HIGH DIMENSIONAL (INVERSE) COVARIANCE MATRIX ESTIMATION

Ming Yuan

School of Industrial and Systems Engineering

Georgia Institute of Technology

myuan@isye.gatech.edu

http://www.isye.gatech.edu/~myuan

OUTLINE

- What – High dimensional covariance matrix estimation and its challenges
- How – Sparsity and graphical models
  - Estimating high dimensional inverse covariance matrix
  - Oracle inequality and adaptivity
- Examples – Gene regulatory networks; Gene set co-expression

Covariance Matrix Estimation

Classical Paradigm

- Problem setup
  - Data – a sample of n independent copies $X^{(1)}, \ldots, X^{(n)}$ of a r.v. $X \in \mathbb{R}^{d \times 1}$
  - Covariance matrix – $\mathrm{cov}(X) = \mathbb{E}((X - \mathbb{E}(X))(X - \mathbb{E}(X))^\mathsf{T})$
- Traditional Estimate
  - Sample covariance matrix

$$\hat{\Sigma}^{\text{Sample}} = \frac{1}{n-1} \sum_{i=1}^{n} (X^{(i)} - \bar{X})(X^{(i)} - \bar{X})^\mathsf{T}$$

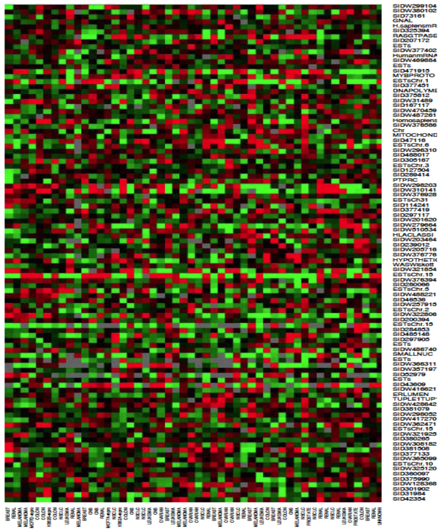
  - Maximum likelihood estimate

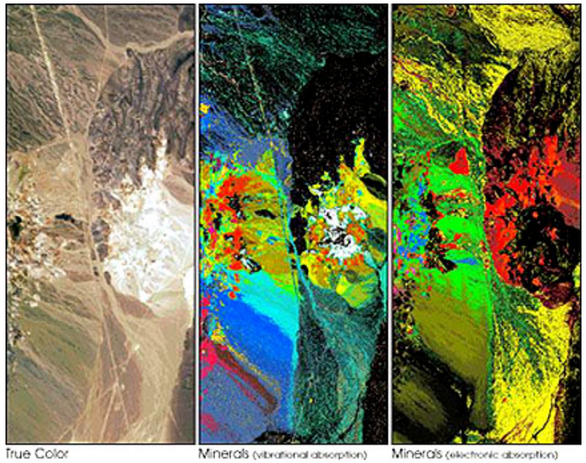
$$\hat{\Sigma}^{\text{MLE}} = \frac{1}{n} \sum_{i=1}^{n} (X^{(i)} - \bar{X})(X^{(i)} - \bar{X})^\mathsf{T}$$

- (Asymptotic) Properties
  - One of main subjects in multivariate data analysis (e.g., Anderson, 2002; Muirhead, 2005)
  - Well understood when d is fixed – Wishart distribution

HIGH DIMENSIONAL PROBLEMS

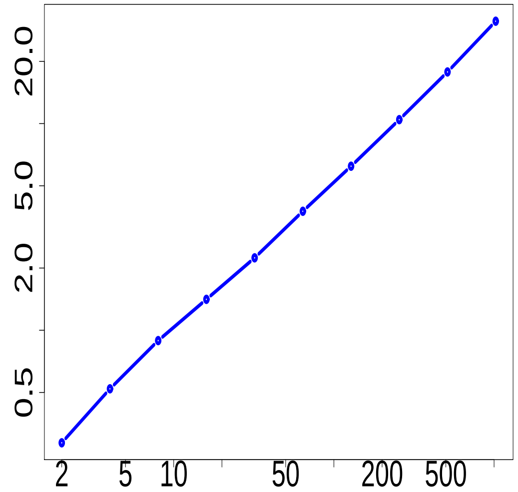
- Classical asymptotic theory: number of parameters d fixed whereas sample size $n \to \infty$
- Modern applications: both d and n may be large
  - Science – e.g., High throughput gene expression studies, $d \sim 10^4$ and $n \sim 10^2$
  - Finance – e.g., Common stocks, $d \approx 6000$ and $n \approx 200$
  - Engineering – e.g., Image analysis, Speech recognition



