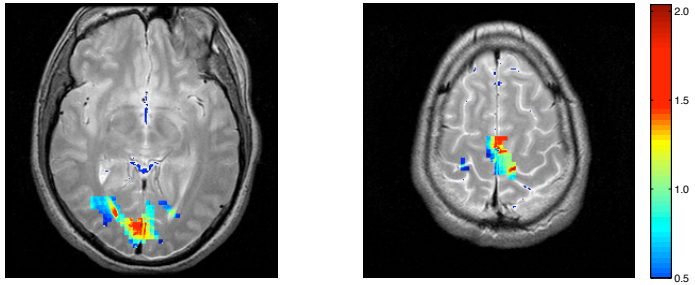
Estimation and Inference in Large Graphical Models

Cun-Hui Zhang, Rutgers University

April 25, 2014

20th Applied Probability Day in Honor/Memory of Larry Shepp
Columbia University

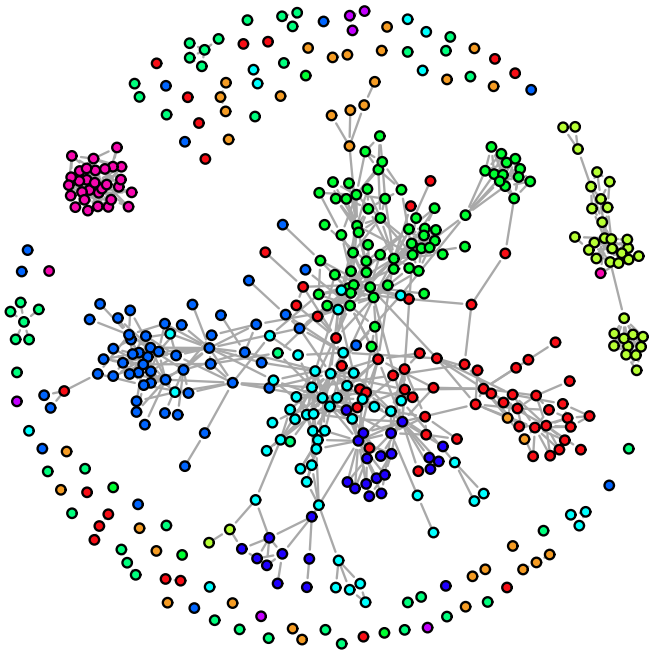


Pick up sticks

Larry Shepp, *Doron* Zeilberger, *Cun-Hui* Zhang

October 23, 2012

Large graphical models:

