Estimation and Inference for Differential Networks

Mladen Kolar (mkolar@chicagobooth.edu)

Collaborators

Byol Kim (U Chicago)

Song Liu (U Bristol)

Motivation

Example: ADHD-200 brain imaging dataset (Biswal et al. 2010)

The ADHD-200 dataset is a collection of resting-state functional MRI of subjects with and without attention deficit hyperactive disorder.

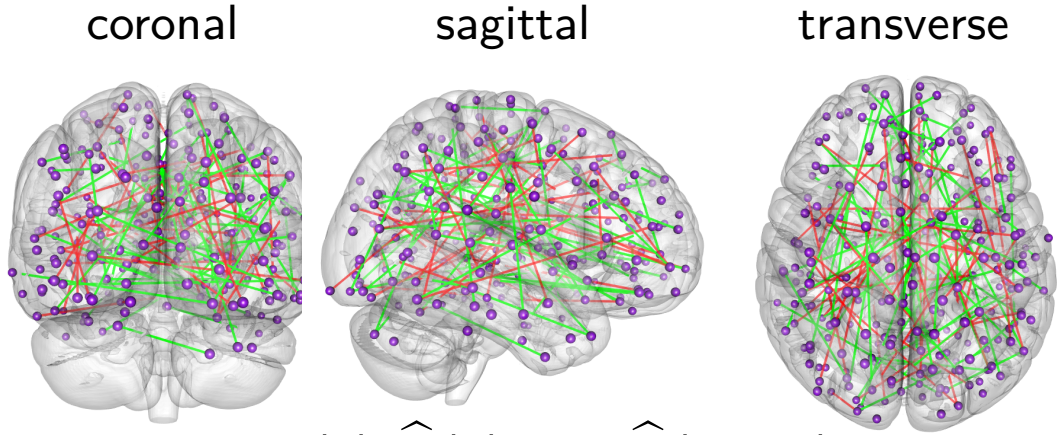
- subjects: 491 controls, 195 cases diagnosed with ADHD of various types
- for each subject there are between 76 and 276 R-fMRI scans
- focus on 264 voxels as the regions of interest (ROI); extracted by Power et al. (2011)

We are interested in understanding how the structure of the neural network varies with age of subjects. (Lu, Kolar, and Liu 2017)

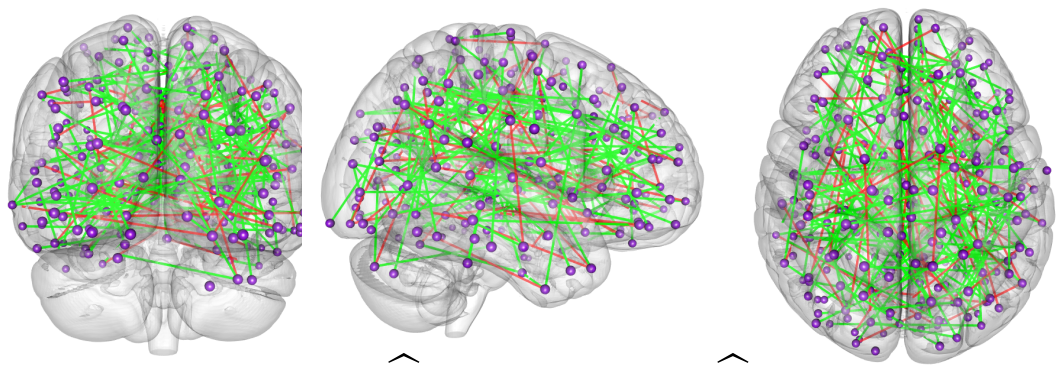
Estimated Brain Networks

The differences between junior, median and senior neural networks.

- The green edges only exist in $\widehat{G}(11.75)$ and the red edges only exist in $\widehat{G}(8)$.
- The green edges only exist in $\widehat{G}(20)$ and the red edges only exist in $\widehat{G}(11.75)$.



(a) $\widehat{G}(8)$ v.s. $\widehat{G}(11.75)$



(b) $\widehat{G}(11.75)$ v.s. $\widehat{G}(20)$

Limitations

While we can estimate the networks and their difference, it is hard to quantify uncertainty in the estimates and perform statistical inference.

The goal here is to develop methodology capable of doing so.