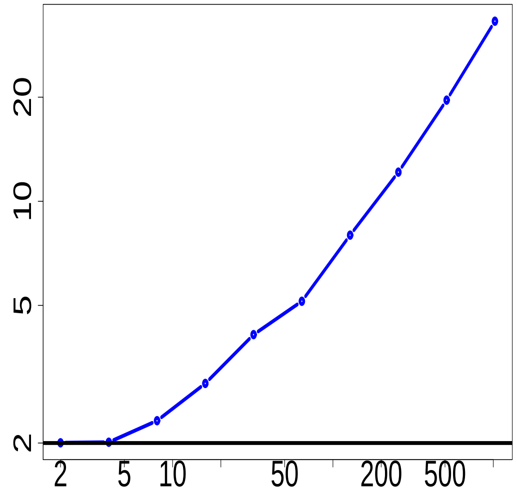


CHALLENGES OF HIGH DIMENSIONALITY

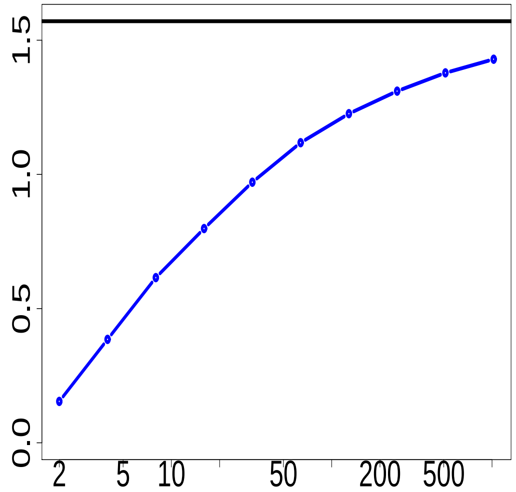
- Sample size $n = 50$
- Dimensionality $d = 2, 2^2, \ldots, 2^{10}$



(a) $\|\hat{\Sigma}^{\text{Sample}} - \Sigma\|$



(b) $\lambda_{\max}(\hat{\Sigma}^{\text{Sample}})$



(c) $\theta(\nu_1, \hat{\nu}_1)$

How to Handle High Dimensionality

- Not all problems are solvable
  - An arbitrary $d \times d$ covariance matrix involves $d(d+1)/2$ parameters
- Parameter reduction through sparsity
  - High ambient dimension; low intrinsic dimension
  - Under a certain parametrization, only a small but unknown subset of parameters are nonzero
- Sparse problems might be tractable
  - Conceptually – What kind of sparsity
  - Methodologically – How to exploit sparsity
  - Theoretically – How sparse

Sparsity in Covariance Matrices

SPARSITY TYPE – SPARSE CHOLESKY FACTORS

- One of the earliest work on sparse covariance matrix estimation (Huang et al., 2006)
- Based on modified Cholesky decomposition for time series analysis (Pourahmadi, 1999; 2000)
  - Modified Cholesky decomposition – $L\Sigma L^{\mathsf{T}} = D$
  - L is lower triangular with ones on the diagonal, D is diagonal
  - Regression interpretation

$$X_i = -\sum_{j<i} L_{ij} X_j + \epsilon_i \qquad \mathrm{cov}(\epsilon) = D$$

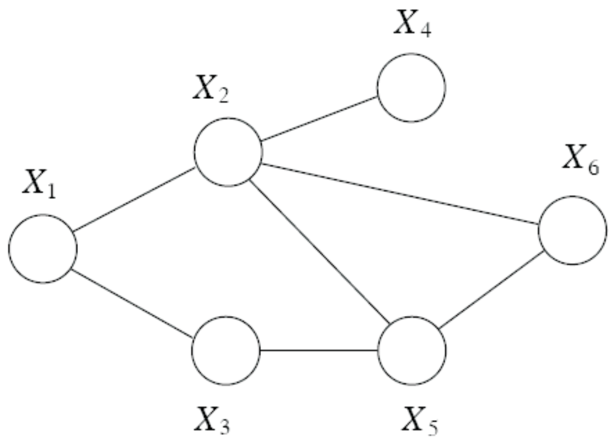
- Imposing sparsity on L – Lasso (Tibshirani, 1996) and other variants