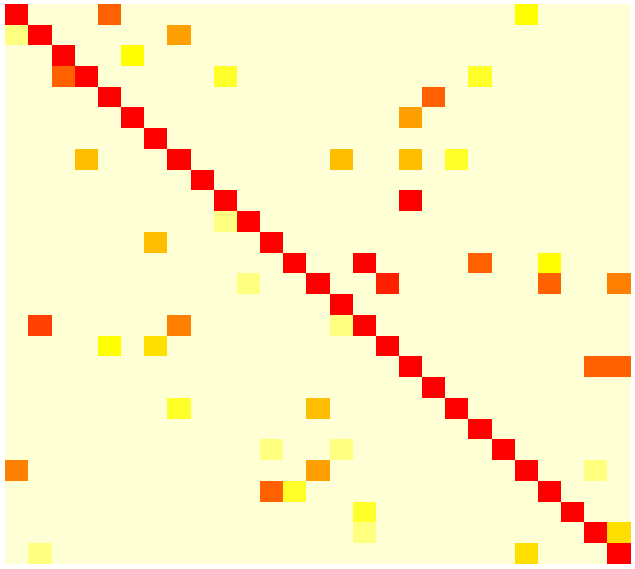


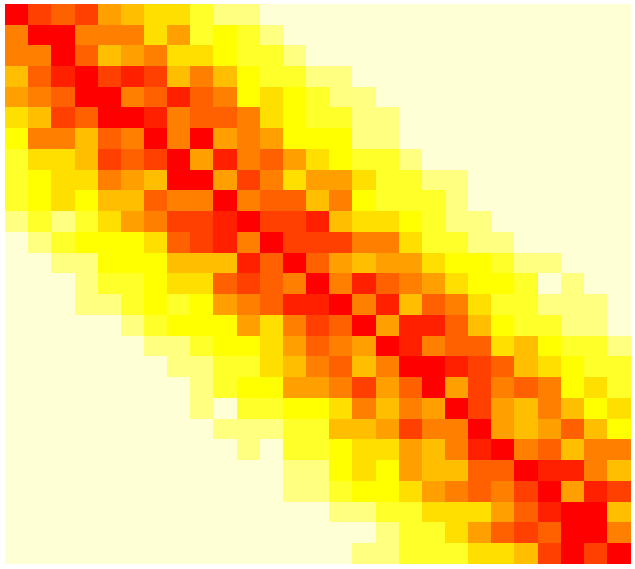
- Variables: $X_1, \ldots, X_p$
- Vertices represent variables
- Edges represent association
- Data: $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_p$ in $\mathbb{R}^n$
- Copulas: $f_j(\boldsymbol{x}_j)$, unknown f_j
- $p \gg n$

Acknowledgements:

- Sun, T. and C. H. Zhang (2013). Sparse matrix inversion with scaled lasso. *Journal of Machine Learning Research 14*, 3385–3418.
- Sun, T. and Zhang, C.-H. (2012). Comments on: Optimal rates of convergence for sparse covariance matrix estimation. *Statistica Sinica 22*, 1354–1358.
- Ren, Z., Sun, T., Zhang, C.-H. and Zhou, H.H. (2013). Asymptotic normality and optimalities in estimation of large Gaussian graphical model, preprint.
- Liu, H., Han, F. and Zhang, C.-H. (2012). Transelliptical graphical modeling under a hierarchical latent variable framework.
- Mitra, R. and Zhang, C.-H. (2014). Multivariate analysis of nonparametric estimates of large correlation matrices. arXiv:1403.6195

Examples of sparsity in $p \gg n$ models





Measurements of association:

- Covariance/correlation
  - Parameters: $\Sigma_{jk} = \mathrm{Cov}(X_j, X_k)$, $\boldsymbol{\Sigma} = (\Sigma_{jk})_{p\times p}$
  - Raw estimator: $\overline{\boldsymbol{\Sigma}} = (\boldsymbol{x}_j^\top \boldsymbol{x}_k / n)_{p\times p}$, $\mathrm{rank}(\overline{\boldsymbol{\Sigma}}) \le n \ll p$
  - Estimation of $\boldsymbol{\Sigma}$ when $\boldsymbol{\Sigma}$ is sparse: thresholding/tapering $\overline{\boldsymbol{\Sigma}}$
  - Matrix denoising problem
- Partial correlation (Markov graphical models)
  - $\mathrm{Corr}(X_j, X_k | X_\ell \notin \{j,k\})$ or $\Theta_{jk} / \sqrt{\Theta_{jj}\Theta_{kk}}$, $\boldsymbol{\Theta} = \boldsymbol{\Sigma}^{-1}$
  - Estimation of $\boldsymbol{\Theta}$ when $\boldsymbol{\Theta}$ is sparse
  - Inversion of an ill-conditioned, noisy matrix $\overline{\boldsymbol{\Sigma}}$
  - Confidence intervals/hypothesis testing for Θ_{jk} and $\Theta_{jk} / \sqrt{\Theta_{jj}\Theta_{kk}}$
- Gaussian and elliptical copulas
  - $Y_j = f_j(X_j)$ with unknown f_j, $j \in [p]$
  - $(X_1, \ldots, X_p) \sim N(0, \boldsymbol{\Sigma})$ or $\boldsymbol{\Sigma}^{1/2}\boldsymbol{U}\xi$ with $\xi \perp\!\!\!\perp \boldsymbol{U}$ & $\boldsymbol{U} \sim \mathrm{unif}(\mathbb{S}^{p-1})$
  - Data: $\boldsymbol{y}_j = f_j(\boldsymbol{x}_j)$, ranks are sufficient
  - Raw estimator: Spearman's rank correlation
  - Raw estimator: Kendall's tau, estimation of $\mathrm{sgn}(Y_j - \widetilde{Y}_j)\mathrm{sgn}(Y_k - \widetilde{Y}_k)$
  - Error bounds for matrices of nonparametric estimates