Probabilistic Graphical Models

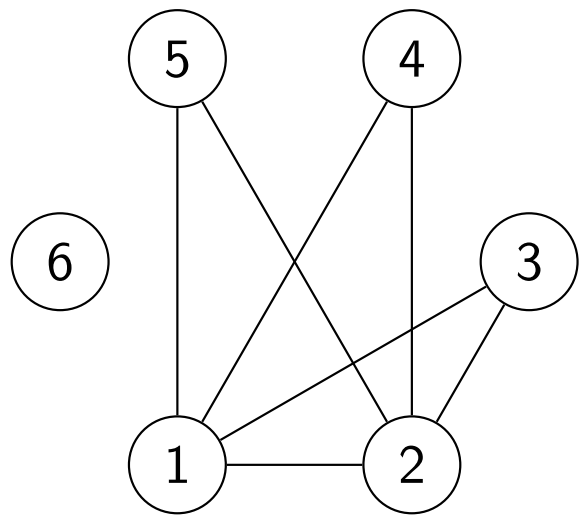
- Graph $G = (V, E)$ with p nodes
- Random vector $\mathbf{X} = (X_1, \ldots, X_p)'$

Represents conditional independence relationships between nodes

Useful for exploring associations between measured variables

$$(a, b) \notin E \iff X_a \perp X_b \mid X_{\overline{ab}}$$

$$X_1 \perp X_6 \mid X_2, \ldots, X_5$$

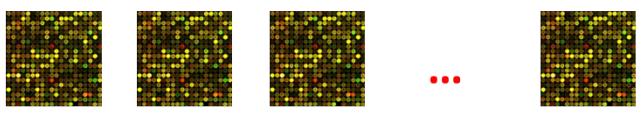


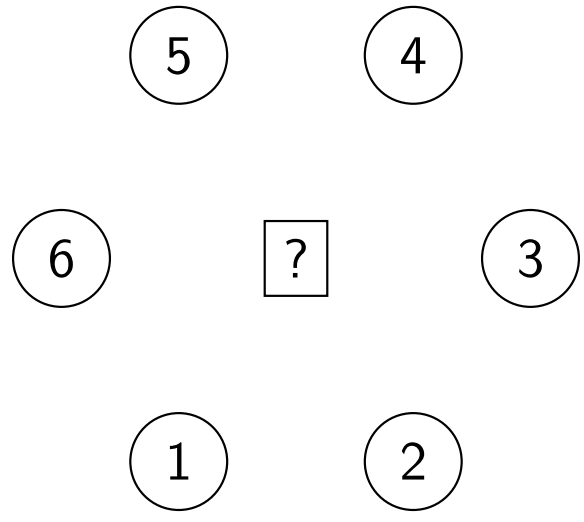
Structure Learning Problem

Given an i.i.d. sample $\mathcal{D}_n = \{\mathbf{x}_i\}_{i=1}^n$ from a distribution $\mathbb{P} \in \mathcal{P}$

Learn the set of conditional independence relationships

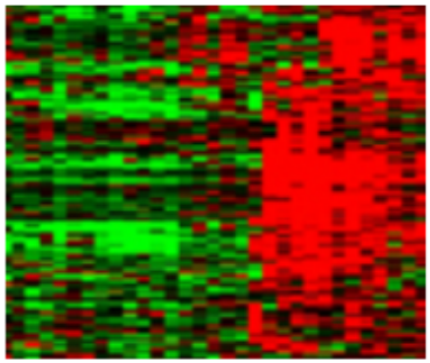
$$\widehat{G} = \widehat{G}(\mathcal{D}_n)$$

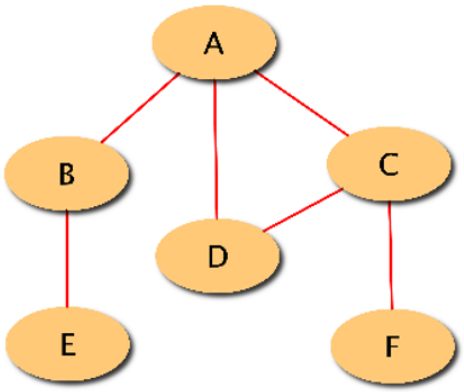


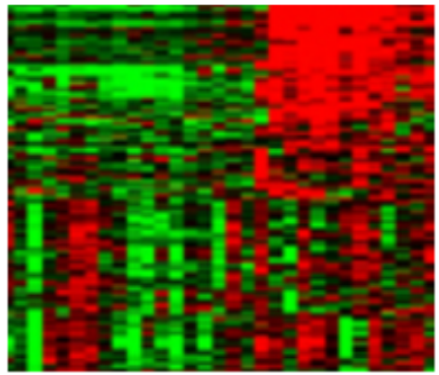


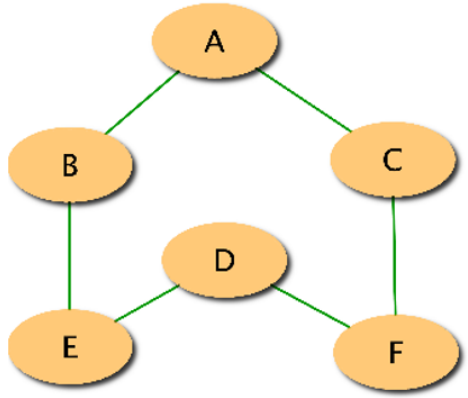
Some literature: Yuan and Lin (2007), O. Banerjee, El Ghaoui, and d'Aspremont (2008), Friedman, Hastie, and Tibshirani (2008), T. T. Cai, Liu, and Luo (2011), Meinshausen and Bühlmann (2006), Ravikumar, Wainwright, and Lafferty (2010), Xue, Zou, and Ca (2012), Yang et al. (2012), Yang et al. (2015), Yang et al. (2013), Yang et al. (2014), ...