Sparsity Type – Sparse Covariance Matrices

- Pioneered by Bickel and Levina (2008a), also motivated by time series setting
- "Bandable" covariance matrices
  - Banded covariance matrix – $\sigma_{ij} = 0$ if $|i - j| \geq k$
  - Approximately banded covariance matrix – i.e., $\sigma_{ij} \sim |i - j|^{-\alpha}$
- Most well-understood
  - Methods – banding (Bickel and Levina, 2008a), tapering (Cai, Zhang and Zhou, 2010), block thresholding (Cai and Yuan, 2011), ...
  - Theory – minimax optimality (Cai, Zhang and Zhou, 2010), adaptivity (Cai and Yuan, 2011)
  - Generalizations – covariance matrix with many zero entries (Bickel and Levina, 2008b; Cai and Zhou, 2010)

Our focus here – Sparse inverse covariance matrix

UNDIRECTED GRAPHICAL MODEL



- $X_{\mathcal{V}}$ is represented by an undirected graph $\mathcal{G}(\mathcal{V}, \mathcal{E})$
  - $\mathcal{V} = \{1, 2, 3, 4, 5, 6\}$ contains vertices corresponding to the random variables
  - the edges $\mathcal{E} = \{(1, 2), (1, 3), \ldots, (5, 6)\}$
- Factorization of probability distribution

$$p(\mathbf{x}_{\mathcal{V}}) = \psi_{12}(x_1, x_2)\psi_{13}(x_1, x_3)\psi_{24}(x_2, x_4)\psi_{25}(x_2, x_5)\psi_{26}(x_2, x_6)\psi_{35}(x_3, x_5)\psi_{56}(x_5, x_6)$$

- Conditional independence, e.g.,

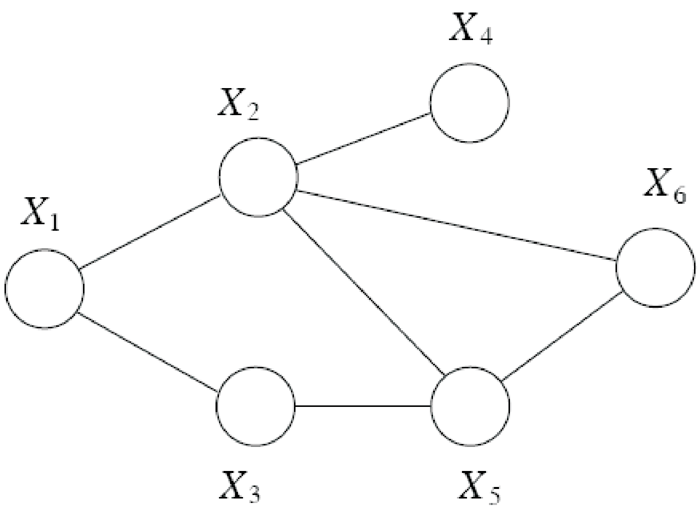
$$X_2 \perp X_3 \big| X_1, X_4, X_5, X_6$$

GAUSSIAN GRAPHICAL MODEL

- Under Normality – $X = (X_1, \ldots, X_d) \sim \mathcal{N}_d(\mu, \Sigma)$

$$\begin{aligned} p(\mathbf{x}_{\mathcal{V}}) &= (2\pi)^{-d/2} |\Sigma|^{-1/2} \exp\left\{ -\sum_{i,j} \sigma^{ij}(x_i - \mu_i)(x_j - \mu_j)/2 \right\} \\ &= (2\pi)^{-d/2} |\Sigma|^{-1/2} \prod_{(i,j):\sigma^{ij} \neq 0} \exp\left\{-\sigma^{ij}(x_i - \mu_i)(x_j - \mu_j)/2\right\} \end{aligned}$$

- Graphical model underlying X implies sparsity in the inverse covariance matrix



$$\Sigma^{-1} = \begin{bmatrix} \sigma^{11} & \sigma^{12} & \sigma^{13} & 0 & 0 & 0 \\ \sigma^{21} & \sigma^{22} & 0 & \sigma^{24} & \sigma^{25} & \sigma^{26} \\ \sigma^{31} & 0 & \sigma^{33} & 0 & \sigma^{35} & 0 \\ 0 & \sigma^{42} & 0 & \sigma^{44} & 0 & 0 \\ 0 & \sigma^{52} & \sigma^{53} & 0 & \sigma^{55} & \sigma^{56} \\ 0 & \sigma^{62} & 0 & 0 & \sigma^{65} & \sigma^{66} \end{bmatrix}$$