Estimation of $\mathbf{\Theta} = \mathbf{\Sigma}^{-1}$

- Graphical Lasso/SCAD (Yuan-Lin 07, Friedman et al 08, Ravikuma et al 08, Rothman et al 08, Lam-Fan 09)
  - Minimize $\langle \mathbf{\Theta}, \overline{\mathbf{\Sigma}} \rangle + \log\det(\mathbf{\Theta}) + \lambda \sum_{j \neq k} \mathrm{pen}(|\Theta_{jk}|)$
  - $\|\widehat{\mathbf{\Theta}} - \mathbf{\Theta}\|_F^2 \lesssim \big(\#\{(j,k) : \Theta_{jk} \neq 0\}\big)(\log p)/n$; $\|\widehat{\mathbf{\Theta}} - \mathbf{\Theta}\|_S$?
- CLIME (Cai et al 10) or Dantzig Selector-by-row (Yuan 10)
  - Minimize $\|\mathbf{\Theta}_{j,*}\|_1$ subject to $\|\overline{\mathbf{\Sigma}}\mathbf{\Theta} - \boldsymbol{I}_{p\times p}\|_{\max} \le \lambda$, $j \in [p]$
  - $\|\widehat{\mathbf{\Theta}} - \mathbf{\Theta}\|_S \lesssim d^* \big(\max_j \|\mathbf{\Theta}_{j,*}\|_1\big)^{\kappa} \sqrt{(\log p)/n}$; $\widehat{\Theta}_{jk}$?
  - $d^* = \max_j \#\{k : \Theta_{jk} \neq 0\} =$ max degree
- Lasso-by-row (Meinshausen-Bühlmann 06; Sun-Z 13)
  - Minimize $(\mathbf{\Theta}\overline{\mathbf{\Sigma}}\mathbf{\Theta})_{jj}\Theta_{jj}^{-3/2} + \Theta_{jj}^{-1/2} + 4\sqrt{(\log p)/n}\|\mathbf{\Theta}_{j,\{j\}^c}/\Theta_{jj}\|_1$, $j \in [p]$
  - $\|\widehat{\mathbf{\Theta}} - \mathbf{\Theta}\|_S \lesssim d^* \sqrt{(\log p)/n}$; $\widehat{\Theta}_{jk}$?
- Analysis
  - $\|\mathbf{\Sigma}\|_S + \|\mathbf{\Theta}\|_S = O(1)$
  - Large deviation of $\max_{j,k} |\overline{\Sigma}_{jk} - \Sigma_{jk}|$
  - Large deviation of $\max_{|A_\ell| \le d^*, \ell \le m} \|(\overline{\mathbf{\Sigma}} - \mathbf{\Sigma})_{A_\ell \times A_\ell}\|_S$

Confidence intervals and hypothesis testing for Θ_{jk}

- Confidence intervals/p-values after model selection
  - In the presence of small signals, it is impossible to construct approximate confidence intervals or p value for Θ_{jk} based on a selected model (LeCam 72; Leeb-Pötscher 06)
  - Conservative methods (Berk et al 09,11; Laber-Murphy 11)
- Semiparametric approach (Z 11)
  - Bias correction and Low-dimensional projection methods (Z-Zhang 11), desparsification (van de Geer et al 13)
  - Thresholded-LDPE with multiplicity adjustment (Z-Zhang 11), resparsification (van de Geer et al 13)
- Estimation of variance in linear regression (Sun-Z 12)
  - Equivalent to the estimation of $1/\Theta_{jj}$
  - Minimize $(\mathbf{\Theta}\overline{\mathbf{\Sigma}}\mathbf{\Theta})_{jj}\Theta_{jj}^{-3/2} + \Theta_{jj}^{-1/2} + 4\sqrt{(\log p)/n}\|\mathbf{\Theta}_{j,\{j\}^c}/\Theta_{jj}\|_1$
  - $\sqrt{n}(\widehat{\Theta}_{jj} - \Theta_{jj})/\widehat{\Theta}_{jj} \to N(0,2)$
  - Analysis: large deviation of $\max_{|A|\le d^*} \|(\overline{\mathbf{\Sigma}} - \mathbf{\Sigma})_{A\times A}\|_S$