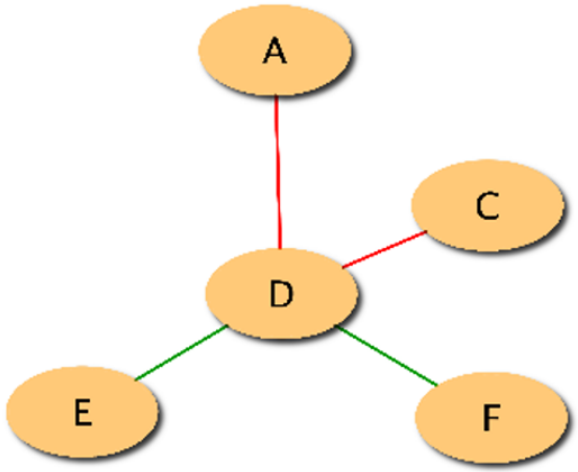
Difference Estimation











Literature Review: Gaussian Graphical Models

Penalized estimation (B. Zhang and Wang 2012, Danaher, Wang, and Witten (2014))

$$\widehat{\Omega}_x, \widehat{\Omega}_y = \arg \max_{\Omega_x \succ \mathbf{0}, \Omega_y \succ \mathbf{0}} \sum_{c \in \{x,y\}} \left(\log |\Omega_c| - \operatorname{tr} \widehat{\Sigma}_c \Omega_c - \lambda \|\Omega_c\|_1 \right) - \lambda_2 \|\Omega_x - \Omega_y\|_1$$

Direct difference estimation (S. D. Zhao, Cai, and Li 2014)

$$\widehat{\Delta} = \arg \min_{\Delta} \|\Delta\|_1 \quad \text{subject to} \quad \|\widehat{\Sigma}_x \Delta \widehat{\Sigma}_y - (\widehat{\Sigma}_y - \widehat{\Sigma}_x)\|_\infty \leq \lambda$$

Literature Review: Inference for Graphical Models

Inference is available for single group testing

- Ren et al. (2015), Lu, Kolar, and Liu (2017), J. Wang and Kolar (2016), Barber and Kolar (2015), Yu, Gupta, and Kolar (2016), Chen et al. (2015), Jankova and van de Geer (2015, 2016)

Xia, Cai, and Cai (2015) combines inference results for each precision matrix based on Ren et al. (2015). We will discuss this result more later.

Inference for Differential Markov Random Fields

The Setting

Let $\mathcal{F} = \{f(\,\cdot\,;\theta)\}$ be a parametric family of probability distributions of the form

$$f(y;\theta) = Z(\theta)^{-1} \exp\Big(\sum_{1\le i\le j\le m} \theta_{ij}\,\psi_{ij}(y_i, y_j)\Big) = Z(\theta)^{-1} \exp\left(\theta^\top \psi(y)\right)$$

Given samples

- $\{x_i\}_{i\in[n_x]}$ from $f(x;\theta_x)$, and
- $\{y_i\}_{i\in[n_y]}$ from $f(y;\theta_y)$

Our goal is to estimate the change

$$\delta_\theta = \theta_x - \theta_y$$

Density Ratio Approach

Kullback-Leibler importance estimation procedure (KLIEP) (Sugiyama et al. 2008)

$$r(y;\delta_\theta) = \frac{f_x(y)}{f_y(y)} = \left(\frac{Z(\theta_x)}{Z(\theta_y)}\right)^{-1} \exp\Big(\big(\underbrace{\theta_x - \theta_y}_{\delta_\theta}\big)^\top \psi(y)\Big)$$

$$\begin{aligned}\delta_\theta &= \arg\min_{\delta_\theta} D_{\mathsf{KL}}(f_x \,\|\, r(\,\cdot\,;\delta_\theta) f_y) \\ &= \arg\min_{\delta_\theta} \Big\{ -\mathbb{E}_x\Big[\delta_\theta^\top \psi(x)\Big] + \log \mathbb{E}_y\Big[\exp\big(\delta_\theta^\top \psi(y)\big)\Big]\Big\}.\end{aligned}$$

Sparse Direct Difference Estimation

Given the data, we replace the expectations with the appropriate sample averages to form the empirical KLIEP loss

$$\ell_{\text{KLIEP}}(\delta_\theta; X, Y) = -\frac{1}{n_x}\sum_{i=1}^{n_x} \delta_\theta^\top \psi(x_i) + \log\left[\frac{1}{n_y}\sum_{j=1}^{n_y} \exp\left(\delta_\theta^\top \psi(y_j)\right)\right]$$

Under a suitable set of assumptions, the penalized estimator

$$\widehat{\delta}_\theta = \arg\min_{\delta_\theta} \ell_{\text{KLIEP}}(\delta_\theta; X, Y) + \lambda\|\delta_\theta\|_1$$

can be shown to be consistent (S. Liu et al. 2017, Fazayeli and Banerjee (2016)).