SPARSITY AND GRAPH

- Complexity of graphs

$$\deg(\Sigma) = \deg(\mathcal{G}) = \max_i \sum_{j \neq i} \mathbf{I}(\sigma^{ij} \neq 0)$$

- Type of sparsity
  - Sparse graph – Σ corresponds to a “low” degree graph

$$\deg(\Sigma) < s$$

  - Approximately sparse graph – Σ can be “approximated” by the first type

$$\max_{1 \leq i \leq d} \sum_{j=1}^{d} \left|\sigma^{ij}\right|^{\alpha} \leq M \qquad (0 < \alpha < 1)$$

EXPLOITING SPARSITY

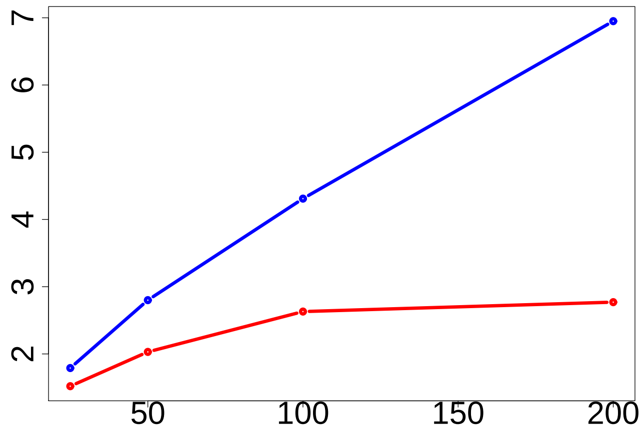
Earlier Attempt – graphical Lasso

- Penalized likelihood

$$\max_{\Sigma \succ 0} \ell(\Sigma) \qquad \text{subject to} \qquad \sum_{i<j} \mathbf{I}(\sigma^{ij} \neq 0) \leq M$$

- Convex relaxation

$$\sum_{i<j} |\sigma^{ij}| \leq M'$$



- A lot of interests since its introduction (Yuan and Lin, 2007)
- Slightly different version considered by Banerjee et al. (2008)
- Efficient algorithm proposed by Friedman et al. (2008)
- Some theory given by Ravikumar et al. (2009)
- Improves $\hat{\Sigma}^{\text{Sample}}$ but ...

Pivotal Estimator?

- Modifying an "initial" estimate
  - For covariance matrix – sample covariance matrix
  - Initial estimate has some good properties

$$\|\hat{\Sigma}^{\text{Sample}} - \Sigma\|_{\max} := \max_{i,j} \left| \hat{\sigma}_{ij}^{\text{Sample}} - \sigma_{ij} \right| = O_p\left(\sqrt{\frac{\log d}{n}} \right)$$

- What about inverse covariance matrix – $\hat{\Sigma}^{-}$? Not good

INVERSE COVARIANCE MATRIX

- Conditional distribution

$$X_1|X_{-1} \sim \mathcal{N}\left(\mu_1 + \Sigma_{1,-1}\Sigma_{-1,-1}^{-1}(X_{-1} - \mu_{-1}), \Sigma_{11} - \Sigma_{1,-1}\Sigma_{-1,-1}^{-1}\Sigma_{-1,1}\right).$$

- Inverse covariance matrix – $\Omega = \Sigma^{-1}$

$$\begin{pmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{pmatrix}^{-1} = \begin{pmatrix} \overbrace{\left(\Sigma_{11} - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21}\right)^{-1}}^{\Omega_{11}} & -\Omega_{11}\Sigma_{12}\Sigma_{22}^{-1} \\ -\Sigma_{22}^{-1}\Sigma_{21}\Omega_{11} & * \end{pmatrix}$$

- Connection

$$\begin{aligned} \mathrm{Var}(X_1|X_{-1}) &= \Omega_{11}^{-1} \\ \mathbb{E}(X_1|X_{-1}) &= \left(\mu_1 + \Sigma_{1,-1}\Sigma_{-1,-1}^{-1}(X_{-1} - \mu_{-1})\right) - X_{-1}^{\mathsf{T}}\Omega_{-1,1}/\Omega_{11} \end{aligned}$$

