Persistent hub discovery in sparse correlation networks

Alfred Hero

University of Michigan, Ann Arbor

April 13, 2010
Acknowledgements

- Bala Rajaratnam (Stanford)
- NSF: ITR CCR-032557
- AFOSR: FA9550-06-1-0324
- ONR: N00014-08-1-1065
- ARO: W911NF-05-1-0403
- Isaac Newton Institute
- DIGITEO, Paris France
Outline

1. Motivation
2. Persistent hub discovery
3. Application
4. Conclusions
Smoker network

Social interaction network (Framingham study, NEJM 2008)

- By 2000 smokers more likely to be at periphery of their networks and in smaller subgroups than non-smokers (see dark circled areas)
- Size of circle: number of cigarettes per day
- Yellow circle: smoker
- Green circle: non-smoker
Correlation analysis of financial time series

WTI Crude vs. US Dollar

Source: Bloomberg

Economic indicators

Currency correlation
Correlation analysis of gene expression arrays

Gene expression profiles

Scatter matrix
Correlation screening and hub discovery

1. Threshold the sample correlation matrix
2. Render adjacency matrix as dependency graph
3. Identify persistent and non-persistent hubs
4. Perform cross-validation (bootstrap, leave-one-out)
Correlation screening and hub discovery

- Correlation screening finds hubs of high sample correlation
- Persistent hub discovery finds hubs surviving both treatments
- Edges shown are survivors after leave-one-out cross-validation
Correlation screening and hub discovery

- Correlation screening finds hubs of high sample correlation
Correlation screening and hub discovery

- Correlation screening finds hubs of high sample correlation
- Persistent hub discovery finds hubs surviving both treatments
Correlation screening and hub discovery

- Correlation screening finds hubs of high sample correlation
- Persistent hub discovery finds hubs surviving both treatments
- Edges shown are survivors after leave-one-out cross-validation
How much confidence can we have in such discoveries?

Confidence mitigated by

- Lack of principles for selecting correlation threshold
- Estimation of correlations between large number $p \gg n$ of variables
  - Affymetrix gene microarray chip gives $\binom{24,000}{2} = 287,988,000$ sample correlations
  - Often number of samples per treatment is less than 10
- Cross validation cannot be relied upon in these situations
How much confidence can we have in such discoveries?

Confidence mitigated by

- Lack of principles for selecting correlation threshold
- Estimation of correlations between large number $p \gg n$ of variables
  - Affymetrix gene microarray chip gives $\binom{24,000}{2} = 287,988,000$ sample correlations
  - Often number of samples per treatment is less than 10
- Cross validation cannot be relied upon in these situations

**Objective**: establish asymptotic (large $p$) theory.
Previous work

- Regularized $l_2$ or $l_F$ covariance estimation
  - Shrinkage towards identity: Ledoit-Wolf (2005)
  - Shrinkage towards banded: Bickel and Levina (2008)
  - Shrinkage towards sparse eigenvector: Johnstone and Lu (2007)

- Gaussian graphical model selection
  - $l_1$ regularized GGM: Meinshausen and Bühlmann (2006)
  - Bayesian estimation: Rajaratnam, Massam, Carvalho (2008)

- Independence testing
  - Sphericity test for multivariate Gaussian: Wilks (1935)
  - Maximal correlation test: Moran (1980)
  - Ranked correlation test: Eagleson (1983)
Previous work

- Regularized $l_2$ or $l_F$ covariance estimation
  - Shrinkage towards identity: Ledoit-Wolf (2005)
  - Shrinkage towards banded: Bickel and Levina (2008)
  - Shrinkage towards sparse eigenvector: Johnstone and Lu (2007)

- Gaussian graphical model selection
  - $l_1$ regularized GGM: Meinshausen and Bühlmann (2006)
  - Bayesian estimation: Rajaratnam, Massam, Carvalho (2008)

- Independence testing
  - Sphericity test for multivariate Gaussian: Wilks (1935)
  - Maximal correlation test: Moran (1980)
  - Ranked correlation test: Eagleson (1983)

New framework: screening, not estimation/independence testing
Previous work

- Regularized $l_2$ or $l_F$ covariance estimation
  - Shrinkage towards identity: Ledoit-Wolf (2005)
  - Shrinkage towards banded: Bickel and Levina (2008)
  - Shrinkage towards sparse eigenvector: Johnstone and Lu (2007)

- Gaussian graphical model selection
  - $l_1$ regularized GGM: Meinshausen and Bühlmann (2006)
  - Bayesian estimation: Rajaratnam, Massam, Carvalho (2008)