Confidence intervals and hypothesis testing for Θ_{jk}

- Estimation of covariance in linear regression (Sun-Z 12; Ren et al 2013)
  - Multivariate linear regression
  - Subject to $\boldsymbol{B}_{A\times A} = \boldsymbol{I}_A$ with $\boldsymbol{B} \in \mathbb{R}^{[p]\times A}$, minimize

$$\mathrm{diag}^{-1}(\sigma_A)(\boldsymbol{B}^\top \overline{\boldsymbol{\Sigma}} \boldsymbol{B})_{A\times A} + \|\sigma_A\|_1 + 4\sqrt{(\log p)/n} \sum_{j\in A} \|\boldsymbol{B}_{*,j}\|_1$$

  - Estimate the $A \times A$ diagonal block of $\boldsymbol{\Theta}$ by

$$\widehat{\boldsymbol{\Theta}}_{A\times A} = (\widehat{\boldsymbol{B}}^\top \overline{\boldsymbol{\Sigma}} \widehat{\boldsymbol{B}})^{-1}_{A\times A}$$

  - $\widehat{\Theta}_{jk}$ with $A = \{j, k\}$
- Validity of statistical inference:

$$\max_{1\le j\le k\le p} \sup_t \left| \mathbb{P}\left\{ \frac{\sqrt{n}(\widehat{\Theta}_{jk} - \Theta_{jk})}{\sqrt{\widehat{\Theta}_{jk}^2 + \widehat{\Theta}_{jj}\widehat{\Theta}_{kk}}} \le t \right\} - \Phi(t) \right| \to 0$$

  provided that $\|\boldsymbol{\Sigma}\|_S + \|\boldsymbol{\Theta}\|_S = O(1)$ and $d^*(\log p) = o(n^{1/2})$
- The Fisher information bound attained

Confidence intervals and hypothesis testing for Θ_{jk}

- Optimality of the proposed method (Ren et al 2013)
  - For $\mathscr{C}_{M_0,d_0} = \{\mathbf{\Theta} : \|\mathbf{\Sigma}\|_S + \|\mathbf{\Theta}\|_S \le M_0, d^* \le d_0\}$ and $c_0 > 0$,

$$\min_{1\le j\le k\le p} \inf_{\widetilde{\Theta}} \sup_{\mathbf{\Theta}\in\mathscr{C}_{M_0,d_0}} \mathbb{P}\left\{\left(\widetilde{\Theta}_{jk} - \Theta_{jk}| \ge c_0 \left(\frac{1}{\sqrt{n}} + \frac{d_0 \log p}{n}\right)\right)\right\} \ge c_0$$

$$\inf_{\widetilde{\Theta}} \sup_{\mathbf{\Theta}\in\mathscr{C}_{M_0,d_0}} \mathbb{P}\left\{\left(\|\widetilde{\Theta}_{jk} - \Theta_{jk}\|_{\max} \ge c_0 \left(\frac{\sqrt{\log p}}{\sqrt{n}} + \frac{d_0 \log p}{n}\right)\right)\right\} \ge c_0$$

  - When $d_0(\log p)/n \le a_0$ for a certain small $a_0 > 0$,

$$\lim_{t\wedge n\to\infty} \max_{1\le j\le k\le p} \sup_{\mathbf{\Theta}\in\mathscr{C}_{M_0,d_0}} \mathbb{P}\left\{\left(\widehat{\Theta}_{jk} - \Theta_{jk}| \ge t \left(\frac{1}{\sqrt{n}} + \frac{d_0 \log p}{n}\right)\right)\right\} = 0$$

$$\sup_{\mathbf{\Theta}\in\mathscr{C}_{M_0,d_0}} \mathbb{P}\left\{\left(\|\widehat{\Theta}_{jk} - \Theta_{jk}\|_{\max} \ge t_0 \left(\frac{\sqrt{\log p}}{\sqrt{n}} + \frac{d_0 \log p}{n}\right)\right)\right\} \to 0, \ \exists t_0$$