Multivariate Linear Regression

$$X_i | X_{-i} \sim \mathcal{N}\left(\mu_i + \Sigma_{i,-i}\Sigma_{-i,-i}^{-1}(X_{-i} - \mu_{-i}), \Sigma_{ii} - \Sigma_{i,-i}\Sigma_{-i,-i}^{-1}\Sigma_{-i,i}\right).$$

- Linear regression – $X_i \sim X_{-i}$:

$$X_i = \alpha_i + X_{-i}^{\mathsf{T}}\theta_{(i)} + e_i$$

  - Intercept

$$\alpha_i = \mu_i - \Sigma_{i,-i}\Sigma_{-i,-i}^{-1}\mu_{-i}$$

  - Coefficient

$$\theta_{(i)} = \Sigma_{-i,-i}^{-1}\Sigma_{-i,i} = -\Omega_{-i,i}/\Omega_{ii}$$

  - Variance of idiosyncratic noise

$$\mathrm{Var}(e_i) = \Sigma_{ii} - \Sigma_{i,-i}\Sigma_{-i,-i}^{-1}\Sigma_{-i,i} = \Omega_{ii}^{-1}$$

Taking advantage of Sparsity

- Translation of sparsity of Ω to regression coefficients

$$\|\theta_{(i)}\|_{\ell_0} = \|\Sigma_{-i,i}\|_{\ell_0} \leq \deg(\Omega)$$

- Exploit regression sparsity
  - Lasso (Tibshirani, 1996)

$$\left\|X_i - (\alpha + X_{-i}^{\mathsf{T}}\theta)\right\|^2 + \lambda\|\theta\|_{\ell_1} \mapsto \min$$

  - Dantzig selector (Candès and Tao, 2007)

$$\min \|\theta\|_{\ell_1} \qquad \text{subject to} \qquad \|(X_{-i} - \mu_{-i})^{\mathsf{T}}(X_i - \mu_i)\|_{\ell_\infty} \leq \delta$$

USEFUL OR NOT

- The obvious – Not working
  - Not symmetric
  - Often "dismissed" as a candidate estimate
  - May expect θ to be a good estimate, but what about Ω?
- The less obvious – Not all bad
  - $\tilde{\Omega}$ is "close" to Ω in terms of matrix ℓ_1 norm
  - Some improvement may lead to better estimates

$$\hat{\Omega} = \underset{\Omega \succeq 0}{\operatorname{argmin}} \|\Omega - \tilde{\Omega}\|_{\ell_1}$$

THEORY

Graphical Models

$$\deg(\Omega) < s$$

- Tuning

$$\delta \sim (n^{-1} \log d)^{1/2}$$

- Closeness in matrix ℓ_1 norm – with overwhelming probability

$$\sup_{\Omega_0 \in \mathcal{M}(s)} \left\| \hat{\Omega} - \Omega_0 \right\|_{\ell_1} \sim s\sqrt{\frac{\log d}{n}}$$

- Optimality

$$\inf_{\bar{\Omega}(\text{data})} \sup_{\Omega_0 \in \mathcal{M}(s)} \mathbb{E} \left\| \bar{\Omega} - \Omega_0 \right\|_{\ell_1} \geq Cs\sqrt{\frac{\log d}{n}}$$

Other Matrix Norms

- Matrix ℓ_∞ norm – $\|A\|_{\ell_\infty} = \|A\|_{\ell_1}$ for symmetric A

$$\sup_{\Omega_0 \in \mathcal{M}(s)} \left\|\hat{\Omega} - \Omega_0\right\|_{\ell_\infty} \sim s\sqrt{\frac{\log d}{n}}$$