Inference For a Low Dimensional Component

Let $\delta_\theta = (\delta_{\theta,1}, \delta_{\theta,2})$. Our goal is construction of an asymptotically Normal estimator of $\delta_{\theta,1}$.

An oracle estimator

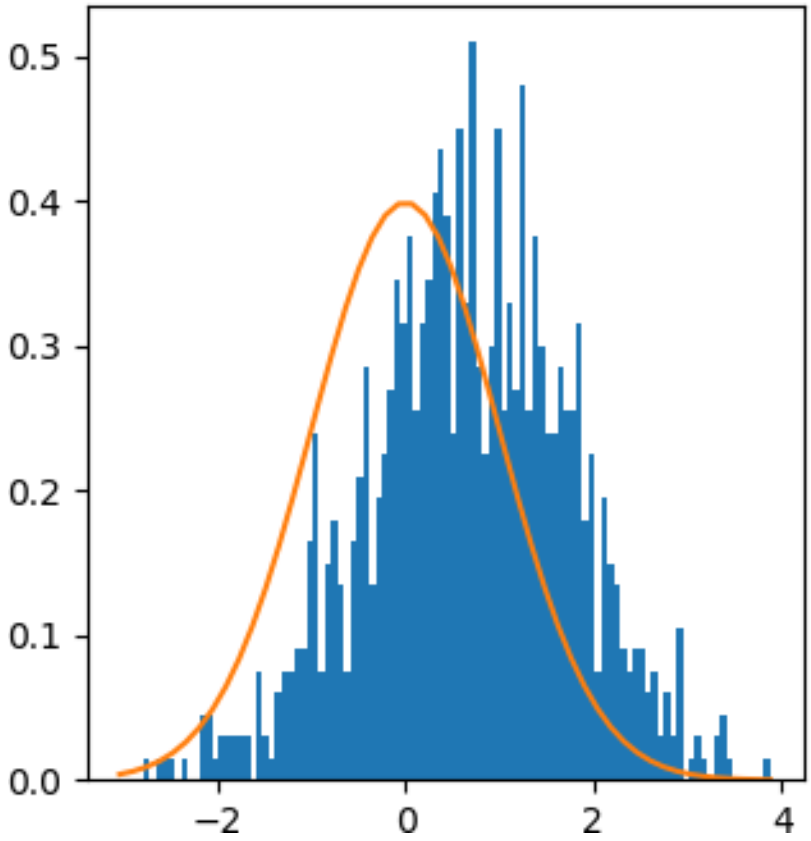
$$\widehat{\delta}_{\theta,1} = \arg\min_{\delta_{\theta,1}} \ell_{\mathsf{KLIEP}}(\delta_{\theta,1}, \delta^*_{\theta,2})$$

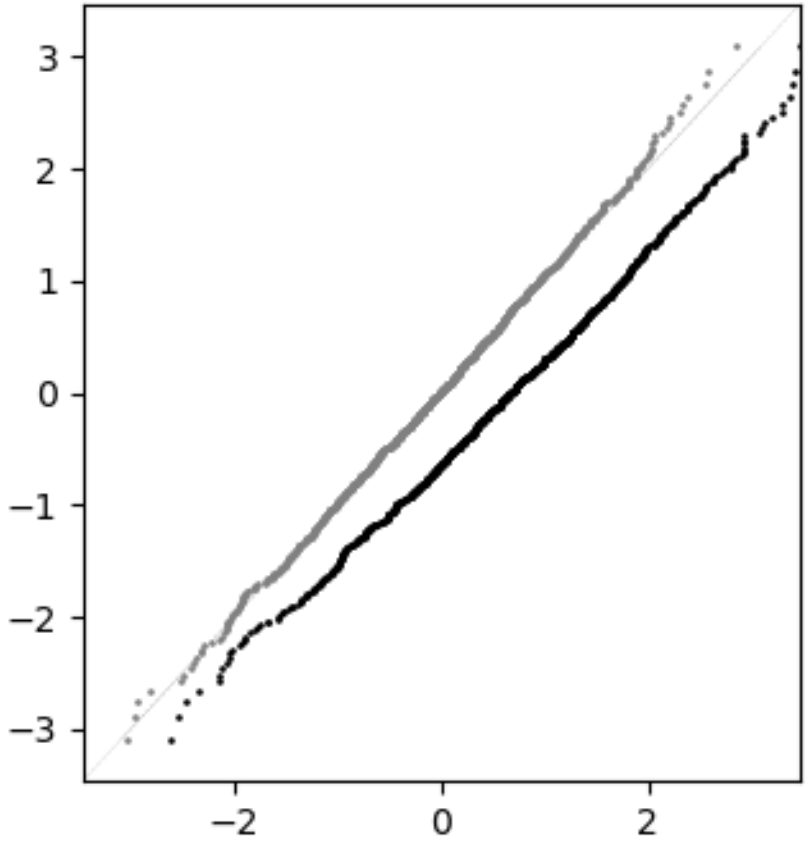
is asymptotically Normal.