- Independence testing
  - Sphericity test for multivariate Gaussian: Wilks (1935)
  - Maximal correlation test: Moran (1980)
  - Ranked correlation test: Eagleson (1983)

New framework: screening, not estimation/independence testing
No particular distribution or sparsity patterns imposed
Outline

1. Motivation

2. Persistent hub discovery

3. Application

4. Conclusions
Sample correlation matrix

- $p$-variate random sample: $\mathbf{X} = [X_1, \ldots, X_p]^T$
- $p \times p$ covariance matrix (unknown): $\Sigma = E[\mathbf{X}\mathbf{X}^T]$
Sample correlation matrix

- $p$-variate random sample: $\mathbf{X} = [X_1, \ldots, X_p]^T$
- $p \times p$ covariance matrix (unknown): $\Sigma = E[\mathbf{XX}^T]$

Given $n$ i.i.d. samples $\mathbf{X} = [\mathbf{X}_1, \ldots, \mathbf{X}_n]^T$ define
Sample covariance matrix:

$$\hat{\Sigma} = \frac{1}{n-1} \sum_{i=1}^{n} (\mathbf{X}_i - \hat{\mu})(\mathbf{X}_i - \hat{\mu})^T$$

Sample correlation matrix:

$$\mathbf{R} = \hat{\mathbf{D}}^{-1/2} \hat{\Sigma} \hat{\mathbf{D}}^{-1/2}$$

where $\hat{\mathbf{D}} = \text{diag}(\hat{\Sigma})$. 
Hub screening = screening over the rows of $\mathbf{R}$

- Define $r_{ij} = (\mathbf{R})_{ij}$ and $\rho$ a user-defined threshold in $[0, 1]$
- Variable $i$ passes hub screen if: $\max_{j \neq i} |r_{ij}| > \rho$
Hub screening = screening over the rows of $\mathbf{R}$

- Define $r_{ij} = (\mathbf{R})_{ij}$ and $\rho$ a user-defined threshold in $[0, 1]$
- Variable $i$ passes hub screen if: $\max_{j \neq i} |r_{ij}| > \rho$
- *Discovered* variables have high correlation with some other variable

HSS $\rho > 0.95$ w/analytes

```
0 200 400 600 800 1000
```

\[ \text{nz} = 6520 \]
Phase transitions in hub screening

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

![Graphs showing phase transitions](image)
Overview of mathematical results

Two types of results for hub and persistent hub screening

- Characterize large $p$ phase transition and its threshold.
- Predict mean discovery rate and $p$-values.
Overview of mathematical results

Two types of results for hub and persistent hub screening

- Characterize large $p$ phase transition and its threshold.
- Predict mean discovery rate and p-values.

Main ingredients in our analysis

- Z-score embedding of sample correlation
- Geometric probability on $(n - 2)$-sphere $S_{n-2} \subset \mathbb{R}^{n-1}$
- Exchangeable process theory handles dependent variables
Z-score representation of sample correlation

- Z-score representation of correlation matrix

\[ R = ZZ^T \]

\[ Z = [Z_1, \ldots, Z_p] = (n - 1)^{-1/2}(I - n^{-1}11^T)XD^{-1/2}. \]
Z-score representation of sample correlation

- Z-score representation of correlation matrix
  \[ R = ZZ^T \]

\[ Z = [Z_1, \ldots, Z_p] = (n - 1)^{-1/2}(I - n^{-1}11^T)XD^{-1/2}. \]

- Z_i standardizes X_i by scale/translation transformation
  \[ Z_i = \frac{X_i - \hat{\mu}_i1}{\hat{s}_i\sqrt{n-1}}, \quad i = 1, \ldots, p \]

\[ \hat{\mu}_i = \frac{1}{n} \sum_{j=1}^{n} X_{ij}, \quad \hat{s}_i^2 = \frac{1}{n-1} \sum_{j=1}^{n} (X_{ij} - \hat{\mu}_j)^2 \]
Z-score representation of sample correlation

- **Z-score representation of correlation matrix**

\[
R = ZZ^T
\]

\[
Z = [Z_1, \ldots, Z_p] = (n - 1)^{-1/2}(I - n^{-1}11^T)XD^{-1/2}.
\]

