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 – \( \text{cov}(X) = \mathbb{E}((X - \mathbb{E}(X))(X - \mathbb{E}(X))^T) \)

- **Traditional Estimate**
  - Sample covariance matrix
    \[
    \hat{\Sigma}_{\text{Sample}} = \frac{1}{n-1} \sum_{i=1}^{n} (X^{(i)} - \bar{X})(X^{(i)} - \bar{X})^T
    \]
  - Maximum likelihood estimate
    \[
    \hat{\Sigma}_{\text{MLE}} = \frac{1}{n} \sum_{i=1}^{n} (X^{(i)} - \bar{X})(X^{(i)} - \bar{X})^T
    \]

- **(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 $n \rightarrow \infty$

- 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 $d = 2, 2^2, \ldots, 2^{10}$
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 – $L \Sigma L^T = D$
  - $L$ is lower triangular with ones on the diagonal, $D$ is diagonal
  - Regression interpretation

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

- 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**

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

- Factorization of probability distribution

$$p(x_V) = \psi_{12}(x_1, x_2)\psi_{13}(x_1, x_3)\psi_{24}(x_2, x_4)\psi_{25}(x_2, x_5)\psi_{26}(x_2, x_6)\psi_{35}(x_3, x_5)\psi_{56}(x_5, x_6)$$

- Conditional independence, e.g.,

$$X_2 \perp X_3 | X_1, X_4, X_5, X_6$$
**GAUSSIAN GRAPHICAL MODEL**

- Under Normality – \( X = (X_1, \ldots, X_d) \sim N_d(\mu, \Sigma) \)

\[
p(x) = (2\pi)^{-d/2} |\Sigma|^{-1/2} \exp \left\{ -\sum_{i,j} \sigma^{ij}(x_i - \mu_i)(x_j - \mu_j)/2 \right\}
\]

\[
= (2\pi)^{-d/2} |\Sigma|^{-1/2} \prod_{(i,j):\sigma^{ij} \neq 0} \exp \left\{ -\sigma^{ij}(x_i - \mu_i)(x_j - \mu_j)/2 \right\}
\]

- 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

\[
\text{deg}(\Sigma) = \text{deg}(G) = \max_i \sum_{j \neq i} I(\sigma_{ij} \neq 0)
\]

- Type of sparsity
  - Sparse graph – \(\Sigma\) corresponds to a “low” degree graph
    \[
    \text{deg}(\Sigma) < s
    \]
  - Approximately sparse graph – \(\Sigma\) can be “approximated” by the first type
    \[
    \max_{1 \leq i \leq d} \sum_{j=1}^{d} |\sigma_{ij}|^\alpha \leq M \quad (0 < \alpha < 1)
    \]
EXPLOITING SPARSITY
EARLIER ATTEMPT – GRAPHICAL LASSO

- Penalized likelihood

\[
\max_{\Sigma > 0} \ell(\Sigma) \quad \text{subject to} \quad \sum_{i<j} I(\sigma_{ij} \neq 0) \leq M
\]

- 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

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

- What about inverse covariance matrix – \(\hat{\Sigma}^{-1}\)? 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 – \( \Omega = \Sigma^{-1} \)

\[
\begin{pmatrix}
\Sigma_{11} & \Sigma_{12} \\
\Sigma_{21} & \Sigma_{22}
\end{pmatrix}^{-1} = \begin{pmatrix}
\Omega_{11} \\
(\Sigma_{11} - \Sigma_{12} \Sigma_{22}^{-1} \Sigma_{21})^{-1} - \Omega_{11} \Sigma_{12} \Sigma_{22}^{-1} \\
-\Sigma_{22}^{-1} \Sigma_{21} \Omega_{11} & *
\end{pmatrix}
\]

- Connection

\[
\begin{align*}
\text{Var}(X_1 | X_{-1}) &= \Omega_{11}^{-1} \\
\mathbb{E}(X_1 | X_{-1}) &= (\mu_1 + \Sigma_{1,-1} \Sigma_{-1,-1}^{-1} (X_{-1} - \mu_{-1})) - X_{-1}^T \Omega_{-1,1} / \Omega_{11}
\end{align*}
\]
**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 – \( X_i \sim X_{-i} \):

  \[ X_i = \alpha_i + X_{-i}^\top \theta_{(i)} + e_i \]

  ▶ **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**

  \[ \text{Var}(e_i) = \Sigma_{ii} - \Sigma_{i,-i} \Sigma_{-i,-i}^{-1} \Sigma_{-i,i} = \Omega_{ii}^{-1} \]
**TAKING ADVANTAGE OF SPARSITY**

