Extracting correlation structure from large random matrices

Alfred Hero

University of Michigan - Ann Arbor

Feb. 17, 2012
1 Background

2 Graphical models

3 Screening for hubs in graphical model

4 Conclusion
Outline

1. **Background**

2. **Graphical models**

3. **Screening for hubs in graphical model**

4. **Conclusion**
Random measurement matrix and assumptions

\[ \mathbf{X} = \begin{bmatrix}
    x_{11} & \cdots & \cdots & x_{1p} \\
    \vdots & \ddots & \ddots & \vdots \\
    x_{n1} & \cdots & \cdots & x_{np}
\end{bmatrix} = \begin{bmatrix}
    \mathbf{X}^1 \\
    \vdots \\
    \mathbf{X}^n
\end{bmatrix} = [\mathbf{X}_1, \ldots, \mathbf{X}_p] \]

Each row of \( \mathbf{X} \) is an independent realization of random vector 
\[ \mathbf{X} = [X_1, \ldots, X_p] \]

For this talk we assume:
- \( \mathbf{X} \) has multivariate Gaussian distribution (not necessary)
- \( \mathbf{X} \) has non-singular covariance matrix \( \Sigma \) (necessary)
- Either the covariance matrix or inverse covariance are sparse (necessary).

A question of interest (Q1): Are there variables \( X_k \) in \( \mathbf{X} \) that are highly correlated with many other variables? This question is surprisingly difficult to answer for small \( n \) large \( p \).
Example: spammer temporal patterns

\[ p = 10,000, \ n = 30 \]

Source: Xu, Kliger and H, Next Wave, 2010

Highly correlated spammers  
spammer correlation graph
Correlation analysis of multiple asset classes

\[ p = 25, n_1 = 80, n_2 = 60 \]

- Left: Average correlation: 0.42, percent of strong relations 33%
- Right: Average correlation: 0.3, percent of strong relations 20%

What asset classes remain connected in Q4-10 and Q1-11?

Source: “What is behind the fall cross assets correlation?” J-J Ohana, 30 mars 2011, Riskelia’s blog.
Example: Acute respiratory infection gene correlation network

\[ p = 20,000, \quad n = 8 \]
Discoveries of variables with high sample correlation

- Number of discoveries exhibit phase transition phenomenon
- This phenomenon gets worse as $p/n$ increases.
Previous work

- Regularized $l_2$ or $l_F$ covariance estimation
  - Sparse eigendecomposition model: Johnstone-Lu (2007)
- Gaussian graphical model selection
  - Bayesian estimation: Rajaratnam-Massam-Carvalho (2008)
- Independence testing
  - Sphericity test for multivariate Gaussian: Wilks (1935)

Covariance and correlation

- Covariance of $X_i, X_j$: $\sigma_{ij} = E[(X_i - E[X_i])(X_j - E[X_j])]$

- Correlation of $X_i, X_j$: $\rho_{ij} = \frac{\sigma_{ij}}{\sqrt{\sigma_{ii}\sigma_{jj}}}$

- Covariance matrix
  $\Sigma = ((\sigma_{ij}))_{i,j=1}^p = E[(X - E[X])^T(X - E[X])]$

- Correlation matrix
  $\Gamma = ((\rho_{ij}))_{i,j=1}^p = \text{diag}(\Sigma)^{-1/2}\Sigma\text{diag}(\Sigma)^{-1/2}$

Fundamental fact:
$$|\rho_{ij}| \leq 1 \text{ and } |\rho_{ij}| = 1 \text{ iff } X_i = aX_j + b.$$ with sign($a$) = sign($\rho_{ij}$)
A correlation network is an undirected graph $G$ with

- vertices $\mathcal{V} = \{X_i, \ldots, X_p\}$
- edges $\mathcal{E} = \{e_{ij} : |\rho_{ij}| > \eta\}$
- i.e., an edge $e_{ij}$ exists between $X_i, X_j$ if the magnitude correlation $|\rho_{ij}|$ exceeds a threshold $\eta$, $\eta \in [0, 1]$.

**Equivalent question (Q1):** for large $\eta$, are there highly connected nodes (hubs) in $G$?
A thresholded correlation matrix and correlation graph

\[ p = 100 \]

**Correlation screening**: find nodes that are connected.

**Hub screening**: find nodes of degree at least \( \delta \).
Outline

1. Background
2. Graphical models
3. Screening for hubs in graphical model
4. Conclusion
Sparse multivariate dependency models

Two types of sparse correlation models:

- **Sparse correlation graphical models:**
  - Most correlation are zero, few marginal dependencies
  - Examples: M-dependent processes, moving average (MA) processes

- **Sparse inverse-correlation graphical models**
  - Most inverse covariance entries are zero, few conditional dependencies
  - Examples: Markov random fields, autoregressive (AR) processes, global latent variables