- **\(Z_i\)** standardizes \(X_i\) by scale/translation transformation

\[
Z_i = \frac{X_i - \hat{\mu}_i 1}{\hat{s}_i \sqrt{n - 1}}, \quad i = 1, \ldots, p
\]

\[
\hat{\mu}_i = \frac{1}{n} \sum_{j=1}^{n} X_{ij}, \quad \hat{s}^2_i = \frac{1}{n-1} \sum_{j=1}^{n} (X_{ij} - \hat{\mu}_j)^2
\]

- \(n\)-dimensional \(Z_i\) lies in \(n - 2\) dimensional subspace

\[
1^T Z_i = 0 \text{ and } \|Z_i\| = 1
\]
Sample correlation and Z-score distances

- More convenient to work with projected Z-score $U_i \in \mathbb{R}^{n-1}$
- Sample correlation between $X_i$ and $X_j$ is equal to projected Z-score inner product

$$r_{ij} = U_i^T U_j$$
Sample correlation and Z-score distances

- More convenient to work with projected Z-score $U_i \in \mathbb{R}^{n-1}$
- Sample correlation between $X_i$ and $X_j$ is equal to projected Z-score inner product
  \[
  r_{ij} = U_i^T U_j
  \]
- This is directly related to Euclidean distance between $U_i$ and $U_j$
  \[
  \|U_i - U_j\| = \sqrt{2(1 - r_{ij})}
  \]
$S_{n-2}$ embedding example: diagonal Gaussian

Projected Z-scores. Diagonal covariance, n=4, p=500
$S_{n−2}$ embedding example: ARMA(2,2) Gaussian
Mathematical analysis

Define $\phi = [\phi_1, \ldots, \phi_p]$ the "discovery" indicator sequence:

$$\phi_i = \begin{cases} 1, & \max_{j \neq i} |r_{ij}| > \rho \\ 0, & \text{o.w.} \end{cases}$$

Define $N$ the number of discoveries:

$$N = \sum_{i=1}^{p} \phi_i$$
Define $\phi = [\phi_1, \ldots, \phi_p]$ the "discovery" indicator sequence:

$$\phi_i = \begin{cases} 1, & \max_{j \neq i} |r_{ij}| > \rho \\ 0, & \text{o.w.} \end{cases}$$

Define $N$ the number of discoveries:

$$N = \sum_{i=1}^{p} \phi_i$$

**Objective:** Find mathematical expressions for $E[N]$ as a function of $p$, $n$, $\rho$. 
Conditional expectation of $\phi_i$ has representation

$$E[\phi_i|U_i] = P(\bigcup_{j \neq i} U_j \in C_\rho, u_i \cup C_\rho, -u_i | U_i)$$
Given $U_i$ define the binary sequence $b = [b_1, \ldots, b_{p-1}]$

$$b_i = \begin{cases} 
1, & U_j \in C_{u_i} \cup C_{-u_i} \\
0, & \text{o.w.}
\end{cases}$$

Then, have equivalent representation

$$E[\phi_i | U_i] = P\left(\sum_{i=1}^{p-1} b_i > 0 | U_i\right)$$
Given $\mathbf{U}_i$ define the binary sequence $\mathbf{b} = [b_1, \ldots, b_{p-1}]$

$$b_i = \begin{cases} 1, & \mathbf{U}_j \in C_{\rho, \mathbf{u}_i} \cup C_{\rho, -\mathbf{u}_i} \\ 0, & \text{o.w.} \end{cases}$$

Then, have equivalent representation

$$E[\phi_i | \mathbf{U}_i] = P\left(\sum_{i=1}^{p-1} b_i > 0 | \mathbf{U}_i \right)$$

Classical result [Thm. 4.5.4] {TW Anderson, 2003}:

**Lemma**

*Let $\mathbf{X}$ be a $p$-variate elliptical vector with diagonal dispersion matrix $\Sigma$. The Z-scores $\{\mathbf{U}_i\}_{i=1}^p$ are i.i.d. random vectors uniformly distributed on $S_{n-2}$.*
Main result: hub screening

**Proposition**

Let the $n \times p$ data matrix $\mathbf{X}$ have i.i.d. rows but possibly dependent columns. Let the sequence $\{\rho_p\}_p$ of correlation thresholds be such that $\rho_p \to 1$ and $p(p - 1)(1 - \rho_p^2)^{(n-2)/2} \to d_n$ for some finite constant $d_n$. Then the mean number of hubs satisfies

$$\lim_{p \to \infty} E[\mathcal{N}] = \kappa_n J(f_{\mathbf{U_*}, \mathbf{U_*^*}}),$$