What about a feasible estimator

$$\widehat{\delta}_{\theta,1} = \arg\min_{\delta_{\theta,1}} \ell_{\mathsf{KLIEP}}(\delta_{\theta,1}, \widehat{\delta}_{\theta,2})?$$

Performance of a Naive Estimator





Feasible Plug-in Estimator

$$0 = \partial_1 \ell_{\text{KLIEP}}(\widehat{\delta}_{\theta,1}, \widehat{\delta}_{\theta,2}) = \partial_1 \ell_{\text{KLIEP}}(\delta^*_{\theta,1}, \delta^*_{\theta,2}) + H_{11}(\widehat{\delta}_{\theta,1} - \delta^*_{\theta,1}) + H_{12}(\widehat{\delta}_{\theta,2} - \delta^*_{\theta,2}) + o_p(1)$$

- $H = \nabla^2 \ell_{\text{KLIEP}}(\delta^*_\theta; Y)$
- $\widehat{\delta}_{\theta,2}$ is a consistent estimator of $\delta^*_{\theta,2}$ with $\|\widehat{\delta}_{\theta,2} - \delta^*_{\theta,2}\|_2 \lesssim_P \sqrt{\frac{s \log(p)}{n}}$

The problem is that the limiting distribution depends on the model selection procedure used to estimate $\delta^*_{\theta,2}$.

ω-Corrected Score Function

Consider a linear combination of the components of the score vector

$$S_\omega(\delta_{\theta,1}, \delta_{\theta,2}) = \partial_1 \ell_{\mathsf{KLIEP}}(\delta_{\theta,1}, \delta_{\theta,2}) - \omega^\top \partial_2 \ell_{\mathsf{KLIEP}}(\delta_{\theta,1}, \delta_{\theta,2})$$

Our estimator $\widetilde{\delta}_{\theta,1}$ is a solution to the following estimating equation

$$0 = S_\omega(\delta_{\theta,1}, \widehat{\delta}_{\theta,2})$$

ω-Corrected Score Function

$$
\begin{aligned}
0 = \sqrt{n} S_\omega(\widetilde{\delta}_{\theta,1}, \widehat{\delta}_{\theta,2}) = \sqrt{n} S_\omega(\delta^*_{\theta,1}, \delta^*_{\theta,2}) \\
&+ \sqrt{n}\left(H_{11} - \omega^T H_{21}\right)\left(\widetilde{\delta}_{\theta,1} - \delta^*_{\theta,1}\right) \\
&+ \sqrt{n}\left(H_{12} - \omega^T H_{22}\right)\left(\widehat{\delta}_{\theta,2} - \delta^*_{\theta,2}\right) + o_P(1)
\end{aligned}
$$

In order to have $\sqrt{n}\left(\widetilde{\delta}_{\theta,1} - \delta^*_{\theta,1}\right) \to_D N(0, \sigma^2_{\delta_{\theta,1}})$ we need:

- $\mathbb{E}[S_\omega(\widetilde{\delta}_{\theta,1}, \widehat{\delta}_{\theta,2})] = 0$
- $\sqrt{n}\left(H_{12} - \omega^T H_{22}\right)\left(\widehat{\delta}_{\theta,2} - \delta^*_{\theta,2}\right)$ vanishes sufficiently fast

ω-Corrected Score Function

Idea from Ning and Liu (2017)

Construct ω as a solution to

$$H_{12} = \omega^T H_{22}$$

That is $\omega = H_{22}^{-1} H_{21}$

Sample estimate obtained as

$$\widehat{\omega} = \arg\min_{\omega} \tfrac{1}{2}\omega^\top H(\widehat{\delta}_\theta; Y)\omega - \omega^\top e_1 + \lambda_2 \|\omega\|_1$$