- Lower bound: $\mathbb{E}_{\mathbf{I}} \frac{d\mathbb{P}_{\mathbf{\Theta}}}{d\mathbb{P}_{\mathbf{I}}} - 1 \to 0$ with suitable random $\mathbf{\Theta} \in \mathscr{C}_{M_0,d_0}$

Copulas

- Data: $\boldsymbol{Y} = (y_{ij})_{n\times p} = (\boldsymbol{y}_1, \ldots, \boldsymbol{y}_p)$, $y_{ij} = f_j(x_{ij})$ with $\uparrow\uparrow f_j$
- Kendall's (1938) tau between vectors $\boldsymbol{y}_j$ and $\boldsymbol{y}_k$ is

$$\widehat{\tau}_{jk} = 2\{n(n-1)\}^{-1} \sum_{1\le i_1 < i_2 \le n} \operatorname{sgn}(y_{i_1,j} - y_{i_2,j}) \operatorname{sgn}(y_{i_1,k} - y_{i_2,k})$$

- Spearman's (1904) rank correlation between $\boldsymbol{y}_j$ and $\boldsymbol{y}_k$ is

$$\widehat{\rho}_{jk} = \operatorname{Corr}(\boldsymbol{r}_j, \boldsymbol{r}_k)$$

  where $\boldsymbol{r}_j$ is the rank vector of $\boldsymbol{y}_j$, $r_{i_0,j} = \sum_{i=1}^{n} I\{y_{ij} \le y_{i_0,j}\}$
- If $\boldsymbol{X} = (x_{ij})_{n\times p}$ has iid rows and continuous elements, then

$$\begin{aligned}
\mathbb{E}\widehat{\tau}_{jk} &= \tau_{jk} = \mathbb{E}\Big\{\operatorname{sgn}(x_{1,j} - x_{2,j}) \operatorname{sgn}(x_{1,k} - x_{2,k})\Big\}\Big(\\
\mathbb{E}\widehat{\rho}_{jk} &= \{(n-2)\rho_{jk} + 3\tau_{jk}\}/(n+1) \\
\rho_{jk} &= 3\,\mathbb{E}\Big\{\operatorname{sgn}(x_{1,j} - x_{2,j}) \operatorname{sgn}(x_{1,k} - x_{3,k})\Big\}\Big(
\end{aligned}$$

- Moreover, $X_j \perp\!\!\!\perp X_k$ implies that $\mathbb{E}\widehat{\tau}_{jk} = \mathbb{E}\widehat{\rho}_{jk} = 0$

Copulas

- Kendall (1948), Kruskal (1958)
- Gaussian copula/nonparanormal: If the rows of $\boldsymbol{X} = (x_{ij})_{n \times p}$ are iid Gaussian with *correlation* matrix $\boldsymbol{\Sigma}$, then

$$\Sigma_{jk} = \sin(\pi \tau_{jk}/2) = 2 \sin(\pi \rho_{jk}/6)$$

- Elliptical copula/trans-elliptical: If the rows of $\boldsymbol{X}$ are iid from

$$\xi \boldsymbol{A} \boldsymbol{U}$$

with uniform $\boldsymbol{U}$ in the sphere $\mathbb{S}^{p-1}$, $\xi \in \mathbb{R}$ independent of $\boldsymbol{U}$ and a deterministic $\boldsymbol{A}$. Then,

$$\Sigma_{jk} = \sin(\pi \tau_{jk}/2)$$

with $\boldsymbol{\Sigma}$ being the correlation matrix corresponding to $\boldsymbol{A}\boldsymbol{A}^T$

- The Kendall-Kruskal formula motivates $\widehat{\boldsymbol{\Sigma}}^{\tau}$ and $\widehat{\boldsymbol{\Sigma}}^{\rho}$ with elements

$$\widehat{\Sigma}^{\tau}_{jk} = \sin\left(\frac{\pi}{2}\widehat{\tau}_{jk}\right), \quad \widehat{\Sigma}^{\rho}_{jk} = 2 \sin\left(\frac{\pi}{6}\widehat{\rho}_{jk}\right).$$

What happens if we carry out high-dimensional multivariate analysis/learning using these nonparametric statistics?