**where $\kappa_n = a_n d_n/2$ and $f_{\mathbf{U_*}, \mathbf{U_*^*}}$ is the following average of the pairwise Z-score density**

$$f_{\mathbf{U_*}, \mathbf{U_*^*}}(\mathbf{u}, \mathbf{v}) = \frac{1}{p} \sum_{i=1}^{p} \frac{1}{p-1} \sum_{j \neq i}^{p} \left( \frac{1}{2} f_{\mathbf{U_i}, \mathbf{U_j}}(\mathbf{u}, \mathbf{v}) + \frac{1}{2} f_{\mathbf{U_i}, \mathbf{U_j}}(\mathbf{u}, -\mathbf{v}) \right).$$
Implication: uniform density is minimax for hub screening

- Factor $J(f_{\mathbf{u}}, f_{\mathbf{v}})$ as Hellinger divergence, Rényi entropy

$$J(f_{\mathbf{u}}, f_{\mathbf{v}}) = |S_{n-2}| \int (f_{\mathbf{v}}(w|w)f_{\mathbf{u}}(w|w))^{1/2} (f_{\mathbf{u}}(w)f_{\mathbf{v}}(w))^{1/2} dw$$

$$\leq |S_{n-2}| \left( \int f_{\mathbf{v}}(w|w)f_{\mathbf{u}}(w|w) \right)^{1/2} \left( \int f_{\mathbf{u}}(w)f_{\mathbf{v}}(w) \right)^{1/2}$$

$$\leq H_2^{1/4}(f_{\mathbf{u}}|\mathbf{v})H_2^{1/4}(f_{\mathbf{v}}|\mathbf{u})H_2^{1/4}(f_{\mathbf{u}})H_2^{1/4}(f_{\mathbf{v}}), \quad (3)$$

- Equalities iff $f_{\mathbf{u}, \mathbf{v}}(u, u) = f_{\mathbf{u}}(u)f_{\mathbf{v}}(u)$ and $f_{\mathbf{u}}(u) = f_{\mathbf{v}}(u)$
- Right side of (3) minimized when $f_{\mathbf{u}}$ is uniform over $S_{n-2}$. 
Implication: phase transition for hub screening

$$M(\rho, n, p)$$

$$\rho_c = \sqrt{1 - cn (p - 1)^{-2}}$$

Critical threshold approximation:

$$\rho_c = \text{max} \{ \rho : \frac{dE[N]}{d\rho} = -1 \}$$
Critical threshold approximation: $\rho_c = \max\{\rho : dE[N]/d\rho = -1\}$

\[
\rho_c = \sqrt{1 - c_n(p - 1)^{-2/(n-4)}}
\]  \hspace{1cm} (4)
Main result: persistent hub screening

Proposition

Let the $n_a \times p$ and $n_b \times p$ data matrices $X^b$ and $X^a$ be independent. Let $\rho_p^a \to 1$ and $\rho_p^b \to 1$ while

$$p^{1/2}(p - 1) \left(1 - (\rho_p^l)^2\right)^{(n_l-2)/2} \to d_{n_l}$$

for $l = a, b$. The mean number $N^{a\wedge b}$ of persistent hubs:

$$\lim_{p \to \infty} E[N^{a\wedge b}] = \kappa_{n}^{a\wedge b} \frac{1}{p} \sum_{i=1}^{p} J(f_{U_i^a, U_{*i}^a}) J(f_{U_i^b, U_{*i}^b}), \quad (5)$$

where $\kappa_{n}^{a\wedge b} = d_{n_a} d_{n_b} a_{n_a} a_{n_b} / 4$ and, for $U \in \{U^a, U^b\}$,

$$f_{U_i, U_{*i}}(u, v) = \frac{1}{p - 1} \sum_{j \neq i} \left( \frac{1}{2} f_{U_i, U_j}(u, v) + \frac{1}{2} f_{U_i, U_j}(u, -v) \right). \quad (6)$$
Implication: phase transition for persistent hub screening
Phase transitions: hub vs persistent hub screening
Consider the average pairwise dependency coefficient

\[ \Delta_p = \sup_{u_1, u_2, u_3} \left| \frac{1}{k, l, i} \left[ f_{U_k, U_l | U_i}(u_1, u_2 | u_3) - f_{U_k | U_i}(u_1 | u_3) f_{U_l | U_i}(u_2 | u_3) \right] \right| \] (7)

Properties:
- \( \Delta_p = 0 \) if diagonal covariance
- \( \Delta_p = O(q/p) \) if q-sparse covariance

