Sparse Permutation Invariant Covariance Estimation: Final Talk

David Prince

Biostat 572

dprince3@uw.edu

May 31, 2012

Electronic Journal of Statistics
Vol. 2 (2008) 494–515
ISSN: 1935-7524
DOI: 10.1214/08-EJS176

Sparse permutation invariant covariance estimation

Adam J. Rothman

University of Michigan
Ann Arbor, MI 48109-1107
e-mail: ajrothma@umich.edu

Peter J. Bickel

University of California
Berkeley, CA 94720-3860
e-mail: bickel@stat.berkeley.edu

Elizaveta Levina*

University of Michigan
Ann Arbor, MI 48109-1107
e-mail: elevina@umich.edu

Ji Zhu

University of Michigan
Ann Arbor, MI 48109-1107
e-mail: jizhu@umich.edu

Review

- Gene-gene interaction networks
- Two key questions:
  1. Which gene products are directly dependent (yes/no for each pair of genes)?
  2. What is the strength and direction of this dependence (numeric for each pair of genes)?
- High-dimensional setting, i.e. n $<<$ p
- Multivariate normality assumption (with standardization)

$$\vec{X} \sim N_p(0, \Sigma)$$

- With $\Omega = \Sigma^{-1}$, $\Omega_{i,j} = 0 \Leftrightarrow X_i$ and X_j are conditionally independent

The sparse permutation invariant covariance estimator (SPICE)

$$\hat{\Omega}_\lambda = \arg\min_{\Omega \succ 0} \{ tr(\Omega \hat{\Sigma}) - \log|\Omega| + \lambda |\Omega^-|_1 \}$$

where :

- $\Omega = \Sigma^{-1}$
- $\hat{\Sigma} = \frac{1}{n} \Sigma_{i=1}^{n} (X_i - \bar{X})(X_i - \bar{X})^T$
- $\Omega^- = \Omega - diagonal(\Omega)$
- λ is the tuning parameter

Background and Theory

- Other approaches to this problem have some shortcomings:
  - Banding is an invalid assumption in this case
  - Approaches that the shrink eigenvalues are not consistent in this setting
- With reasonable assumptions we have:

$$\|\hat{\Omega}_\lambda - \Omega_0\|_F = O_P(\sqrt{\frac{(p+s)\log p}{n}})$$

$$\|\Omega_\lambda - \Omega_0\| = O_P(\sqrt{\frac{(s+1)\log p}{n}})$$

Standardization and tuning parameter selection

- Standardization

$$\Sigma = W\Gamma W$$

  where Γ is the correlation matrix and $W = diag(\Sigma)^{\frac{1}{2}}$
- Tuning parameter selection
  - We do not generally know λ's value, so we must use data
  - bounds for λ from Friedman et al. (2007)
  - Smaller λ values induce less sparsity, bigger λ values induce more sparsity
  - Criteria used: minimizing the negative log likelihood, minimizing classification error,

Covariance matrices under consideration

Banded Covariance Structures

- $\Omega_1 : \sigma_{jk} = 0.7^{|j-k|}$
- $\Omega_2 : \omega_{jk} = I(|j-k| = 0) + 0.4 \times I(|j-k| = 1)$
  $+0.2 \times I(|j-k| = 2) + 0.2 \times I(|j-k| = 3) + 0.1 \times I(|j-k| = 4)$

Varying sparsity

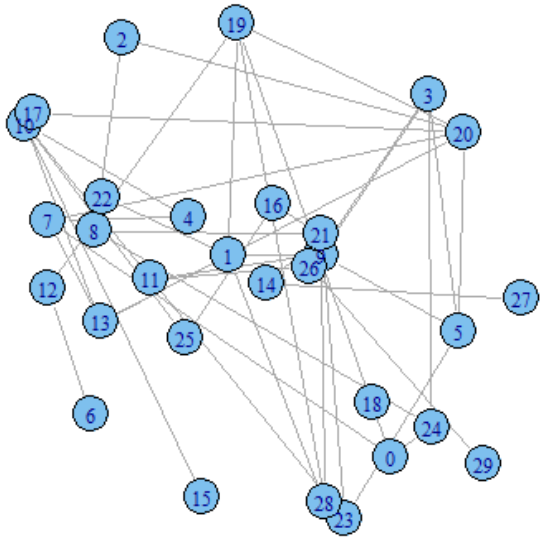
- $\Omega_3 = B + \delta I$, where $\forall\ b_{ij}, i \neq j$,

$$\begin{aligned} P(b_{ij} = .5) &= \alpha \\ P(b_{ij} = 0) &= 1 - \alpha \end{aligned}$$

