inequality, and (eq5) holds by conditioning $\mathcal{E}$ and the assumption $\lambda_0 \leq \lambda/2$. It follow from (S2.3) that

$$\Delta \in \mathcal{C}_{\mathcal{A},3}. \tag{S2.4}$$

Then, by the restricted eigenvalue condition on $\widehat{\mathbf{\Sigma}}_{\mathrm{r}}$ and (S2.3) we deduce,

$$\phi_0\|\Delta\|_2^2 \leq \langle \Delta, \widehat{\mathbf{\Sigma}}_{\mathrm{r}}\Delta\rangle \leq \frac{3}{2}\lambda\|\Delta_{\mathcal{A}}\|_1 \leq \frac{3}{2}\sqrt{s}\lambda\|\Delta\|_2,$$

i.e.,

$$\|\Delta\|_2 \leq \frac{3}{2\phi_0}\sqrt{s}\lambda \leq \frac{6}{\phi_0}(\frac{C_1(C_3+C_4)}{\sqrt{C}}\sqrt{\frac{s\log p}{n}}+\frac{C_1+\sqrt{C}C_2}{\sqrt{C}}\sigma_\epsilon\sqrt{\frac{s\log p}{n}}+\frac{C_1C_3}{\sqrt{C}}\sqrt{\frac{s\log p}{n_{\mathrm{r}}}}).$$

The above induction is conditioning on $\mathcal{E}$. We need give a lower bound on $\mathbb{P}[\mathcal{E}]$. Indeed,

$$2\|(\widehat{\mathbf{\Sigma}}-\widehat{\mathbf{\Sigma}}_{\mathrm{r}})\boldsymbol{\beta}^*_{\mathcal{A}}+\widehat{\mathbf{\Sigma}}\boldsymbol{\beta}^*_{\mathcal{I}}+\tilde{\boldsymbol{\epsilon}}\|_\infty \leq 2\|(\widehat{\mathbf{\Sigma}}-\widehat{\mathbf{\Sigma}}_{\mathrm{r}})\boldsymbol{\beta}^*_{\mathcal{A}}\|_\infty+2\|\widehat{\mathbf{\Sigma}}\boldsymbol{\beta}^*_{\mathcal{I}}\|_\infty+2\|\tilde{\boldsymbol{\epsilon}}\|_\infty$$

Then, it follows Lemma 4 and Lemma 5 and Lemma 6 that $\mathbb{P}[\mathcal{E}] \geq 1 - 3/p^2 - 1/p^3$. This completes the proof of (i) of Theorem 1.

(ii). Let $\widehat{\mathbf{\Sigma}}_{\mathrm{new}} = \mathbf{X}^T_{\mathrm{new}}\mathbf{X}_{\mathrm{new}}/n_{\mathrm{new}}$. Then,

$$\begin{aligned}
&\|\mathbf{X}_{\mathrm{new}}(\widehat{\boldsymbol{\beta}}^{\mathrm{c}}-\boldsymbol{\beta}^*_{\mathcal{A}})\|_2^2/n_{\mathrm{new}} = \langle\widehat{\mathbf{\Sigma}}_{\mathrm{new}}(\widehat{\boldsymbol{\beta}}^{\mathrm{c}}-\boldsymbol{\beta}^*_{\mathcal{A}}),\widehat{\boldsymbol{\beta}}^{\mathrm{c}}-\boldsymbol{\beta}^*_{\mathcal{A}}\rangle\\
&= \langle\Delta,\widehat{\mathbf{\Sigma}}_{\mathrm{r}}\Delta\rangle+\langle\Delta,(\widehat{\mathbf{\Sigma}}_{\mathrm{new}}-\widehat{\mathbf{\Sigma}}_{\mathrm{r}})\Delta\rangle\\
&\leq \frac{3}{2}\lambda\|\Delta_{\mathcal{A}}\|_1+\|\Delta\|_1^2\|\widehat{\mathbf{\Sigma}}_{\mathrm{new}}-\widehat{\mathbf{\Sigma}}_{\mathrm{r}}\|_\infty\\
&\leq \frac{3}{2}\lambda\|\Delta_{\mathcal{A}}\|_1+\|\Delta\|_1^2\frac{2C_1}{\sqrt{C}}(\sqrt{\frac{\log p}{n_{\mathrm{r}}}}+\sqrt{\frac{\log p}{n_{\mathrm{new}}}})\\
&\leq \mathcal{O}(\sigma_\epsilon\sqrt{\frac{s\log p}{n}}+\sqrt{\frac{s\log p}{n_{\mathrm{r}}}})\|\Delta_{\mathcal{A}}\|_2+\mathcal{O}(\sqrt{\frac{\log p}{n_{\mathrm{r}}}}+\sqrt{\frac{\log p}{n_{\mathrm{new}}}})s^2\|\Delta_{\mathcal{A}}\|_2^2\\
&\leq \mathcal{O}((\sigma_\epsilon\sqrt{\frac{s\log p}{n}}+\sqrt{\frac{s\log p}{n_{\mathrm{r}}}})^2)(1+s^2(\sqrt{\frac{\log p}{n_{\mathrm{r}}}}+\sqrt{\frac{\log p}{n_{\mathrm{new}}}}))
\end{aligned}$$

where, the first inequality uses (S2.3) and Cauchy-Schwartz inequality, and the second one is due to Lemma 3, and the third inequality follow from (S2.4) and Cauchy-Schwartz inequality, the fourth inequality uses Theorem 1. □

References

Bibliography

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

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