- Bounding spectral norm – for symmetric A

$$\|A\|_{\ell_2}^2 \leq \|A\|_{\ell_1} \|A\|_{\ell_\infty} = \|A\|_{\ell_1}^2$$

  - Therefore

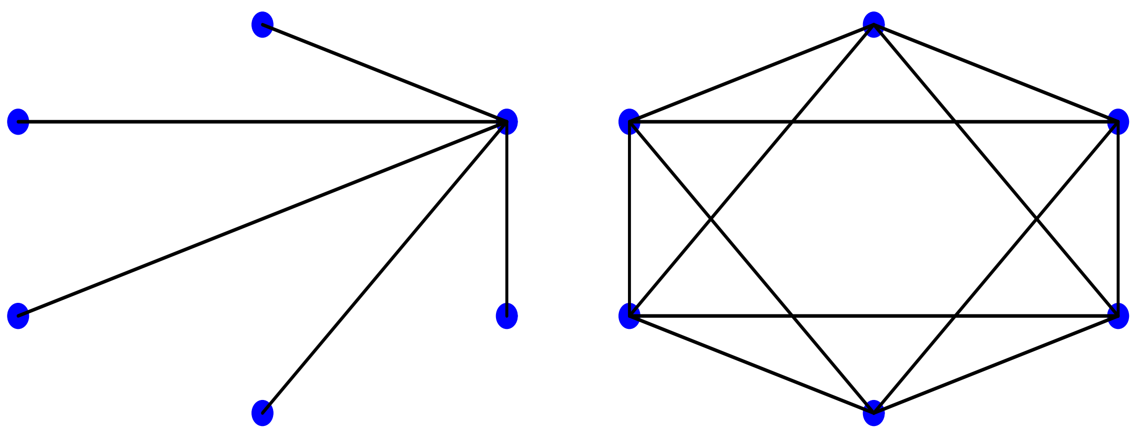
$$\sup_{\Omega_0 \in \mathcal{M}(s)} \left\|\hat{\Omega} - \Omega_0\right\|_{\ell_2} \sim s\sqrt{\frac{\log d}{n}}$$

ESTIMABILITY AND SPARSITY

- When $\deg(\mathcal{G}) = o(n^{1/2}\log^{-1/2} d)$, Ω or Σ can be "consistently" estimated

$$\|\hat{\Sigma} - \Sigma\|_{\ell_q}, \|\hat{\Omega} - \Omega\|_{\ell_q} = O_p\left(s\sqrt{\frac{\log d}{n}}\right)$$

- If $\deg(\mathcal{G}) \gg n^{1/2}\log^{-1/2} d$, Ω or Σ can not be "consistently" estimated



(a) More Difficult (b) Easier

$$n \gg s^2 \log d$$

- Impact of gene set size (d) is less significant than the connectivity (s)
- More samples are necessary if there is a "hub" gene

BEYOND GRAPHICAL MODELS

$$\left\|\hat{\Omega}-\Omega_0\right\|_{\ell_q} \leq C \inf_{\Omega}\left(\left\|\Omega-\Omega_0\right\|_{\ell_1}+\beta_n(\Omega, \delta)\right)$$

- Sparsity bound
  - If
$$\delta \sim (n^{-1} \log d)^{1/2}$$
  - Then
$$\beta_n(\Omega, \delta) = \deg(\Omega)\delta$$
- Matrix norm – ℓ_1, ℓ_2 and ℓ_∞
- Example – Take $\Omega = \Omega_0$ for graphical models

ADAPTIVITY – APPROXIMATE SPARSITY

$$\sum_{j=1}^{d} |\Omega_{ij}|^{\alpha} \leq M$$

- Construct an approximation to Ω

$$\bar{\Omega}_{ij} = \Omega_{ij} \mathbf{1} \left(|\Omega_{ij}| > \zeta \right)$$

- Tuning

$$\delta \sim \sqrt{\frac{\log d}{n}}$$

- Applying oracle inequality – matrix ℓ_1, ℓ_2 and ℓ_∞ norms

$$\sup_{\Omega_0 \in \mathcal{M}(\alpha, M)} \|\hat{\Omega} - \Omega_0\|_{\ell_q} \sim M \left(\frac{\log d}{n} \right)^{\frac{1-\alpha}{2}}$$