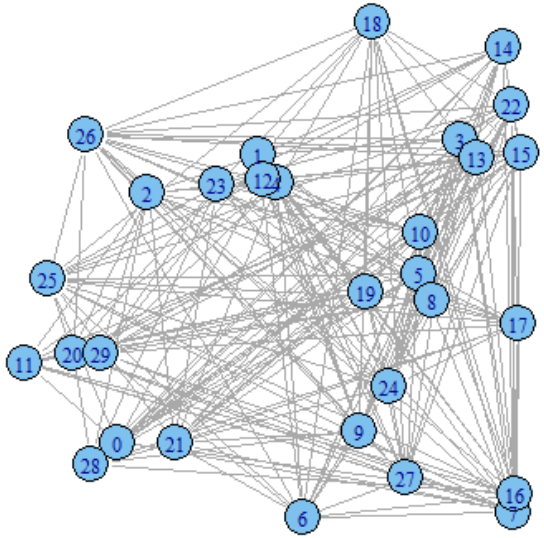
  With $\alpha = 0.1$ and δ chosen so that $\Omega_3 \succ 0$
- Ω_4: uses the same set-up as Ω_3 except $\alpha = 0.5$

Graphical model

Ω_3



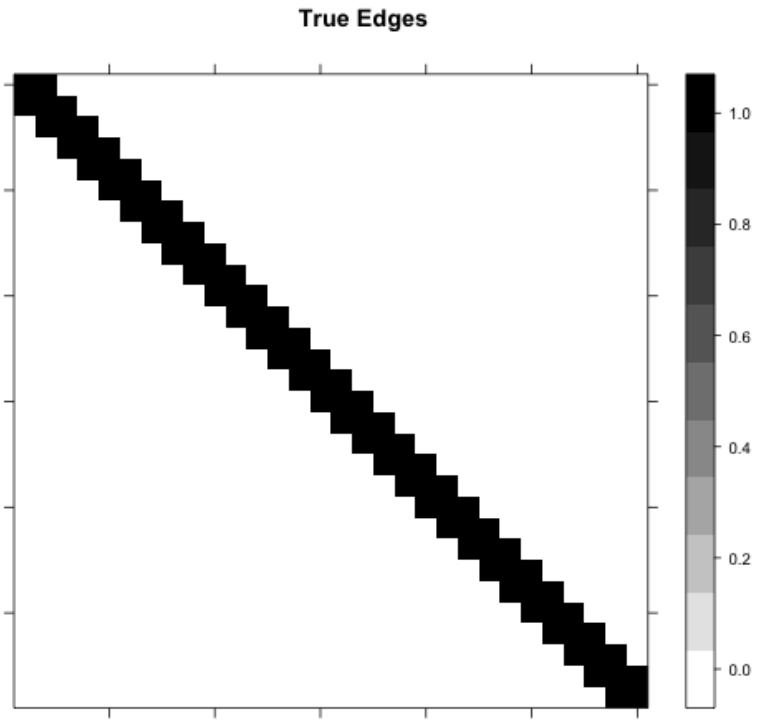
Ω_4

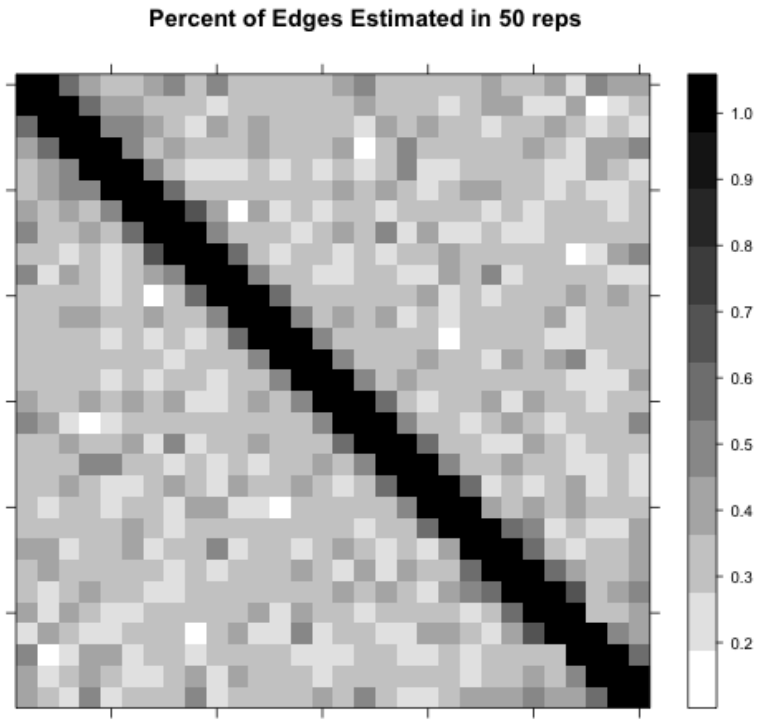


Answering question 1: which variables are conditionally dependent?