- needs to converge sufficiently fast

Our Practical Procedure

Compute an initial estimate $\widetilde{\delta}_\theta$:

$$\widetilde{\delta}_\theta \leftarrow \arg\min_{\delta_\theta} \ell_{\mathsf{KLIEP}}(\delta_\theta; X, Y) + \lambda_1 \|\delta_\theta\|_1$$

$$\widetilde{\delta}_\theta \leftarrow \arg\min_{\delta_\theta} \ell_{\mathsf{KLIEP}}(\delta_\theta; X, Y) : \mathrm{supp}(\delta_\theta) \subseteq \mathrm{supp}(\widetilde{\delta}\theta)$$

Compute a sparse approximation $\widetilde{\omega}$ to the corresponding row of the inverse of the Hessian:

$$\widetilde{\omega} \leftarrow \arg\min_{\omega} \tfrac{1}{2}\omega^\top H(\widetilde{\delta}_\theta; Y)\omega - \omega^\top e_1 + \lambda_2 \|\omega\|_1$$

Re-estimate on the union of the supports from Steps 1 and 2:

$$(\widehat{\delta}_{\theta,1}, \widehat{\delta}_{\theta,2}) \leftarrow \arg\min_{\delta_\theta} \ell_{\mathsf{KLIEP}}(\delta_\theta; X, Y) : \mathrm{supp}(\delta_\theta) \subseteq \mathrm{supp}(\widetilde{\delta}_\theta) \cup \mathrm{supp}(\widetilde{\omega})$$

Technical Conditions

High Level Technical Conditions:

- consistency for initial parameter estimation
- deviaton for the gradient and Hessian
- local smoothness of the loss
- conditions for the central limit theorem and estimation of the variance

Technical Conditions

- $\|\psi(y)\|_\infty = M < \infty$ with probability 1
- $\lambda_{\min}(H) \geq \underline{\lambda} > 0$
- ω is approximately sparse
- $\forall\, \Delta \in \mathcal{E} \cap \mathcal{B}(\mathbf{0}, \|\delta_\theta^*\|_1)$

$$\frac{1}{n_y^2} \sum_{1 \leq j < j' \leq n_y} \langle \psi(y_j) - \psi(y_{j'}), \Delta \rangle^2 \geq C^2 \left(\kappa_1 \|\Delta\|^2 - \kappa_2 \frac{\log p}{n_y} \|\Delta\|_1^2 \right)$$

with probability $1 - \delta_{n_y}$

Main Result

Suppose the regularity conditions hold. Let $n = n_x \wedge n_y$ be the smaller sample size and $\rho = \lim_{n\to\infty} n/(n_x \vee n_y)$.

With appropriately chosen regularization parameters λ_1, λ_2 out estimator $\widehat{\delta}_{\theta,1}$ satisfies

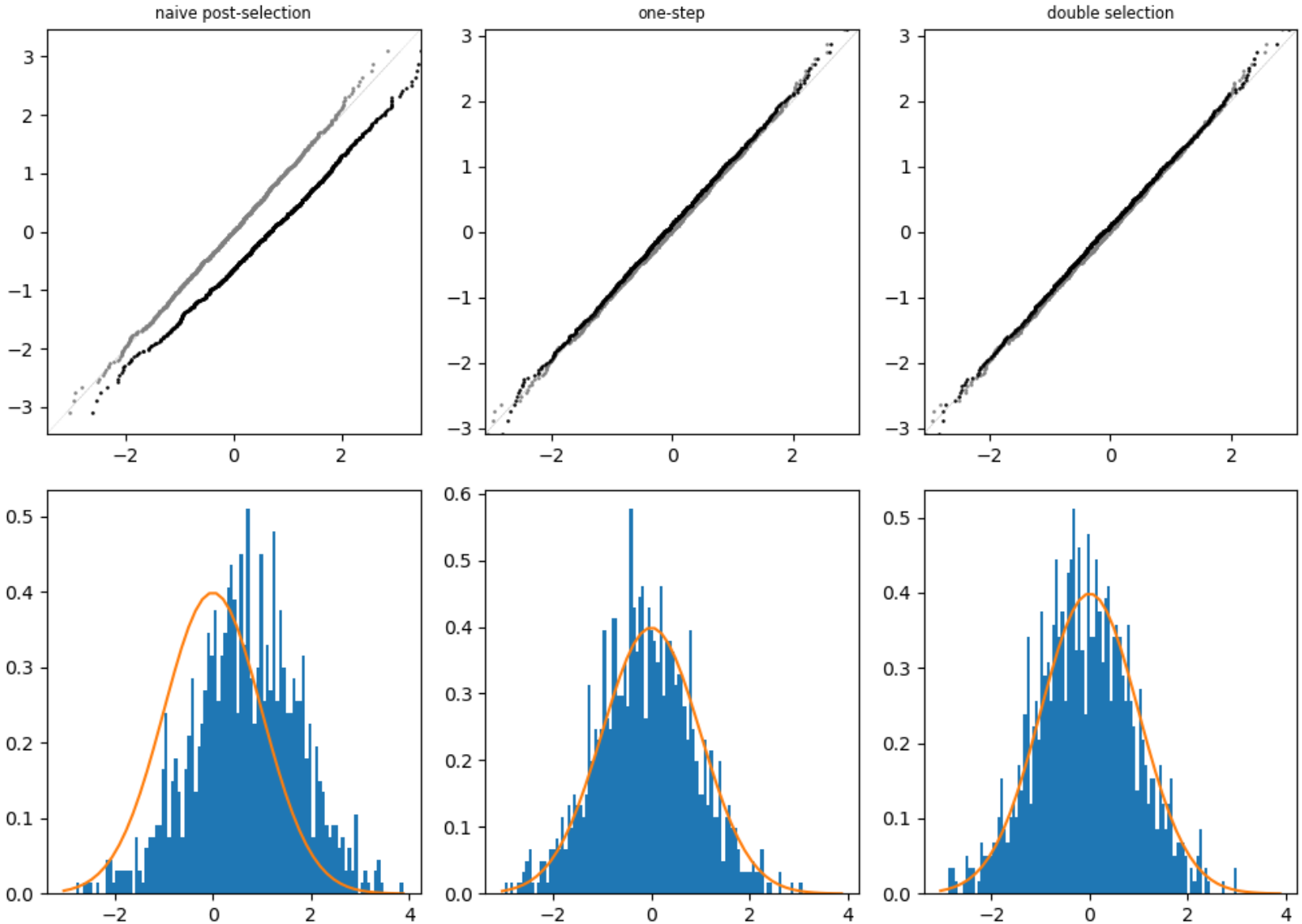
$$\sqrt{n}\sigma^{-1}(\widehat{\delta}_{\theta,1} - \delta^*_{\theta,1}) \to_D N(0, 1+\rho)$$