- Sometimes correlation matrix and its inverse are both sparse
- Sometimes only one of them is sparse
Gaussian graphical models - GGM - (Lauritzen 1996)

Multivariate Gaussian model

\[
p(x) = \frac{|K|^{1/2}}{(2\pi)^{p/2}} \exp \left( -\frac{1}{2} \sum_{i,j=1}^{p} x_i x_j [K]_{ij} \right)
\]

where \( K = [\text{cov}(X)]^{-1} \): \( p \times p \) precision matrix

- \( G \) has an edge \( e_{ij} \) iff \( [K]_{ij} = 0 \)
- Adjacency matrix \( A \) of \( G \) obtained by hard-thresholding \( K \)

\[
A = h(K), \quad h(u) = \frac{1}{2}(\text{sgn}(|u| - \rho) + 1)
\]

\( \rho \) is arbitrary positive threshold
Partial correlation representation of GGM

Equivalent representation for $A$ is $A = h(\Gamma)$

- $\Gamma$ is partial correlation matrix

$$\Gamma = [\text{diag}(K)]^{-1/2}K[\text{diag}(K)]^{-1/2}$$

- Properties

$$||[\Gamma]_{i,j}|| \leq 1, \quad ||[\Gamma]_{i,j}|| = 0 \iff ||[K]_{i,j}|| = 0$$
Block diagonal Gaussian graphical model

**Figure:** Left: partial correlation matrix $\mathbf{A}$. Right: associated graphical model
Two coupled block Gaussian graphical model
Multiscale Gaussian graphical model
Banded Gaussian graphical model
Screening for hubs in $\mathcal{G}$

Figure: Star components - hubs of degree $d = 1, \ldots, 5, \ldots$

- Single treatment: count number of hub nodes in $\mathcal{G}$
  \[ N_d = \sum_{i=1}^{p} I(d_i \geq d) \]

- Different treatments: Count number of hub node coincidences in $\mathcal{G}^a$ and $\mathcal{G}^b$
  \[ N_d^{a\land b} = \sum_{i=1}^{p} I(d_i^a \geq d)I(d_i^b \geq d) \]
Screening hubs in $G$ from random samples

**Problem:** Find hubs in $G$ given $n$ i.i.d. samples $\{X^j\}_{j=1}^n$

**Solution:** Threshold the sample partial correlation matrix

$$P = \left[\text{diag}(R^{-1})\right]^{-1/2}R^{-1}\left[\text{diag}(R^{-1})\right]^{-1/2}$$

$R$ is the sample correlation matrix

$$R = \left( \frac{\text{cov}(X_i, X_j)}{\sqrt{\text{var}(X_i)\text{var}(X_j)}} \right)^p_{i,j=1}$$

$$= \left[\text{diag} (\text{cov}(X))\right]^{-1/2}\text{cov}(X)[\text{diag} (\text{cov}(X))]^{-1/2}$$
### Issues

#### Difficulties

- When $n < p$ sample covariance matrix $\widehat{\text{cov}}(X)$ is not invertible.
- False matches can occur at any threshold level $\rho \in [0, 1)$.
- The number of false matches abruptly increases in $p$.

#### Proposed solutions: for $n < p$

- We define a rank deficient version of partial correlation
- We derive finite sample $p$-values for the number of false matches
- We derive expressions for phase transition thresholds.
- Theory applies to both correlation graphs and concentration graphs
Z-scores

Z-scores associated with $X_i$:

$$1\hat{\sigma}_i(X_i^l - \hat{\mu}_i), \quad l = 1, \ldots, n$$

$$\hat{\mu}_i = n^{-1}\sum_{l=1}^{n} X_i^l, \quad \hat{\sigma}_i^2 = (n - 1)^{-1}\sum_{l=1}^{n} (X_i^l - \hat{\mu}_i)^2$$

Define matrix of projected Z-scores

$$U = [U_1, \ldots, U_p], \quad U_i \in S_{n-2} \subset \mathbb{R}^{n-1}$$

- Sample correlation matrix representation

$$R = U^T U, \quad r_{ij} = U_i^T U_j$$

- Sample partial correlation representation

$$P = Y^T Y, \quad Y = [UU^T]^{-1}UD^{-1/2}U[UU^T]^{-2}$$
Z-scores lie on sphere $S_{n-2}$

Correlation is related to distance between Z-scores

$$\|U_i - U_j\| = \sqrt{2(1 - r_{ij})}$$
Example: Z-scores for diagonal Gaussian

Projected Z-scores. Diagonal covariance, n=4, p=500

Diagonal covariance, p=500
Example: Z-scores for ARMA(2,2) Gaussian
Hub discoveries: define number of vertices having degree \( d_i \geq \delta \)

\[
N_{\delta,\rho} = \sum_{i=1}^{p} \phi_{\delta,i}
\]

\[
\phi_{\delta,i} = \begin{cases} 
1, & \text{card}\{j : j \neq i, |Z_i^T Z_j| \geq \rho\} \geq \delta \\
0, & \text{o.w.}
\end{cases}
\]

\[
Z_i = \begin{cases} 
U_i, & \text{correlation} \\
Y_i, & \text{partial correlation}
\end{cases}
\]
Asymptotic discovery rate

Assume that rows of $\mathbf{X}$ are i.i.d. with bounded elliptically contoured density and sparse graphical model.