- Over repeated sampling, how does SPICE perform?
- What happens as p increases? As sparsity increases?
- Measuring whether SPICE estimates the graphical model:
  True positive rate: true non-zeros estimated as non-zero
  True negative rate: true zeros estimated as zero

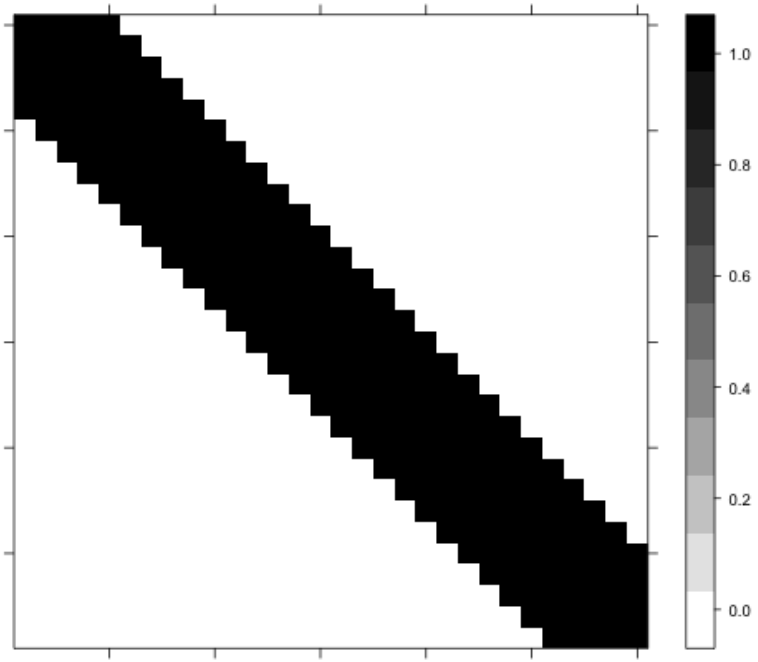
Simulation study results: Ω_1: p=30 and n=100