where

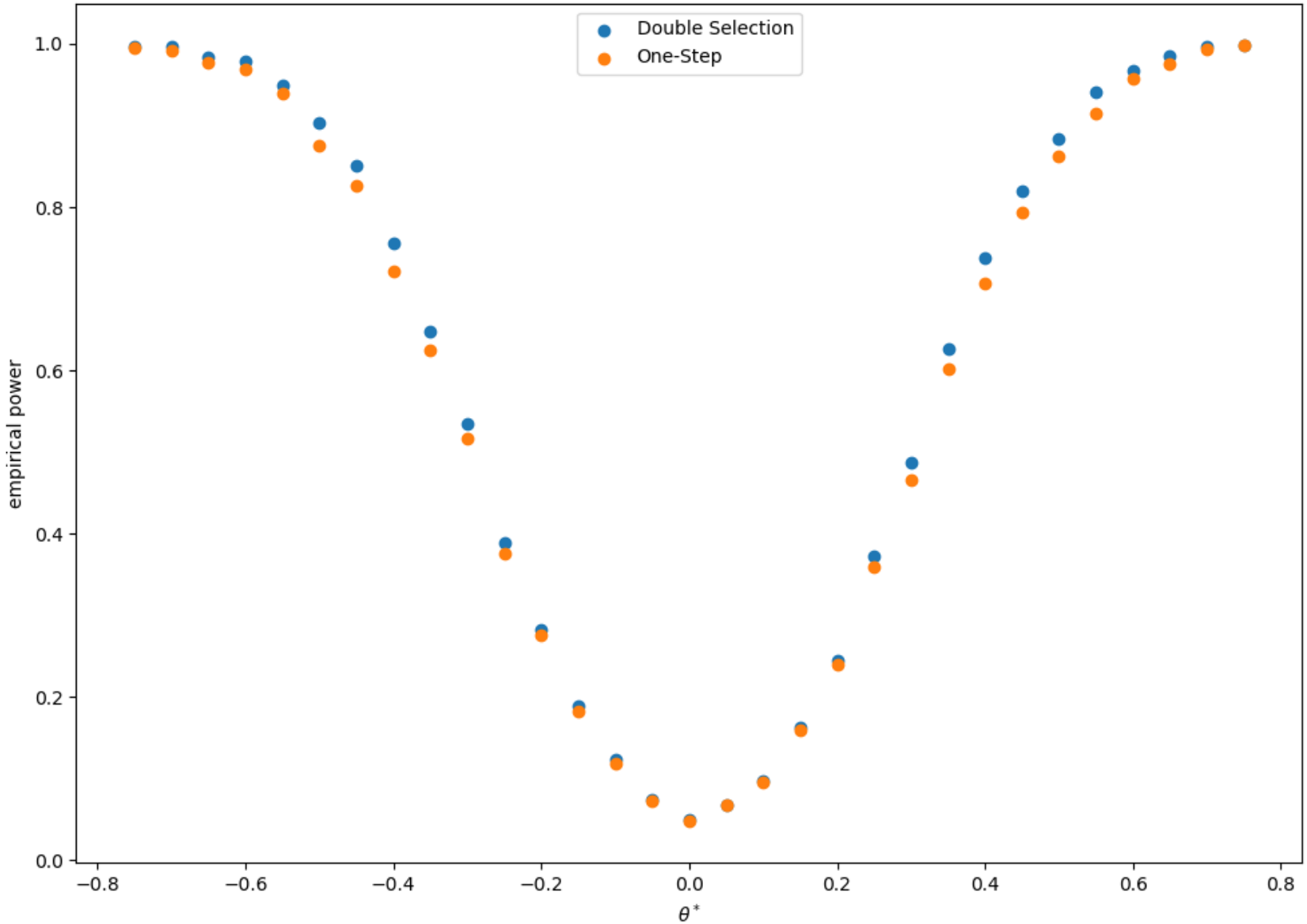
$$\sigma^2 = (\Sigma_{11} - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21})^{-1}.$$