Proposition

If the joint density of the Z-scores satisfies \( \Delta_p \to 0 \) then \( N, N^{a \wedge b} \) are asymptotically Poisson with rate given limits (1) and (5).
## Application: hub discovery

<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 (\(\alpha\)), TPR (\(\beta\)), and minimum detectable correlation threshold (\(\rho_1\)). To obtain entries \(\rho_1\ \rho\) used Poisson approximation of Proposition for \(\alpha\), \(\rho\) and Fisher-Z Gaussian approximation for \(\beta\), \(\rho_1\). Here \(p = 1000\) and \(\beta = 0.8\) on Gaussian sample having diagonal covariance.
**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 \(\beta\).
Application: persistent hub discovery

**Figure:** Comparison between predicted (diamonds) and actual (numbers) operating points ($\alpha, \beta$) for persistent hub screening.
Beverage Data from Gene Expression Omnibus (GEO) NCBI

- Subjects: 6 individuals at 5 time points (0 to 12 hours)
- Treatments: intake of:
  - A: alcohol ($n_1 = 20$)
  - G: grape juice ($n_2 = 22$)
  - H: water ($n_3 = 23$)
  - W: red wine ($n_4 = 22$)
- 108 Affymetrix HU133 Genechip peripheral blood samples
- Each sample contains $p = 22,283$ gene probes
Application: observed Z-scores

Figure: 3 dimensional projections of the Z-scores for the experimental beverage data under each of the treatments A,G,H,W. For visualization the 22,238 variables (gene probes) were downsampled by a factor of 8 and a randomly selected set of four samples in each treatment were used to produce these figures.
Application: persistent correlation discoveries

<table>
<thead>
<tr>
<th></th>
<th>42</th>
<th>50</th>
<th>82</th>
<th>424</th>
</tr>
</thead>
<tbody>
<tr>
<td>{A, G, H, W}</td>
<td>493</td>
<td>748</td>
<td>1069</td>
<td>677</td>
</tr>
<tr>
<td>{A, G, H}</td>
<td>3313</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

**Table:** Number of genes discovered by auto-screening (top row) and persistency screening (lower three rows) for various combinations of treatments in the experimental data. Auto-screening threshold determined using Poisson approximation to Type I error of level $10^{-5}$. 
**Application: set-inclusion diagram**

**Figure:** Set-inclusion graph between genes discovered by correlation screening in various combinations of treatments. Size of node is proportional to the log of number of associated correlation screening discoveries given in Table 2. A directed edge from node $i$ to node $j$ exists if at least 90% of the genes discovered in node $i$ are also discovered in node $j$ and the thickest edges indicate 100% set inclusion. The asymmetry of diagram indicates that treatments have different effects on gene expression.
Figure: Heatmap of 4444 genes discovered in at least one of the set inclusion tests shown in Table 2.
Application: persistent covariance network

Figure: 774 gene subnetwork of the 3313 gene persistent-correlation network across all four treatments corresponding to the last row of Table 2. Two nodes in this network are linked by an edge if for all 4 treatments their sample correlation is above the $10^{-5}$ FWER correlation-screening threshold.
Outline

1. Motivation
2. Persistent hub discovery
3. Application
4. Conclusions
Conclusions

- Correlation hub screening is important in applications
Conclusions

- Correlation hub screening is important in applications
- Screening negatively affected by false positive phase transition as function of threshold
Conclusions

- Correlation hub screening is important in applications
- Screening negatively affected by false positive phase transition as function of threshold
- Large $p$ expressions for critical PT threshold $\rho_c$ is available
Conclusions

- Correlation hub screening is important in applications
- Screening negatively affected by false positive phase transition as function of threshold
- Large $p$ expressions for critical PT threshold $\rho_c$ is available
- Effect of dependency on phase transitions is mediated by Rényi 2-entropy of average marginal density on sphere
Correlation hub screening is important in applications

Screening negatively affected by false positive phase transition as function of threshold

Large $p$ expressions for critical PT threshold $\rho_c$ is available

Effect of dependency on phase transitions is mediated by Rényi 2-entropy of average marginal density on sphere

Key concepts:
- Stochastic representation of sample correlation on sphere
- Exchangeable processes
Conclusions

- Correlation hub screening is important in applications
- Screening negatively affected by false positive phase transition as function of threshold
- Large $p$ expressions for critical PT threshold $\rho_c$ is available
- Effect of dependency on phase transitions is mediated by Rényi 2-entropy of average marginal density on sphere
- Key concepts:
  - Stochastic representation of sample correlation on sphere
  - Exchangeable processes
- This work has been submitted to Annals of Statistics.