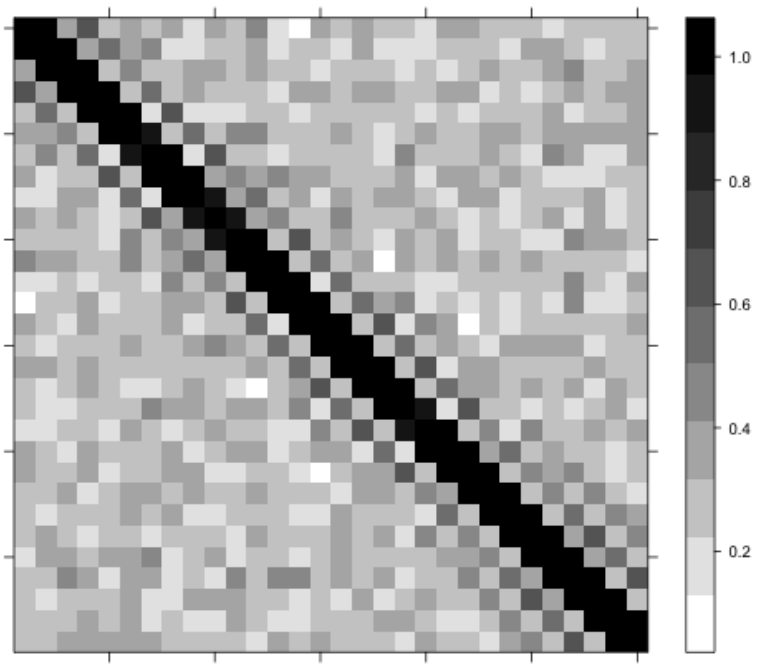


Simulation study results: Ω_2: p=30 and n=100

True edges

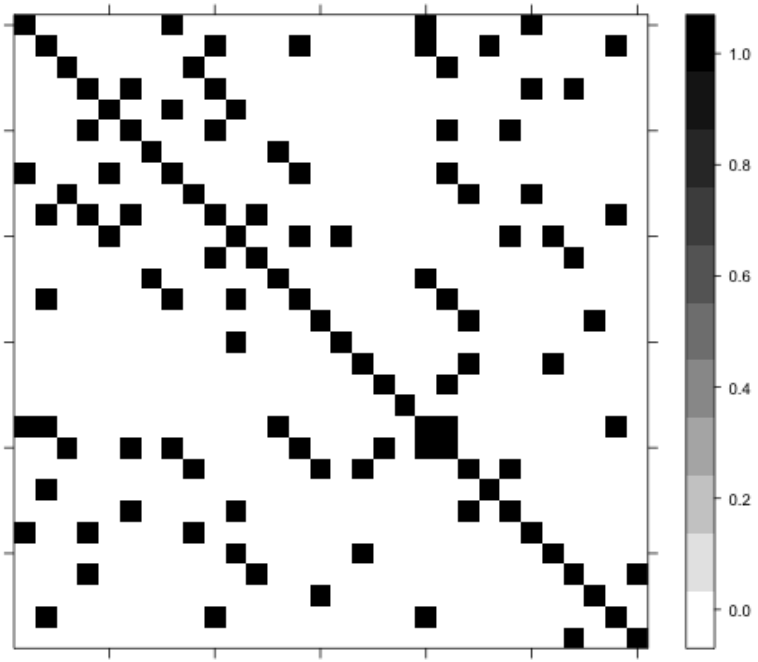


Proportion edges (50 reps)