- Translation of sparsity of $\Omega$ to regression coefficients

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

- Exploit regression sparsity
  - **Lasso** (Tibshirani, 1996)

\[
\|X_i - (\alpha + X_{-i}^T\theta)\|^2 + \lambda\|\theta\|_{\ell_1} \rightarrow \min
\]

  - **Dantzig selector** (Candès and Tao, 2007)

\[
\min \|\theta\|_{\ell_1} \quad \text{subject to} \quad \|(X_{-i} - \mu_{-i})^T(X_i - \mu_i)\|_{\ell_\infty} \leq \delta
\]
USEFUL OR NOT

• The obvious – Not working
  ▶ Not symmetric
  ▶ Often “dismissed” as a candidate estimate
  ▶ May expect $\theta$ to be a good estimate, but what about $\Omega$?

• The less obvious – Not all bad
  ▶ $\tilde{\Omega}$ is “close” to $\Omega$ in terms of matrix $\ell_1$ norm
  ▶ Some improvement may lead to better estimates

$$\hat{\Omega} = \arg\min_{\Omega \succeq 0} \| \Omega - \tilde{\Omega} \|_{\ell_1}$$
THEORY


**Graphical Models**

\[ \text{deg}(\Omega) < s \]

- Tuning

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

- Closeness in matrix \( \ell_1 \) 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

\[
\inf_{\tilde{\Omega} \text{ (data)}} \sup_{\Omega_0 \in \mathcal{M}(s)} \mathbb{E} \left| \left| \tilde{\Omega} - \Omega_0 \right| \right|_{\ell_1} \geq C s \sqrt{\frac{\log d}{n}}
\]
**Other Matrix Norms**

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

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

- **Bounding spectral norm** – for symmetric $A$

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

  Therefore

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

- When \( \text{deg}(\mathcal{G}) = o(n^{1/2} \log^{-1/2} d) \), \( \Omega \) or \( \Sigma \) 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 \( \text{deg}(\mathcal{G}) \gg n^{1/2} \log^{-1/2} d \), \( \Omega \) or \( \Sigma \) can not be “consistently” estimated

\[
n \gg s^2 \log d
\]

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

(a) More Difficult

(b) Easier
BEYOND GRAPHICAL MODELS

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

- Sparsity bound
  - If
    \[ \delta \sim (n^{-1} \log d)^{1/2} \]
  - Then
    \[ \beta_n(\Omega, \delta) = \text{deg}(\Omega)\delta \]

- Matrix norm – \( \ell_1, \ell_2 \) and \( \ell_\infty \)

- Example – Take \( \Omega = \Omega_0 \) for graphical models
Adaptivity — Approximate Sparsity

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

- Construct an approximation to \( \Omega \)
  \[ \tilde{\Omega}_{ij} = \Omega_{ij} 1 (|\Omega_{ij}| > \zeta) \]

- Tuning
  \[ \delta \sim \sqrt{\frac{\log d}{n}} \]

- Applying oracle inequality – matrix \( \ell_1, \ell_2 \) and \( \ell_\infty \) norms
  \[ \sup_{\Omega_0 \in M(\alpha,M)} \|\hat{\Omega} - \Omega_0\|_{\ell_q} \sim M \left( \frac{\log d}{n} \right)^{\frac{1-\alpha}{2}} \]

- Optimality
  \[ \inf_{\Omega} \sup_{\Omega_0 \in M(\alpha,M)} P \left\{ \|\tilde{\Omega} - \Omega_0\|_{\ell_1} \geq CM \left( \frac{\log d}{n} \right)^{\frac{1-\alpha}{2}} \right\} > 0 \]
NUMERICAL EXPERIMENTS
13,182 publicly available microarray samples from Affymetrix HGU133a platform
  - Downloaded from GEO and Array Express
  - Contains 2,717 tissue types
  - 22,283 probes $\rightarrow$ 12,719 genes

(a) Gene Expression Network  (b) Tissue Network
GENE SET DIFFERENTIAL CO-EXPRESSION

(a) control vs. disease genes
(b) control vs. disease genes
(c) Highly coexpressed over normal samples
   Coexpression subnetwork
   G1234
   G1
   G2
   G3
   G4

Mutated Regulator
Disruption of Pathway Activity
G0
G1
G2
G3
G4

Regular Pathway Activity
G0
G1
G2
G3
G4

NOT coexpressed over cancer samples
Coexpression subnetwork
G1234
G1
G2
G3
G4
**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.