Will the "best" results match those of the sample correlation matrix $\widehat{\boldsymbol{\Sigma}}^s$ based on Gaussian data $\boldsymbol{X}$? Here

$$\widehat{\boldsymbol{\Sigma}}^s = \boldsymbol{D}^{-1/2}\boldsymbol{X}^T\boldsymbol{X}\boldsymbol{D}^{-1/2},\ \boldsymbol{D} = \mathrm{diag}(\boldsymbol{X}^T\boldsymbol{X}).$$

Examples:

- Estimation of sparse $\boldsymbol{\Sigma}$
- PCA
- Estimation of sparse $\boldsymbol{\Sigma}^{-1}$
- Statistical inference about $(\boldsymbol{\Sigma}^{-1})_{jk}$ in graphical models

- In Liu-Lafferty-Wasserman (2009), Xue-Zou (2012a,b), Liu et al (2012), Han-Liu (2012), Liu-Han-Z (2012) and more, high-dimensional methodologies were proposed and analyzed based on error bounds on

$$\|\widehat{\mathbf{\Sigma}} - \mathbf{\Sigma}\|_{\max} = \max_{j,k} |\widehat{\Sigma}_{jk} - \Sigma_{jk}|$$

- Since $\widehat{\Sigma}^{\tau}_{jk}$ and $\widehat{\Sigma}^{\rho}_{jk}$ are Lipschitz in U-statistics with bounded kernels, its large deviation property is comparable to that of the sample correlation:

$$\mathbb{P}\Big\{ \|\widehat{\mathbf{\Sigma}} - \mathbf{\Sigma}\|_{\max} > Ct \Big\} \le p^2 \max_{j,k} \mathbb{P}\Big\{ |\widehat{\Sigma}_{jk} - \Sigma_{jk}| > Ct \Big\} \le 2p^2 e^{-nt^2/2}$$

- Thus, if the performance of an estimator $\widehat{\theta} = \theta(\widehat{\mathbf{\Sigma}}^s)$ is guaranteed by such large deviation bounds for $\|\widehat{\mathbf{\Sigma}}^s - \mathbf{\Sigma}\|_{\max}$, then the same guarantee is valid if $\widehat{\mathbf{\Sigma}}^s$ is replaced by $\widehat{\mathbf{\Sigma}}^{\tau}$ or $\widehat{\mathbf{\Sigma}}^{\rho}$ in copula models where $\boldsymbol{X}$ is not available.
- Examples of applications
  - Thresholding raw $\widehat{\mathbf{\Sigma}}$ for sparse $\mathbf{\Sigma}$ (Bickel-Levina, 2008)
  - Extension of $\|\cdot\|_F$ bounds for G-Lasso or SCAD
  - Extension of $\|\cdot\|_S$ for the CLIME

Spectrum error bounds for $\widehat{\mathbf{\Sigma}}^{\tau}$ (Han-Liu 13; Wegkamp-Zhao 13)

- Wirte $\widehat{\boldsymbol{T}} = (\widehat{\tau}_{jk})_{p\times p}$, $\widehat{\mathbf{\Sigma}}^{\tau} = (\sin((\pi/2)\widehat{\tau}_{jk}))_{p\times p}$ and $\mathbf{\Sigma}^{\tau} = (\sin((\pi/2)\mathbb{E}\,\widehat{\tau}_{jk}))_{p\times p}$
- Use the convexity of $\operatorname{trace}(e^{\boldsymbol{M}})$ for $\boldsymbol{M} = \boldsymbol{M}^{\top}$ to decouple U-statistics,