**Poisson limit**: (H., Rajaratnam, 2011):

**Theorem**

*For large $p$*

$$P(N_{\delta,\rho} > 0) \approx \begin{cases} 1 - \exp(-\lambda_{\delta,\rho}/2), & \delta = 1 \\ 1 - \exp(-\lambda_{\delta,\rho}), & \delta > 1 \end{cases}.$$

$$\lambda_{\delta,\rho} = p \left( \frac{p - 1}{\delta} \right) (P_0(\rho, n))^{\delta}$$

$$P_0(\rho, n) = 2B\left(\frac{n-2}{2}, 1/2\right) \int_{\rho}^{1} (1 - u^2)^{\frac{n-4}{2}} du$$
Sparse covariance

Ellipticity and sparsity assumption guarantee universal Poisson rate

\[ E[N_{\delta,\rho}] = \lambda_{\delta,\rho} \left( 1 + O \left( (q/p)^2 \right) \right). \]

Figure: Left: row \( k \)-sparse covariance. Right: block \( k \)-sparse covariance. \( k = 10 \) and \( p = 100 \).

For non-elliptical and non-sparse case, Poisson limit holds with

\[ E[N_{\delta,\rho}] = \lambda_{\delta,\rho} \mathcal{J}(f_{\mathcal{U}}). \]
False discovery probability: \( P(N_{\delta,\rho} > 0) \approx 1 - \exp(-\lambda_{\delta,\rho}) \)

- For \( p = 10 \) and \( \delta = 1 \)
- For \( p = 10000 \) and \( \delta = 1 \)
Mean discovery rate ($\delta = 1$)

Critical threshold: $\rho_c \approx \max\{\rho : \frac{dE[N_{\delta,\rho}]}{d\rho} = -1\}$

$$\rho_c = \sqrt{1 - c_{\delta,n}(p - 1)^{-2/(n-4)}}$$
Phase transitions as function of $\delta$, $p$
Poisson convergence rates

Assume

- $\rho, n, p$ are such that $p(p - 1)^\delta (1 - \rho^2)^{(n-2)/2} = O(1)$

$$\left| P(N_{\delta, \rho} = 0) - e^{-\lambda_{\delta, \rho}} \right| \leq O \left( \max \left\{ p^{-1/\delta}, p^{1/(n-2)}, \Delta_{p,n,k,\delta} \right\} \right)$$

$\Delta_{p,n,k,\delta}$ is dependency coefficient between $\delta$-nearest-neighbors of $Z_i$ and its $p - k$ furthest neighbors
Where does Poisson convergent limit come from?

Specialize to case of $\delta = 1$:

Define

- $\phi_{ij}$: indicator function of an edge between nodes $i$ and $j$
- $N_e = \sum_{j \neq i} \phi_{ij}$: the total number of edges
- $N = \sum_{i=1}^{\rho \mapsto \infty} \max_{j: j \neq i} \phi_{ij}$: the total number of connected nodes

Key properties:

- $N$ is even integer
- $\{N = 0\} \iff \{N_e = 0\}$
- $N_e$ converges to a Poisson random variable $N^*$ with rate $\Lambda^* = E[N_e]$

\[
\lim_{p \to \infty, \rho \to 1} P(N_e = k) = \frac{(\Lambda^*)^k}{k!} e^{-\Lambda^*}, \quad k = 0, 1, \ldots
\]
Validation: correlation screening with spike-in

<table>
<thead>
<tr>
<th>( n ) ( \alpha )</th>
<th>0.010</th>
<th>0.025</th>
<th>0.050</th>
<th>0.075</th>
<th>0.100</th>
</tr>
</thead>
<tbody>
<tr>
<td>10</td>
<td>0.99/0.99</td>
<td>0.99/0.99</td>
<td>0.99/0.99</td>
<td>0.99/0.99</td>
<td>0.99/0.99</td>
</tr>
<tr>
<td>15</td>
<td>0.96/0.96</td>
<td>0.96/0.95</td>
<td>0.95/0.95</td>
<td>0.95/0.94</td>
<td>0.95/0.94</td>
</tr>
<tr>
<td>20</td>
<td>0.92/0.91</td>
<td>0.91/0.90</td>
<td>0.91/0.89</td>
<td>0.90/0.89</td>
<td>0.90/0.89</td>
</tr>
<tr>
<td>25</td>
<td>0.88/0.87</td>
<td>0.87/0.86</td>
<td>0.86/0.85</td>
<td>0.85/0.84</td>
<td>0.85/0.83</td>
</tr>
<tr>
<td>30</td>
<td>0.84/0.83</td>
<td>0.83/0.81</td>
<td>0.82/0.80</td>
<td>0.81/0.79</td>
<td>0.81/0.79</td>
</tr>
<tr>
<td>35</td>
<td>0.80/0.79</td>
<td>0.79/0.77</td>
<td>0.78/0.76</td>
<td>0.77/0.76</td>
<td>0.77/0.75</td>
</tr>
</tbody>
</table>