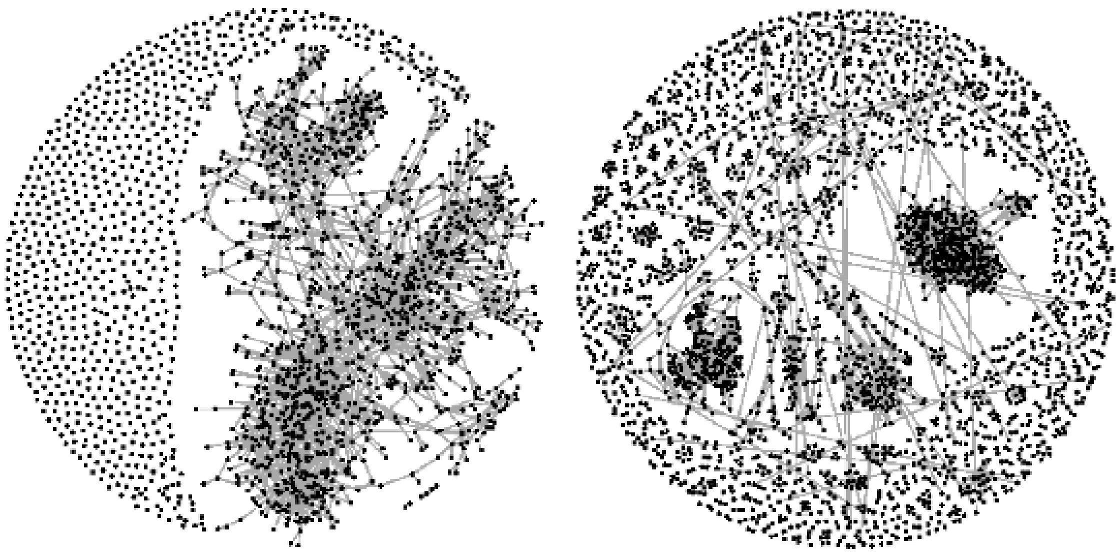
- Optimality

$$\inf_{\bar{\Omega}} \sup_{\Omega_0 \in \mathcal{M}(\alpha, M)} \mathbb{P} \left\{ \left\| \bar{\Omega} - \Omega_0 \right\|_{\ell_1} \geq CM \left(\frac{\log d}{n} \right)^{\frac{1-\alpha}{2}} \right\} > 0$$