Furthermore, σ^2 can be consistently estimated.

Simulation Plot



Power Plot


Double Selection
One-Step
empirical power
1.0
0.8
0.6
0.4
0.2
0.0
−0.8
−0.6
−0.4
−0.2
0.0
0.2
0.4
0.6
0.8
θ^*

Simultaneous Inference

For a potentially large I, we would like to simultaneously test

$$H_{0k} : \delta^*_{\theta,k} = \delta_{\theta,0k}, \quad k \in I \subseteq [p]$$

Or construct simultaneous confidence intervals

$$\mathbb{P}\left\{\delta^*_{\theta,k} \in (\widehat{\delta}_{\theta,k} - \widehat{\sigma}_k t_\alpha(I)/\sqrt{n}, \widehat{\delta}_{\theta,k} + \widehat{\sigma}_k t_\alpha(I)/\sqrt{n}) \quad \forall k \in I\right\} \geq 1 - \alpha$$

Simulation Assisted Simultaneous Inference

Inference based on the following test statistic

$$\max_{k \in I} \sqrt{n} |\widehat{\delta}_{\theta,k} - \delta^*_{\theta,k}|.$$

Approximate the distribution of the test statistic with

$$W = \max_k \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \left\langle \widehat{\omega}_k, -\psi(x_i) + \frac{1}{n_y} \sum_{j=1}^{n_y} \psi(y_j) \widehat{r}(y_j; \delta^*_\theta) \right\rangle \xi_i,$$

where $\{\xi_i\}_{i=1}^n$ are i.i.d. standard normal random variables.

The bootstrap critical value is the empirical quantile

$$t_\alpha(I) = \inf \{ t \in \mathbb{R} : \mathbb{P}\{W \leq t \mid \mathcal{X}, \mathcal{Y}\} \geq 1 - \alpha \}.$$

Future Work

Lower bounds

Can we say something about uncertainty when confounders are present?

- The problem is how to account for the fact that we are estimating the effect of confounders.

References

Banerjee, O., L. El Ghaoui, and A. d'Aspremont. 2008. "Model Selection Through Sparse Maximum Likelihood Estimation." *J. Mach. Learn. Res.* 9 (3): 485–516.

Barber, Rina Foygel, and Mladen Kolar. 2015. "ROCKET: Robust Confidence Intervals via Kendall's Tau for Transelliptical Graphical Models." *ArXiv E-Prints, arXiv:1502.07641*, February.

Biswal, Bharat B, Maarten Mennes, Xi-Nian Zuo, Suril Gohel, Clare Kelly, Steve M Smith, Christian F Beckmann, Jonathan S Adelstein, Randy L Buckner, and Stan Colcombe. 2010. "Toward Discovery Science of Human Brain Function." *Proceedings of the National Academy of Sciences* 107 (10). National Acad Sciences: 4734–9.

Cai, T. Tony, W. Liu, and X. Luo. 2011. "A Constrained ℓ_1 Minimization Approach to Sparse Precision Matrix Estimation." *J. Am. Stat. Assoc.* 106 (494): 594–607.

Chen, Mengjie, Zhao Ren, Hongyu Zhao, and Harrison H. Zhou. 2015. "Asymptotically Normal and Efficient Estimation of Covariate-Adjusted Gaussian Graphical Model." *Journal of the American Statistical Association* 0 (ja): 00–00. doi:10.1080/01621459.2015.1010039.

Danaher, P., P. Wang, and Daniela M. Witten. 2014. "The Joint Graphical Lasso for Inverse Covariance Estimation Across Multiple Classes." *J. R. Stat. Soc. B* 76 (2). Wiley-Blackwell: 373–97.