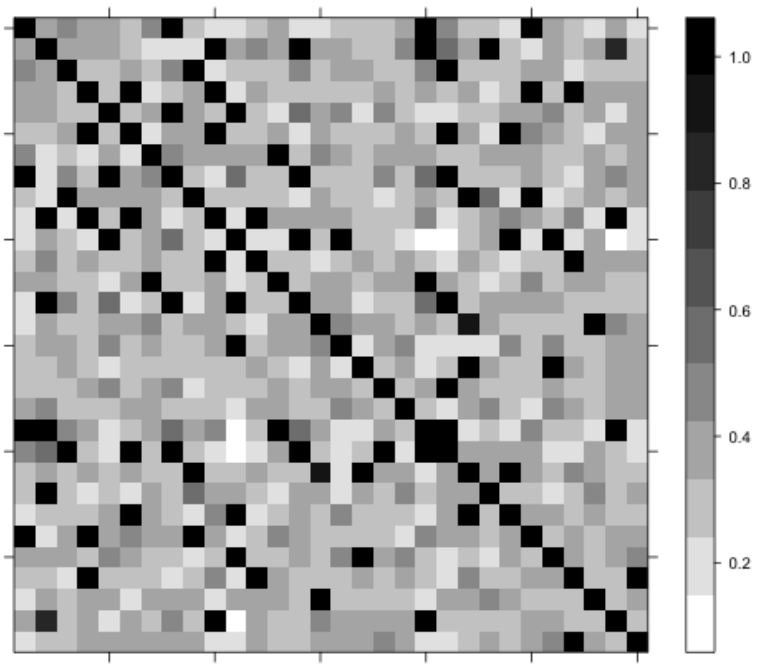


Simulation study results: Ω_3, p=30 and n=100

True edges



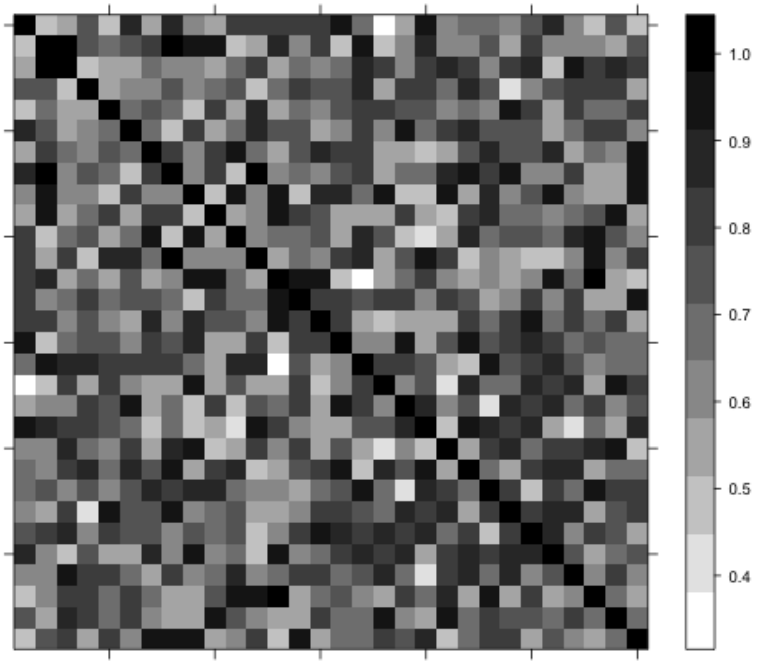
Proportion edges (50 reps)



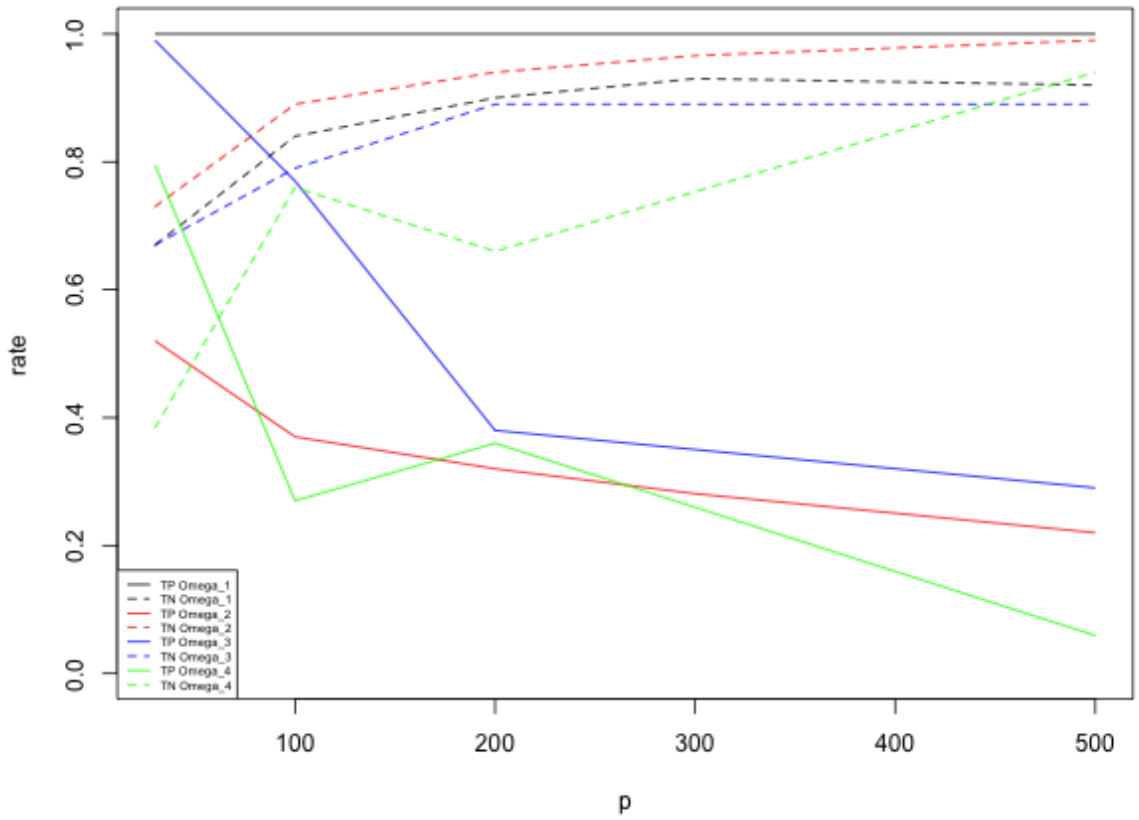
Simulation study results: Ω_4, p=30 and n=100

