Variable screening with multiple studies

Tianzhou Ma[1], Zhao Ren[2] and George C. Tseng[2]

[1]*University of Maryland,* [2]*University of Pittsburgh*

Supplementary Material

S1. Proof of Lemma 1

Proof. Part (i) immediately follows from Lemma 2 (i) equation (25) in Cai and Liu (2011). To prove part (ii), we need to bound the three terms on the right side of the following inequality,

$$\max_{j,k} |\hat{\theta}_j^{(k)} - \theta_j^{(k)}| \le \max_{j,k} |\hat{\theta}_j^{(k)} - \tilde{\theta}_j^{(k)}| + \max_{j,k} |\tilde{\theta}_j^{(k)} - \check{\theta}_j^{(k)}| + \max_{j,k} |\check{\theta}_j^{(k)} - \theta_j^{(k)}|, \tag{S1.1}$$

where $\tilde{\theta}_j^{(k)} := \frac{1}{n}\sum_{i=1}^{n}(X_{ij}^{(k)}Y_i^{(k)} - \tilde{\rho}_j^{(k)})^2$ with $\tilde{\rho}_j^{(k)} = \frac{1}{n}\sum_{i=1}^{n} X_{ij}^{(k)}Y_i^{(k)}$, and $\check{\theta}_j^{(k)} := \frac{1}{n}\sum_{i=1}^{n}(X_{ij}^{(k)}Y_i^{(k)} - \rho_j^{(k)})^2$. Note that $E(\check{\theta}_j^{(k)}) = \theta_j^{(k)}$.

By the marginal sub-Gaussian distribution assumption in assumption (C1), we have that $(X_{ij}^{(k)}Y_i^{(k)} - \rho_j^{(k)})^2$ has mean $\theta_j^{(k)}$ and finite Orlicz $\psi_{1/2}$-norm (see, e.g., Adamczak et al. (2011)). Thus we can apply equation (3.6) of Adamczak et al. (2011), i.e.,

$$P(\max_{j,k} \sqrt{n}|\check{\theta}_j^{(k)} - \theta_j^{(k)}| > t) \le 2Kp\exp(-c\min[\frac{t^2}{n}, t^{1/2}]),$$

with $t = (C_\theta/3)\sqrt{n\log p}$ for a large enough constant $C_\theta > 0$ depending on M_1, η, b and M only to obtain that,

$$P(\max_{j,k} |\check{\theta}_j^{(k)} - \theta_j^{(k)}| > (C_\theta/3)\sqrt{\frac{\log p}{n}}) = O(p^{-M}). \tag{S1.2}$$

We have used the assumption $\log p = o(n^{1/3})$ in assumption (C2) to make sure $\frac{t^2}{n} \le t^{1/2}$.

By applying equation (1) in supplement of Cai and Liu (2011), we obtain that,

$$P(\max_{j,k} |\tilde{\theta}_j^{(k)} - \hat{\theta}_j^{(k)}| > (C_\theta/3)\sqrt{\frac{\log p}{n}}) = O(p^{-M}). \tag{S1.3}$$

In addition, by a similar truncation argument as that in the proof of Lemma 2 in Cai and Liu (2011) and equation (7) therein, we obtain that by picking a large enough $C_\theta > 0$,

$$P(\max_{j,k} |\tilde{\theta}_j^{(k)} - \check{\theta}_j^{(k)}| > (C_\theta/3)\sqrt{\frac{\log p}{n}}) = O(p^{-M}). \tag{S1.4}$$

We complete the proof by combining (S1.1)-(S1.4) with a union bound argument.

□

S2. Proof of Lemma 2

Proof. It is easy to check that $E(\check{H}_j^{(k)}) = 0$ and $\mathrm{var}(\check{H}_j^{(k)}) = 1$. The marginal sub-Gaussian distribution assumption in assumption (C1) implies that $\check{H}_j^{(k)}$ has finite Orlicz ψ_1-norm (i.e., sub-exponential distribution with finite constants). Therefore, $(\check{H}_j^{(k)})^2 - 1$ is centered random variable with finite Orlicz $\psi_{1/2}$-norm. Note that $\check{H}_j^{(k)}$ are independent for $k \in [K]$. The result follows from equation (3.6) of Adamczak et al. (2011). □

S3. Proof of Lemma 3

Proof. Note that $\check{H}_j^{(k)} - \hat{H}_j^{(k)} = \frac{\sqrt{n}\bar{X}_j^{(k)}\sqrt{n}\bar{Y}^{(k)}}{\sqrt{n\theta_j^{(k)}}}$. By assumption (C1), we have that $E(\sqrt{n}\bar{X}_j^{(k)}) = E(\sqrt{n}\bar{Y}^{(k)}) = 0$, $\mathrm{var}(\sqrt{n}\bar{X}_j^{(k)}) = \mathrm{var}(\sqrt{n}\bar{Y}^{(k)}) = 1$, and both $\sqrt{n}\bar{X}_j^{(k)}$ and $\sqrt{n}\bar{Y}^{(k)}$ are sub-Gaussian with bounded constants. Therefore, the first equation follows from Bernstein inequality (e.g., Definition 5.13 in Vershynin (2010)) applied to centered sub-exponential variable $\sqrt{n}\bar{X}_j^{(k)} \cdot \sqrt{n}\bar{Y}^{(k)}$, noting $\theta_j^{(k)} \geq \tau_0$ by assumption (C1). The second equation follows from the first one, $\log^3 p = o(n)$, and a Bernstein inequality (e.g., Corollary 5.17 in Vershynin (2010)) applied to the sum of centered sub-exponential variables $\check{H}_j^{(k)}$. □

Bibliography

Adamczak, R., Litvak, A. E., Pajor, A., and Tomczak-Jaegermann, N. (2011). Restricted isometry property of matrices with independent columns and neighborly polytopes by random sampling. *Constructive Approximation*, 34(1):61–88.

Cai, T. and Liu, W. (2011). Adaptive thresholding for sparse covariance matrix estimation. *Journal of the American Statistical Association*, 106(494):672–684.

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

Tianzhou Ma

Department of Epidemiology and Biostatistics, University of Maryland, College Park, MD 20742.

E-mail: tma0929@umd.edu

Zhao Ren (Corresponding Author)

Department of Statistics, University of Pittsburgh, Pittsburgh, PA 15261.

E-mail: zren@pitt.edu

George C. Tseng

Department of Biostatistics, University of Pittsburgh, Pittsburgh, PA 15261.

E-mail: ctseng@pitt.edu