NUMERICAL EXPERIMENTS

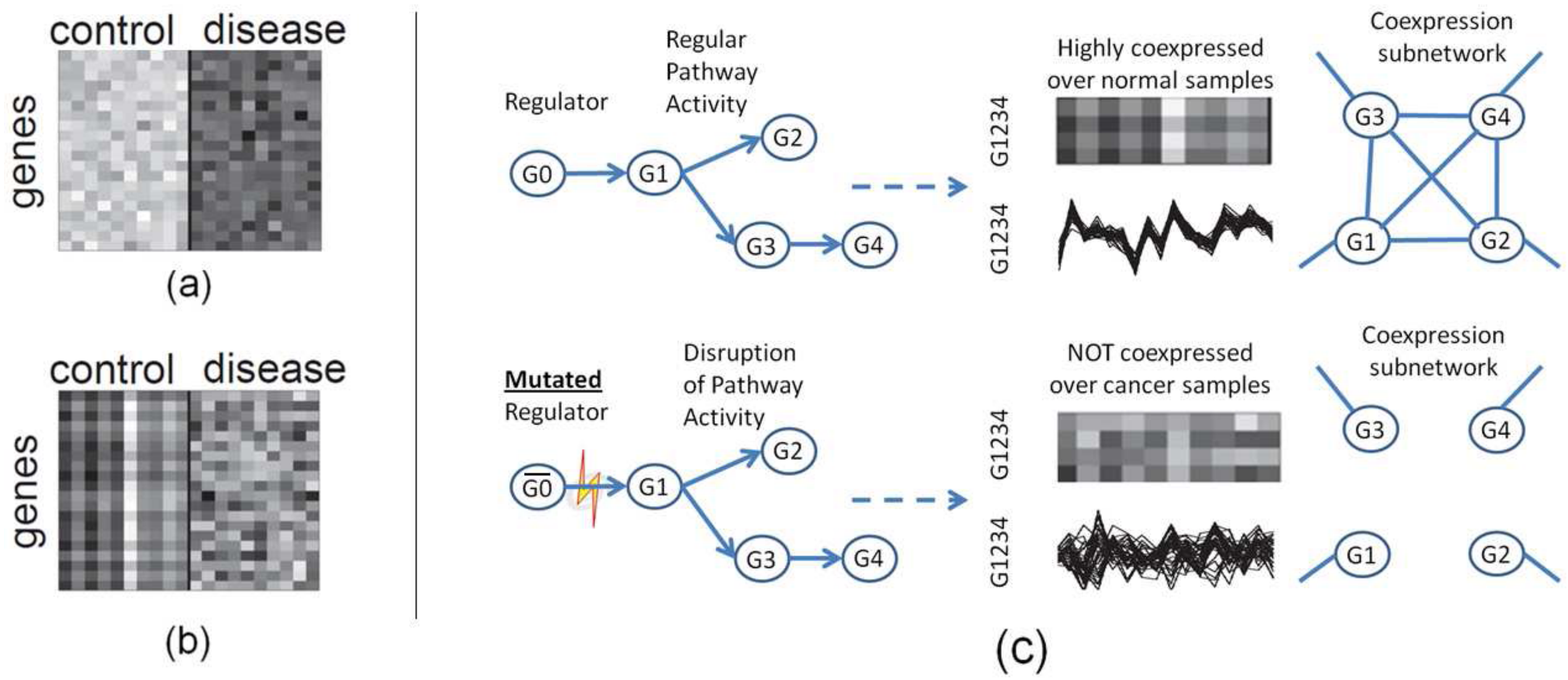
GENE/TISSUE NETWORK

- 13,182 publicly available microarray samples from Affymetrixs HGU133a platform
  - Downloaded from GEO and Array Express
  - Contains 2,717 tissue types
  - 22,283 probes $\Longrightarrow$ 12,719 genes



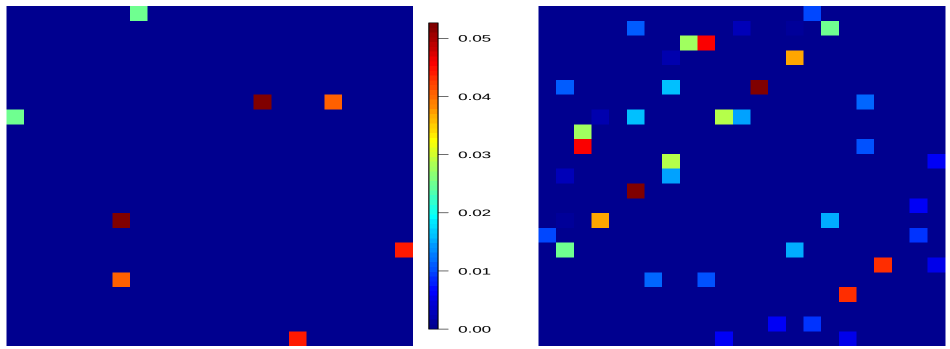
(a) Gene Expression Network (b) Tissue Network

GENE SET DIFFERENTIAL CO-EXPRESSION

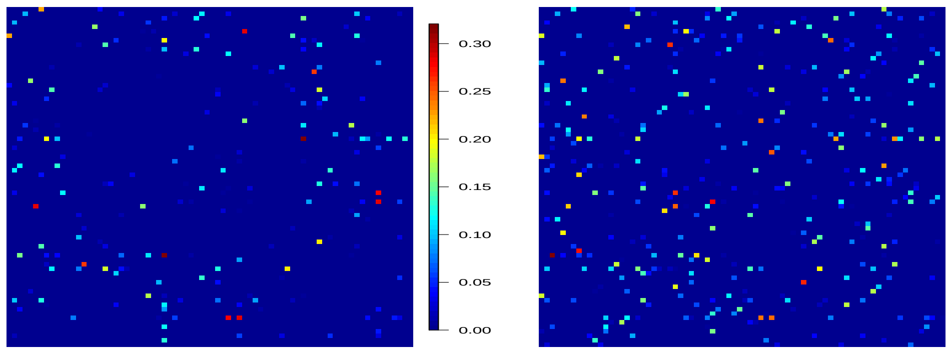


DIFFERENTIAL CO-EXPRESSION

- Lung cancer data (Beer et al., 2002)
  - Tumor tissue (86)
  - Normal tissue (44)
- Gene set definition (Choi and Kendziorski, 2009)
  - GO categories (3471)
  - KEGG pathways (178)
  - Size ranging from 3 to 3703
- Preliminary “analysis”
  - Inverse covariance matrices estimated
  - Distance in terms of spectral norm used as statistics
  - Normalized with $s(n^{-1}\log d)^{1/2}$



(a) Regulation of DNA Binding (GO:0051101; 23)



(b) Immune System Development (GO:0002520; 76)

CONCLUSIONS

- When it comes to high dimensional (inverse) covariance matrix estimation, sparse problems are more manageable
- Sparsity of covariance matrix can be exploited in multiple ways, with inverse covariance matrix connected with graphical models
- Taking advantage of the connection between multivariate normal and multivariate linear regression, a computationally feasible approach is proposed to harness sparsity in inverse covariance matrix
- The proposed approach can effectively and adaptively recover “approximately” sparse inverse covariance matrices
- Although focusing on multivariate normal, marginal subgaussianity is sufficient
- (Inverse) covariance matrix estimation is often not the ultimate goal of statistical analysis. Further research is needed in understanding its role in procedures such as PCA, LDA and etc.