Truth

Proportion edges (50 reps)



High Dimensional, n=100



(TP: true non-zeros est. as non-zero; TN: true zeros est. as zero)

Summary of results in answering question 1

- Comparing Ω_1 to Ω_2, SPICE detects stronger conditional dependencies more often
- Comparing Ω_3 to Ω_4, SPICE discriminates better in the sparse setting (also consider strength of conditional dependencies)
- Noisy
- The real world: increasing p and its impact on true positive and true negative rates

Answering question 2: what is the strength and direction of the conditional dependence?

Average Kullback-Leibler Loss over 50 replications

p	LW	SPICE	LW	SPICE
	Ω_1		Ω_2	
30	3.70 (0.27)	1.69 (0.20)	2.89 (0.19)	2.53 (0.23)
100	27.63 (0.72)	8.79 (0.41)	14.07 (0.27)	10.60 (0.43)
200	79.02 (0.87)	21.82 (0.61)	31.56 (0.43)	22.89 (0.63)
300	139.41 (1.41)	36.45 (0.90)	49.89 (0.53)	35.94 (0.72)
	Ω_3		Ω_4	
30	3.45 (0.28)	1.87 (0.21)	3.27 (0.38)	3.98 (0.29)
100	19.61 (1.25)	14.83 (0.55)	16.73 (0.78)	17.76 (0.44)
200	41.25 (1.91)	37.00 (0.78)	35.77 (0.85)	66.08 (0.60)

Colon tumor classification example

- p=2,000 genes and n=62 tissue samples (40 tumorous)
- Determine the 50, 100 most discriminating genes based on expression levels
- Uses a covariance estimator in the LDA rule, for $k = 0, 1$:

$$\arg\max_k \{x^T \hat{\Omega} \hat{\mu}_k - \frac{1}{2} \hat{\mu}_k^T \hat{\Omega} \hat{\mu}_k + log \hat{\pi}_k\}$$

- Compares the classification error rates for a testing set of 20

Mean classification error percentage (SE) over 50 splits

estimator	p=50	p=100
Ledoit-Wolf	15.6 (7.8)	17.2 (5.5)
SPICE (normal)	12.1 (6.5)	18.7 (8.4)
SPICE (error)	14.7 (7.3)	16.9 (8.5)

Conclusions

- Over many replications SPICE discriminates in the sparse settings considered (Q1)
- Performs better in terms of Kullback Leibler Loss than the Ledoit-Wolf estimator in these sparse settings (Q2)
- Generalizability is uncertain; sparsity is a reasonable assumption for many biological networks