$$\mathbb{E}\operatorname{trace}\left(\exp(\lambda(\widehat{\boldsymbol{T}} - \mathbb{E}\,\widehat{\boldsymbol{T}}))\right) \le \mathbb{E}\operatorname{trace}\left(\exp\left(\frac{\lambda}{\lfloor n/2\rfloor}\sum_{i=1}^{\lfloor n/2\rfloor}\boldsymbol{h}(\boldsymbol{x}^{2i}, \boldsymbol{x}^{2i-1})\right)\right)$$

  where $\boldsymbol{h}$ is the kernel of the U-statistics $\widehat{\boldsymbol{T}}$ and $\boldsymbol{x}^i$ are the iid rows of $\boldsymbol{X}$
- Use a matrix Bernstein inequality (Oliveira 10; Troop 11) to bound the tail

$$\mathbb{P}\left\{\|\widehat{\boldsymbol{T}} - \mathbb{E}\,\widehat{\boldsymbol{T}}\|_S > t\right\} \le 2p\exp\frac{-\lfloor n/2\rfloor t^2/2}{p\|\mathbf{\Sigma}\|_S + \|\mathbf{\Sigma}\|_S^2 + 2pt}$$

- Requirement of $n \gg p\log p$ for $\|\cdot\|_S$ convergence, with an extra $\log p$
- Lack of concentration for controlling $\|\cdot\|_S$, e.g. compared with

$$\mathbb{P}\left\{\left\|\boldsymbol{X}^{\top}\boldsymbol{X}/n - \boldsymbol{I}\right\|_S \ge (\sqrt{d/n} + t)(1 + \sqrt{d/n} + t)\right\} \le 2e^{-nt^2/2}$$

  for the sample covariance matrix of $N(0, \boldsymbol{I})$ data (Davidson-Szarek, 01)

Spectrum error bounds in Gaussian copula/nonparanormal models (Mitra-Z 14)

- For both nonparametric estimates $\widehat{\mathbf{\Sigma}} = \widehat{\mathbf{\Sigma}}^{\tau}$ and $\widehat{\mathbf{\Sigma}} = \widehat{\mathbf{\Sigma}}^{\rho}$

$$\mathbb{E} \left\| \widehat{\mathbf{\Sigma}} - \mathbf{\Sigma} \right\|_S \le C_0 \left\| \mathbf{\Sigma} \right\|_S \left(\sqrt{p/n} + p/n \right)$$

- For $\mathscr{A}_{M,d,m} = \{A : \|\mathbf{\Sigma}_{A\times A}\|_S \le M, |A| \le d\}$ with $|\mathscr{A}_{M,d,m}| \le m$,

$$\mathbb{P}\left\{ \max_{A \in \mathscr{A}_{M,d,m}} \|(\widehat{\mathbf{\Sigma}} - \mathbf{\Sigma})_{A\times A}\|_S \ge C_M(\Delta_{d,m,t} + \Delta_{d,m,t}^2 + \Delta'_{d,m,t}) \right\} \le e^{-nt}$$

  where C_M depends on M only, $\Delta_{d,m,t} = \sqrt{(d + \log m)/n + t}$ and $\Delta'_{d,m,t} = \sqrt{(\log m)/n + t} + d(\log p)/n + t$

- Hoeffding and further decompositions; Lipschitz functions of $N(0, \boldsymbol{I})$

- Examples of applications
  - Consistency of PCA when $p/n \to 0$
  - d-sparse PCA when $d(\log p)/n \to 0$
  - Fast convergence for estimating "d-bandable" $\mathbf{\Sigma}$ (Cai-Z-Zhou, 10)

Large deviation problems in the elliptical model and in general

- Sub-Gaussian property of

$$(\operatorname{sgn}(X_1), \ldots, \operatorname{sgn}(X_p)),$$

  the sign sub-Gaussian condition (Han-Liu, 13)

- Concentration in the spectrum norm of

$$\left(n^{-1} \sum_{i=1}^{n} \left(g(x_{ij}, x_{ik}) - \mathbb{E}\, g(X_j, X_k) \right) \right)_{p \times p}$$

  with a smooth $g(x, y)$ satisfying $\|g\|_\infty \leq 1$. We solved this problem with a strong condition on $\{g_{jk}\}$ which happens to hold for Gaussian copulas

Thanks!