**Table:** Achievable limits in FPR (\( \alpha \)) for TPR = 0.8 (\( \beta \)), as function of \( n \), minimum detectable threshold, and correlation threshold (\( \rho_1/\rho \)). To obtain entries \( \rho_1/\rho \) a Poisson approximation determined \( \rho = \rho(\alpha) \) and a Fisher-Z Gaussian approximation determined \( \rho_1 = \rho_1(\beta) \). Here \( p = 1000 \) on Gaussian sample having diagonal covariance with a spike-in correlated pair.
Validation: correlation screening with spike-in

**Figure:** Comparison between predicted (diamonds) and actual (numbers) operating points \((\alpha, \beta)\) using Poisson approximation to false positive rate \((\alpha)\) and Fisher approximation to false negative rate \((\beta)\). Each number is located at an operating point determined by the sample size \(n\) ranging over \(n = 10, 15, 20, 25, 30, 35\). These numbers are color coded according to the target value of \(\beta\).
Hub screening p-value computation algorithm

- Hub screening p-value algorithm:
  - Step 1: Compute critical phase transition threshold $\rho_{c,1}$ for discovery of connected vertices
  - Step 2: Generate partial correlation graph with threshold $\rho^* > \rho_{c,1}$
  - Step 3: Compute p-values for each vertex of degree $\delta = k$ found
    \[
    p_{v_k}(i) = P(N_{k,\rho(i)} > 0) = 1 - \exp(-\lambda_{k,\rho(i,k)})
    \]
    where $\rho(i, k)$ is sample correlation between $X_i$ and its $k$-th NN.
  - Step 4: Render these p-value trajectories as a “waterfallplot”
Example: NKI gene expression dataset

Netherlands Cancer Institute (NKI) early stage breast cancer

- \( p = 24,481 \) gene probes on Affymetrix HU133 GeneChip
- 295 samples (subjects)
- Peng et al used 266 of these samples to perform covariance selection
  - They preprocessed (Cox regression) to reduce number of variables to 1,217 genes
  - They applied sparse partial correlation estimation (SPACE)
- Here we apply hub screening directly to all 24,481 gene probes
- Theory predicts phase transition threshold \( \rho_{c,1} = 0.296 \)
Mean discovery rate validation for sham NKI dataset

<table>
<thead>
<tr>
<th>observed degree</th>
<th># predicted ( (E[N_{\delta, \rho^*}]) )</th>
<th># actual ( (N_{\delta, \rho^*}) )</th>
</tr>
</thead>
<tbody>
<tr>
<td>( d_i \geq \delta = 1 )</td>
<td>8531</td>
<td>8492</td>
</tr>
<tr>
<td>( d_i \geq \delta = 2 )</td>
<td>1697</td>
<td>1635</td>
</tr>
<tr>
<td>( d_i \geq \delta = 3 )</td>
<td>234</td>
<td>229</td>
</tr>
<tr>
<td>( d_i \geq \delta = 4 )</td>
<td>24</td>
<td>28</td>
</tr>
<tr>
<td>( d_i \geq \delta = 5 )</td>
<td>2</td>
<td>4</td>
</tr>
</tbody>
</table>

**Table:** Fidelity of the predicted (mean) number of false positives and the observed number of false positives in the realization of the sham NKI dataset experiment shown in Fig. 7.
Waterfall plot of p-values for sham NKI dataset

Figure: Waterfall plot of log p-values for concentration hub screening of a sham version of the NKI dataset.
Waterfall plot of p-values for actual NKI dataset with selected discoveries shown
Waterfall plot of p-values of NKI dataset with discoveries of Peng et al shown.
Outline

1. Background
2. Graphical models
3. Screening for hubs in graphical model
4. Conclusion
Final remarks

Data-driven discovery of correlation graphs is fraught with danger.

- Sample starved regime: Number of variables $= p \gg n = \text{number of samples}$
  - Mean number of discoveries: exhibit sharp phase transition
  - Critical phase transition threshold exists
  - Poisson-type limits hold on the number of discoveries

- Study of theoretical performance limits are essential

References: