Correlation mining in massive data

Alfred Hero

University of Michigan - Ann Arbor

April 11, 2013

Outline

Correlation mining

The objective of **correlation mining** is to discover interesting or unusual patterns of dependency among a large number of variables (sequences, signals, images, videos).

Related to:

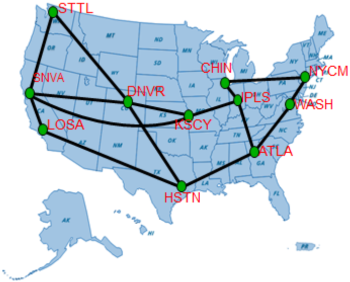
- Pattern mining, anomaly detection, cluster analysis
- Graph analytics, community detection, node/edge analysis
- Gaussian Graphical models (GGM) - Lauritzen 1996

"Big Data" aspects:

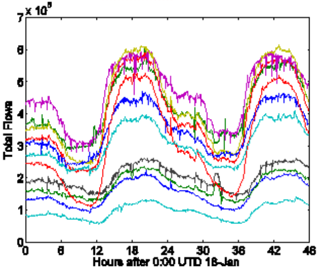
- Large numbers of signals, images, videos
- Observed correlations between signals are incomplete and noisy
- Number of samples $\ll$ number of objects of interest

Correlation mining for Internetwork anomaly detection

11 node Abilene network



11 x 576 NetFlow measurements



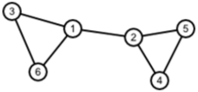
6x6 spatial correlation matrix

$$\mathbf{R} = \begin{bmatrix} 8 & 2 & -2 & 0.1 & 0.2 & 3 \\ 2 & 8 & 0.1 & 2 & -3 & 0.1 \\ -2 & 0.1 & 8 & 0.4 & 0.1 & 4 \\ 0.1 & 2 & 0.4 & 8 & 1.1 & 0.2 \\ 0.2 & -3 & 0.1 & 1.1 & 8 & 0.5 \\ 3 & 0.1 & 4 & 0.2 & 0.5 & 8 \end{bmatrix}$$

6x6 adjacency matrix

$$\mathbf{A} = \begin{bmatrix} 0 & 1 & 1 & 0 & 0 & 1 \\ 1 & 0 & 0 & 1 & 1 & 0 \\ 1 & 0 & 0 & 0 & 0 & 1 \\ 0 & 1 & 0 & 0 & 1 & 0 \\ 0 & 1 & 0 & 1 & 0 & 0 \\ 1 & 0 & 1 & 0 & 0 & 0 \end{bmatrix}$$

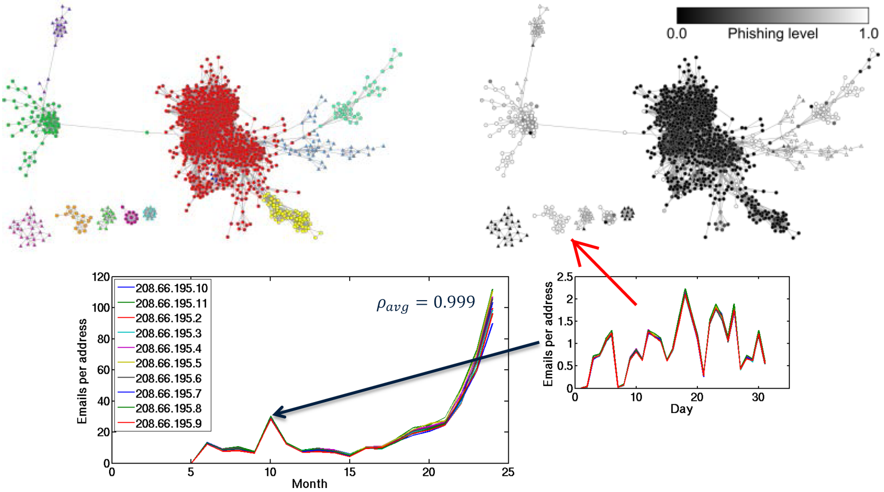
6 node correlation graph



Patwari, H and Pacholski, "Manifold learning visualization of network traffic data," SIGCOMM 2005

Correlation mining for SPAM community detection

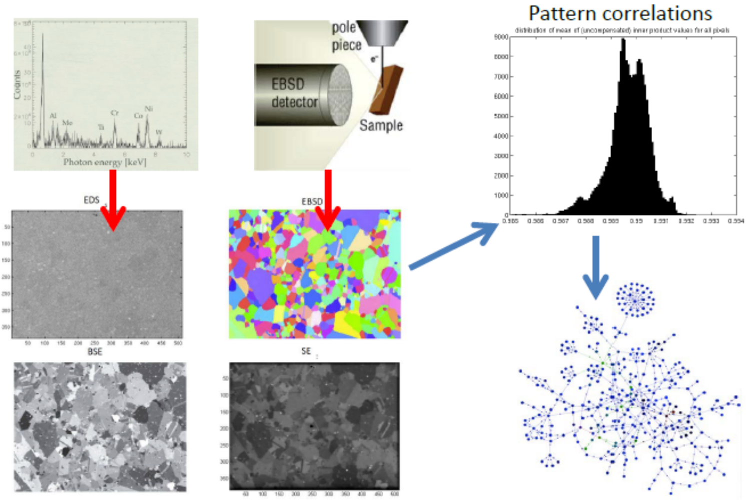
$p = 100,000, \; n = 30$



K. S. Xu et al. Revealing social networks of spammers through spectral clustering. *Proc. ICC*, 2009.

Correlation mining for materials science

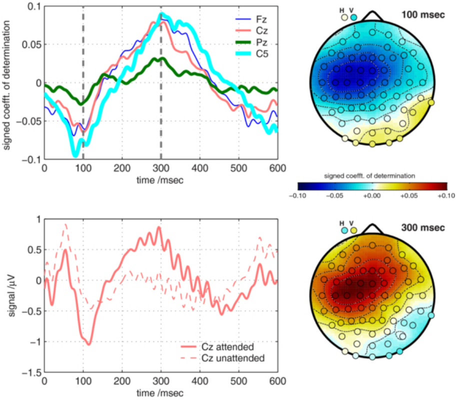
$p = 1,000,000,000, \; n = 1000$ to $100,000$

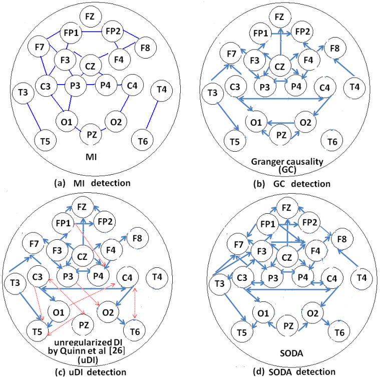


Park, Wei, DeGraef, Shah, Simmons and H, "EBSD image segmentation using a physics-based model," submitted

Correlation mining for neuroscience

$p = 100, \; n_1 = 50, \; n_2 = 50$

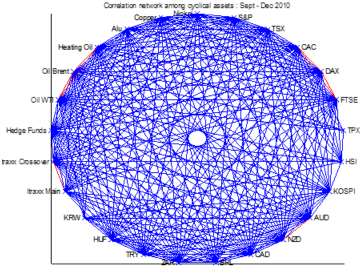


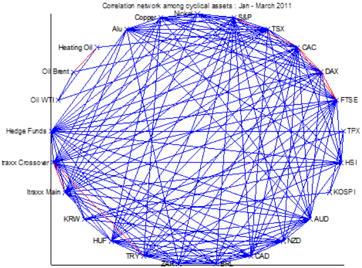


Xu, Syed and H, “EEG spatial decoding with shrinkage optimized directed information assessment,” ICASSP 2012

Correlation mining for finance

$p = 40,000, n_1 = 60, n_2 = 80$





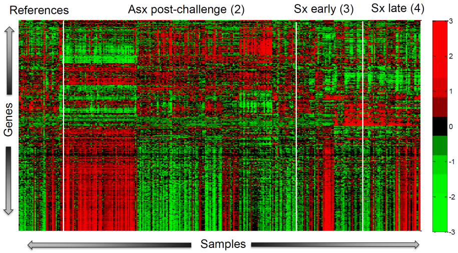
Source: "What is behind the fall in cross assets correlation?" J-J Ohana, 30 mars 2011, Riskelia's blog.

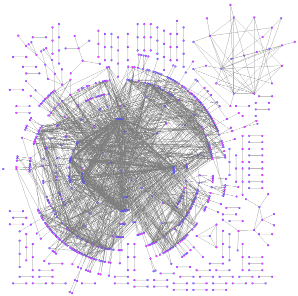
- Left: Average correlation: 0.42, percent of strong relations 33%
- Right: Average correlation: 0.3, percent of strong relations 20%

Hubs of high correlation influence the market. What hubs changed or persisted in Q4-10 and Q1-11?

Correlation mining for biology: gene-gene network

$p = 24,000, \ n = 270$





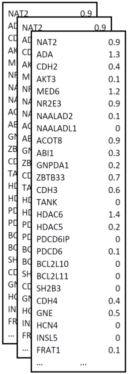
Source: Huang, . . ., and H, PLoS Genetics, 2011

Gene expression correlation graph

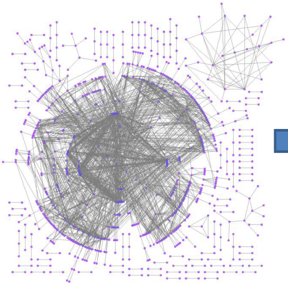
Q: What genes are "hubs" in this correlation graph?

Correlation mining for biology: gene-gene network

Experiment
- p probes
- n replicates

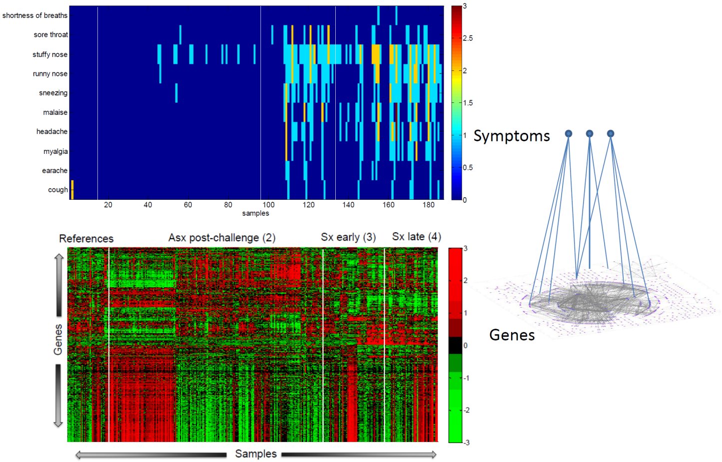


NAT2	0.9
ADA	1.3
CDH2	0.4
AKT3	0.1
MED6	1.2
NR2E3	0.9
NAALAD2	0.1
NAALADL1	0
ACOT8	0.9
ABI1	0.3
GNPDA1	0.2
ZBTB33	0.7
CDH3	0.6
TANK	0
HDAC6	1.4
HDAC5	0.2
PDCD6IP	0
PDCD6	0.1
BCL2L10	0
BCL2L11	0
SH2B3	0
CDH4	0.4
GNE	0.5
HCN4	0
INSL5	0
FRAT1	0.1
...	...



Sub-graph	d	$N_{=d}$	$N_{\geq d}$
	1	29	39
	2	9	20
	3	0	1
	4	1	1
	5	0	0
...	...	...	...

Correlation mining for predictive medicine: bipartite graph



Q: What genes are predictive of certain symptom combinations?

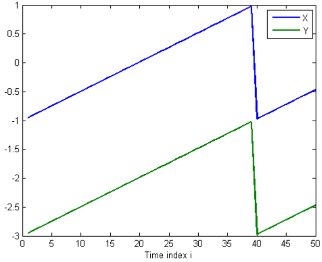
Firouzi, Rajaratnam and H, "Predictive correlation screening," AISTATS 2013

Outline

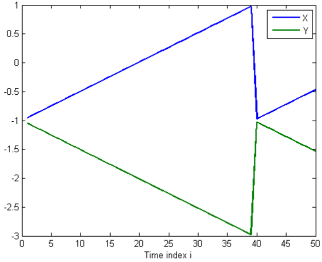
Sample correlation: $p = 2$ variables $n = 50$ samples

Sample correlation:

$$\widehat{\text{corr}}_{X,Y} = \frac{\sum_{i=1}^{n}(X_i - \overline{X})(Y_i - \overline{Y})}{\sqrt{\sum_{i=1}^{n}(X_i - \overline{X})^2 \ \sum_{i=1}^{n}(Y_i - \overline{Y})^2}} \in [-1, 1]$$



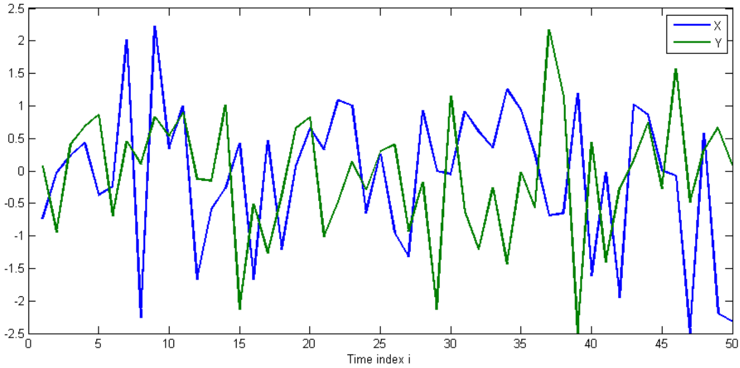
,



Positive correlation =1

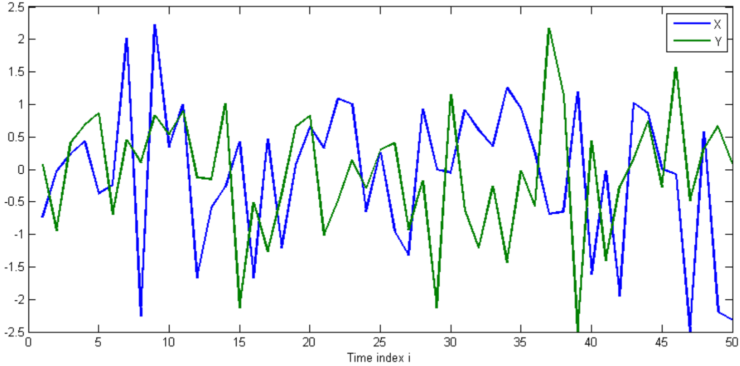
Negative correlation =-1

Sample correlation for random sequences: $p = 2$, $n = 50$



Q: Are the two time sequences X_i and Y_j correlated, e.g. $|\widehat{\text{corr}}_{XY}| > 0.5$?

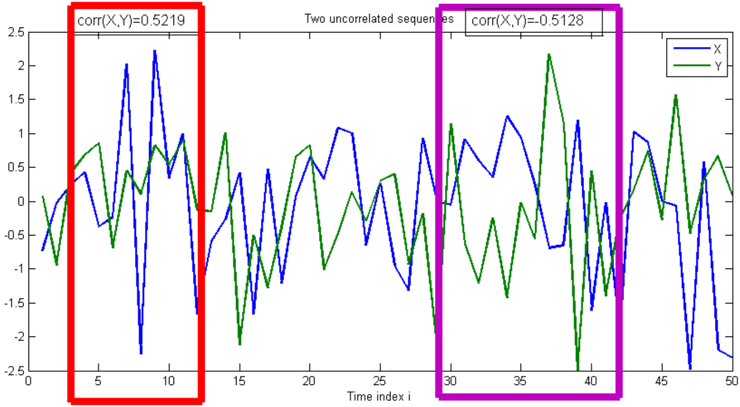
Sample correlation for random sequences: $p = 2$, $n = 50$



Q: Are the two time sequences X_i and Y_j correlated?

A: No. Computed over range $i = 1, \ldots 50$: $\widehat{\text{corr}}_{XY} = -0.0809$

Sample correlation for random sequences: $p = 2$, $n < 15$



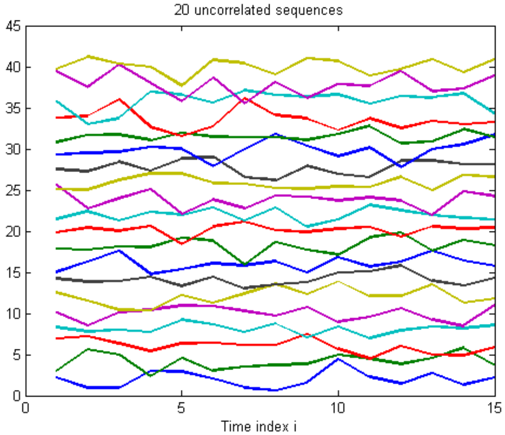
Q: Are the two time sequences X_i and Y_j correlated?

A: Yes. $\widehat{\text{corr}}_{XY} > 0.5$ over range $i = 3, \ldots 12$ and $\widehat{\text{corr}}_{XY} < -0.5$ over range $i = 29, \ldots, 42$.

Real-world example: reported correlation divergence

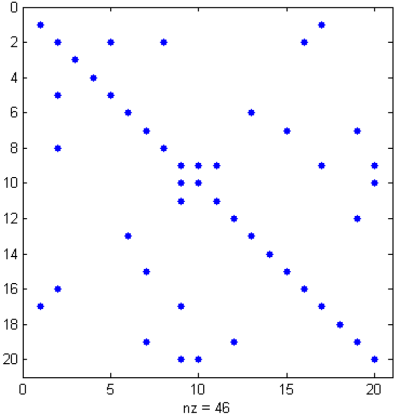
Source: FuturesMag.com `www.futuresmag.com/.../Dom%20FEB%2024.JPG`

Correlating a set of $p = 20$ sequences



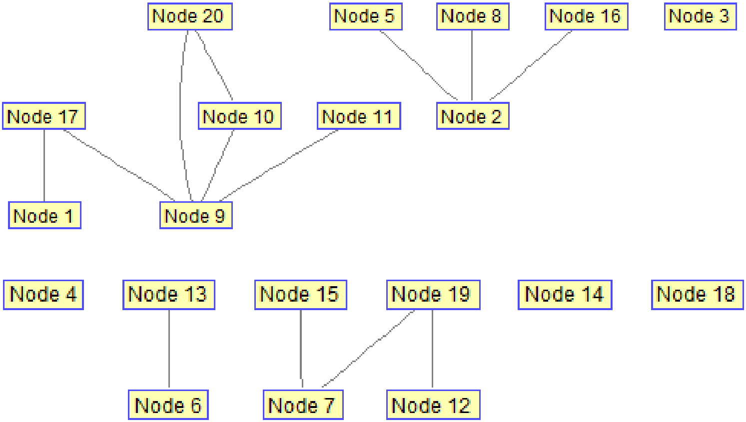
Q: Are any pairs of sequences correlated? Are there patterns of correlation?

Thresholded (0.5) sample correlation matrix



Apparent patterns emerge after thresholding each pairwise correlation at ± 0.5.

Associated sample correlation graph

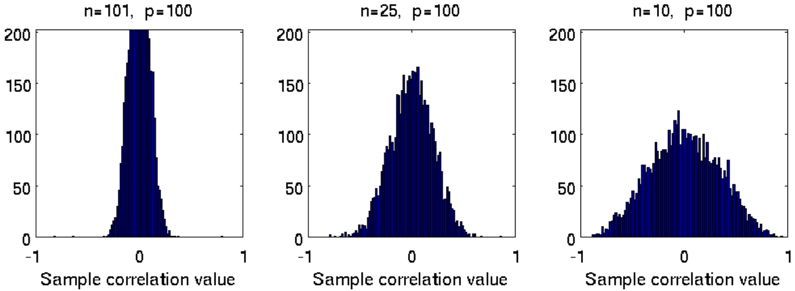


Graph has an edge between node (variable) i and j if ij-th entry of thresholded correlation is non-zero.

Sequences are actually uncorrelated Gaussian.

The problem of false discoveries: phase transitions

- Number of discoveries exhibit phase transition phenomenon
- This phenomenon gets worse as p/n increases.
- Example: false discoveries of high correlation for uncorrelated Gaussian variables



Outline

Random matrix measurement model

	Variable 1	Variable 2	...	Variable p
Sample 1	X_{11}	X_{12}	...	X_{1p}
Sample 2	X_{21}	X_{22}	...	X_{2p}
$\vdots$	$\vdots$	$\vdots$	...	$\vdots$
Sample n	X_{n1}	X_{n2}	...	X_{np}

$n \times p$ measurement matrix $\mathbb{X}$ has i.i.d. elliptically distributed rows

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

Columns of $\mathbb{X}$ index variables while rows index i.i.d. samples

$p \times p$ covariance (dispersion) matrix associated with each row is $\mathrm{cov}(\mathbf{X}^i) = \mathbf{\Sigma}$

Sparse multivariate dependency models

Two types of sparse (ensemble) correlation models:

- Sparse correlation ($\mathbf{\Sigma}$) graphical models:
  - Most correlation are zero, few marginal dependencies
  - Examples: M-dependent processes, moving average (MA) processes
- Sparse inverse-correlation ($\mathbf{K} = \mathbf{\Sigma}^{-1}$) 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
- Often only one of them is sparse

Refs: Meinshausen-Bühlmann (2006), Friedman (2007), Bannerjee (2008), Wiesel-Eldar-H (2010) .

Gaussian graphical models - GGM - (Lauritzen 1996)

Multivariate Gaussian model

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

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

- $\mathcal{G}$ has an edge e_{ij} iff $[\mathbf{K}]_{ij} \neq 0$
- Adjacency matrix $\mathbf{B}$ of $\mathcal{G}$ obtained by thresholding $\mathbf{K}$

$$\mathbf{B} = h(\mathbf{K}), \quad h(u) = \frac{1}{2}(\mathrm{sgn}(|u| - \rho) + 1)$$

ρ is arbitrary positive threshold

Banded Gaussian graphical model $\mathcal{G}$

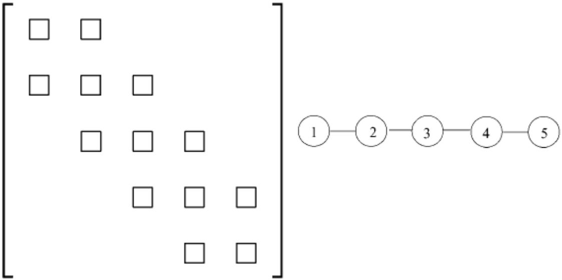
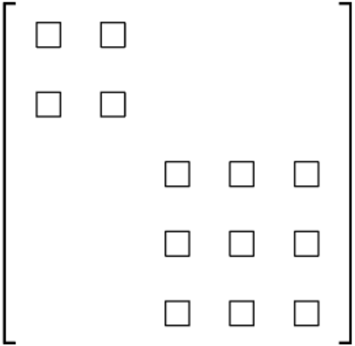


Figure: Left: inverse covariance matrix **K**. Right: associated graphical model

Example: Autoregressive (AR) process: $X_{n+1} = -aX_n + W_n$ for which $\mathbf{X} = [X_1, \ldots, X_p]$ satisfies $[\mathbf{I} - \mathbf{A}]\mathbf{X} = \mathbf{W}$ and $\mathbf{K} = \mathrm{cov}^{-1}(\mathbf{X}) = \sigma_W^2[\mathbf{I} - \mathbf{A}][\mathbf{I} - \mathbf{A}]^T$.

Block diagonal Gaussian graphical model $\mathcal{G}$



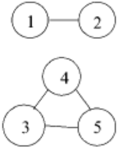
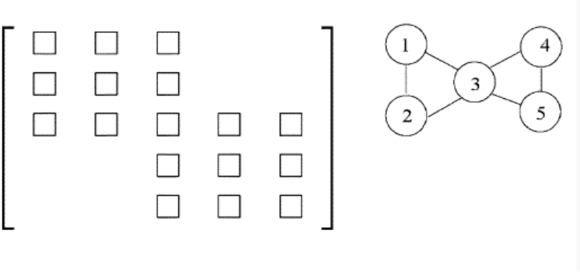


Figure: Left: inverse covariance matrix **K**. Right: associated graphical model

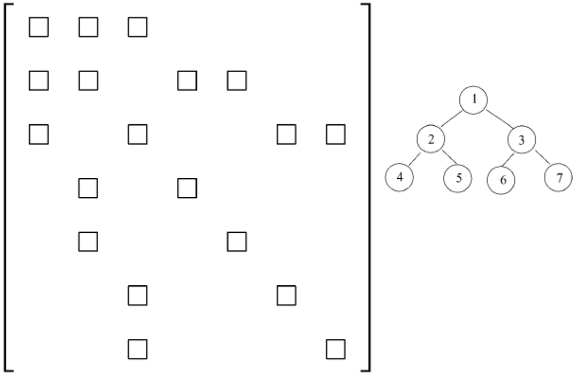
Example: $X_n = [Y_n, Z_n]$, Y_n, Z_n independent processes.

Two coupled block Gaussian graphical model $\mathcal{G}$



Example: $X_n = [Y_n + U_n, U_n, Z_n + U_n]$, Y_n, Z_n, U_n independent AR processes.

Multiscale Gaussian graphical model $\mathcal{G}$



Spatial graphical model: Poisson random field

Let $p^t(x, y)$ be a space-time process satisfying Poisson equation

$$\frac{\nabla^2 p^t}{\nabla x^2} + \frac{\nabla^2 p^t}{\nabla y^2} = W^t$$

where $W^t = W^t(x, y)$ is driving process.
For small Δ_x, Δ_y, p satisfies the difference equation:

$$X^t_{i,j} = \frac{(X^t_{i+1,j} + X^t_{i-1,j})\Delta^2 y + (X^t_{i,j+1} + X^t_{i,j-1})\Delta^2 x - W^t_{i,j}\Delta^2 x \Delta^2 y}{2(\Delta^2 x + \Delta^2 y)}$$

In matrix form, as before: $[\mathbf{I} - \mathbf{A}]\mathbf{X}^t = \mathbf{W}^t$ and

$$\mathbf{K} = \mathrm{cov}^{-1}(\mathbf{X}^t) = \sigma^2_W [\mathbf{I} - \mathbf{A}][\mathbf{I} - \mathbf{A}]^T$$

A is sparse "pentadiagonal" matrix.

Random field generated from Poisson equation

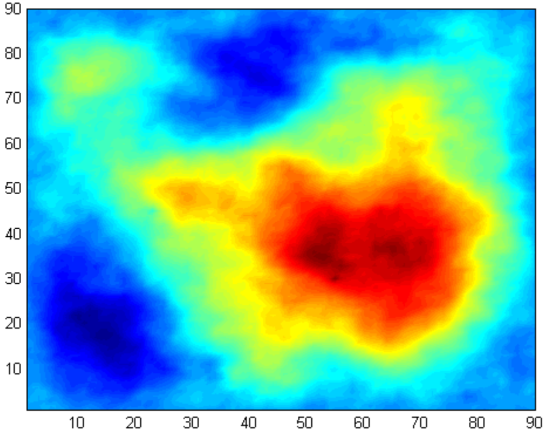


Figure: Poisson random field. $\mathbf{W}^t = \mathbf{N}_{iso} + sin(\omega_1 t)\mathbf{e}_1 + sin(\omega_2 t)\mathbf{e}_2$ ($\omega_1 = 0.025$, $\omega_2 = 0.02599$, SNR=0dB).

Empirical partial correlation map for spatial random field

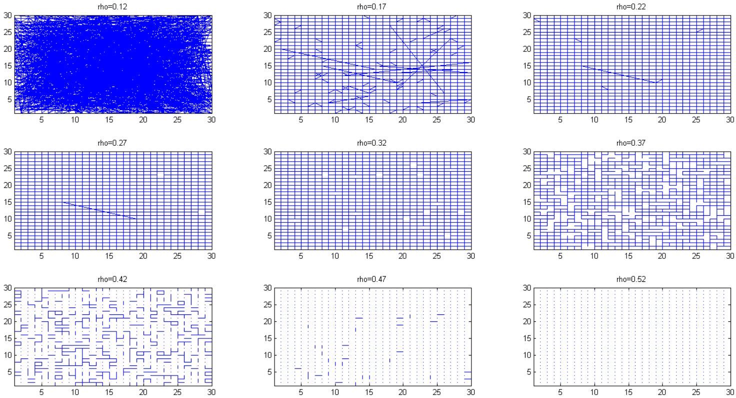


Figure: Empirical parcorr at various threshold levels. p=600, n=1500

Empirical correlation map of spatial random field

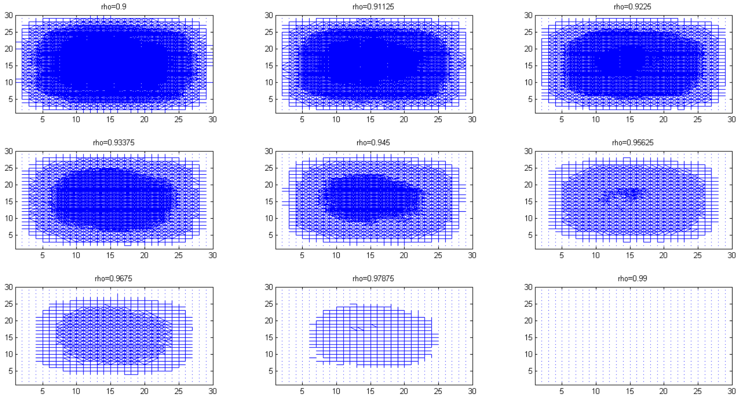


Figure: Empirical corr at various threshold levels. p=600, n=1500

Outline

1. Corrrelation mining
2. Caveats: why we need to be careful!
3. Dependency models
4. Correlation Mining Theory
5. Experiments
6. Summary and perspectives

Correlation mining: theory

Given

- Number of nodes $= p$
- Number of samples $= n$
- Correlation threshold $= \rho$
- $p \times p$ matrix of sample correlations
- Sparse graph assumption: $\#$ true edges $\ll p^2$

Questions

- Can we predict critical phase transition threshold ρ_c?
- What level of confidence/significance can one have on discoveries for $\rho > \rho_c$?
- Are there ways to predict the number of required samples for given threshold level and level of statistical significance?

Relevant work

- Regularized l_2 or $l_{\mathcal{F}}$ covariance estimation
  - Banded covariance model: Bickel-Levina (2008)
  - Sparse eigendecomposition model: Johnstone-Lu (2007)
  - Stein shrinkage estimator: Ledoit-Wolf (2005), Chen-Weisel-Eldar-H (2010)
- Gaussian graphical model selection
  - l_1 regularized GGM: Meinshausen-Bühlmann (2006), Wiesel-Eldar-H (2010).
  - Bayesian estimation: Rajaratnam-Massam-Carvalho (2008)
- Independence testing
  - Sphericity test for multivariate Gaussian: Wilks (1935)
  - Maximal correlation test: Moran (1980), Eagleson (1983), Jiang (2004), Zhou (2007), Cai and Jiang (2011)
- Correlation hub screening (H, Rajaratnam 2011, 2012)
  - Fixed n, asymptotic in p, covers partial correlation too.
  - Discover degree $\geq k$ hubs $\equiv$ test maximal k-NN correlation.

Hub screening theory (H and Rajaratnam 2012)

Empirical hub discoveries: For threshold ρ and degree parameter δ define number $N_{\delta,\rho}$ of vertices in sample partial-correlation graph with degree $d_i \geq \delta$

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

$$\phi_{\delta,i} = \begin{cases} 1, & \mathrm{card}\{j : j \neq i, |\mathbf{\Omega}_{ij}| \geq \rho\} \geq \delta \\ 0, & o.w. \end{cases}$$

$$\mathbf{\Omega} = \mathrm{diag}(\mathbf{R}^{\dagger})^{-1/2}\mathbf{R}^{\dagger}\mathrm{diag}(\mathbf{R}^{\dagger})^{-1/2}$$

is sample partial correlation matrix and $\mathbf{R}$ is sample correlation matrix

$$\mathbf{R} = \mathrm{diag}(\hat{\mathbf{\Sigma}})^{-1/2}\hat{\mathbf{\Sigma}}\mathrm{diag}(\hat{\mathbf{\Sigma}})^{-1/2}$$

Asymptotic familywise false-discovery rate

Asymptotic limit on false discoveries: (H and Rajaratnam 2012): Assume that rows of $\mathbb{X}$ are i.i.d. with bounded elliptically contoured density and block sparse covariance (null hypothesis).

Theorem

Let p and $\rho = \rho_p$ satisfy $\lim_{p\to\infty} p^{1/\delta}(p-1)(1-\rho_p^2)^{(n-2)/2} = e_{n,\delta}$. *Then*

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

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

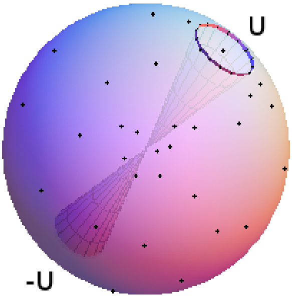
$$P_0(\rho,n) = 2B((n-2)/2,1/2)\int_{\rho}^{1}(1-u^2)^{\frac{n-4}{2}}du$$

Elements of proof

- Z-score representations for sample (partial) correlation (**P**) **R**

$$\mathbf{R} = \mathbf{U}^T\mathbf{U}, \quad \mathbf{P} = \mathbf{U}^T[\mathbf{U}\mathbf{U}^T]^{-2}\mathbf{U}, \quad (\mathbf{U} = n-1 \times p)$$

- $P_0(\rho, n)$: probability that a uniformly distributed vector $\mathbf{Z} \in S_{n-2}$ falls in cap$(r, \mathbf{U})\cap$cap$(r, -\mathbf{U})$ with $r = \sqrt{2(1-\rho)}$.
- As $p \to \infty$, $N_{\delta,\rho}$ behaves like a Poisson random variable: $P(N_{\delta,\rho} = 0) \to e^{-\lambda_{\delta,\rho,n}}$



Poisson-like convergence rate

Under assumption that

- $p^{1/\delta}(p-1)(1-\rho_p^2)^{(n-2)/2} = O(1)$

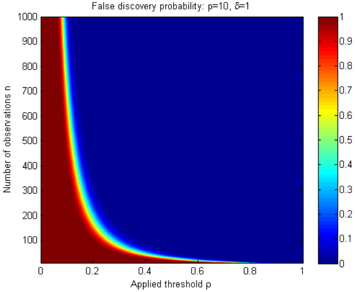
can apply Chen-Stein to obtain bound

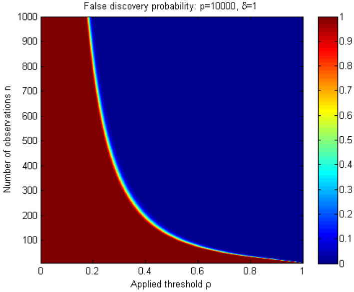
$$\left|P(N_{\delta,\rho}=0) - e^{-\lambda_{\delta,\rho,n}}\right| \leq O\left(\max\left\{p^{-1/\delta}, p^{-1/(n-2)}, \Delta_{p,n,k,\delta}\right\}\right)$$

$\Delta_{p,n,k,\delta}$ is dependency coefficient between δ-nearest-neighbors of $\mathbf{Y}_i$ and its $p-k$ furthest neighbors

Predicted phase transition for false hub discoveries

False discovery probability: $P(N_{\delta,\rho} > 0) \approx 1 - \exp(-\lambda_{\delta,\rho,n})$

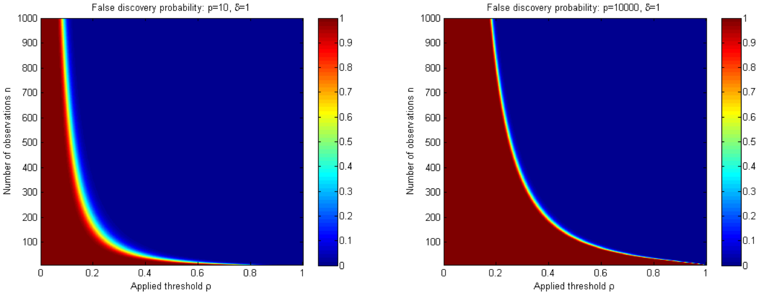




p=10	$(\delta = 1)$	p=10000

Predicted phase transition for false hub discoveries

False discovery probability: $P(N_{\delta,\rho} > 0) \approx 1 - \exp(-\lambda_{\delta,\rho,n})$

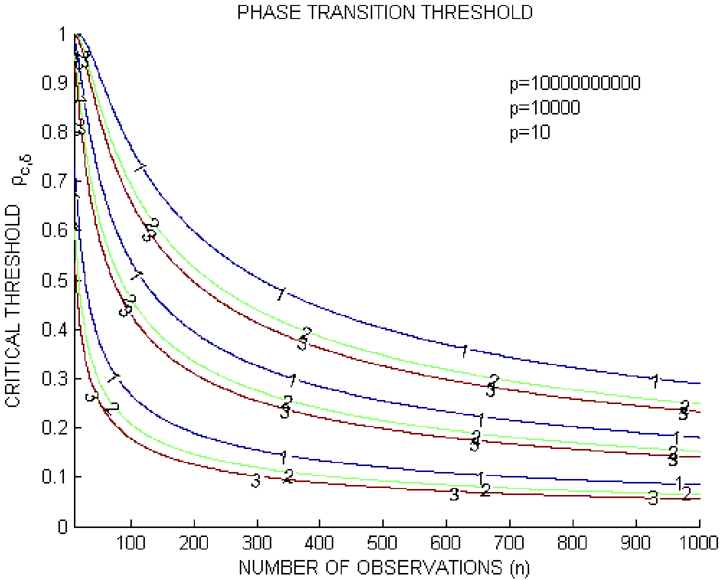


p=10	$(\delta = 1)$	p=10000

Critical threshold:

$$\rho_c = \sqrt{1 - c_{\delta,n}(p-1)^{-2\delta/\delta(n-2)-2}}$$

Phase transitions as function of δ, p



Experimental Design Table (EDT): mining connected nodes

$n\diagdown\alpha$	0.010	0.025	0.050	0.075	0.100
10	0.99\0.99	0.99\0.99	0.99\0.99	0.99\0.99	0.99\0.99
15	0.96\0.96	0.96\0.95	0.95\0.95	0.95\0.94	0.95\0.94
20	0.92\0.91	0.91\0.90	0.91\0.89	0.90\0.89	0.90\0.89
25	0.88\0.87	0.87\0.86	0.86\0.85	0.85\0.84	0.85\0.83
30	0.84\0.83	0.83\0.81	0.82\0.80	0.81\0.79	0.81\0.79
35	0.80\0.79	0.79\0.77	0.78\0.76	0.77\0.76	0.77\0.75

Table: Design table for spike-in model: $p = 1000$, detection power $\beta = 0.8$. Achievable limits in FPR (α) as function of n, minimum detectable correlation ρ_1, and level α correlation threshold (shown as $\rho_1 \backslash \rho$ in table).

Experimental validation

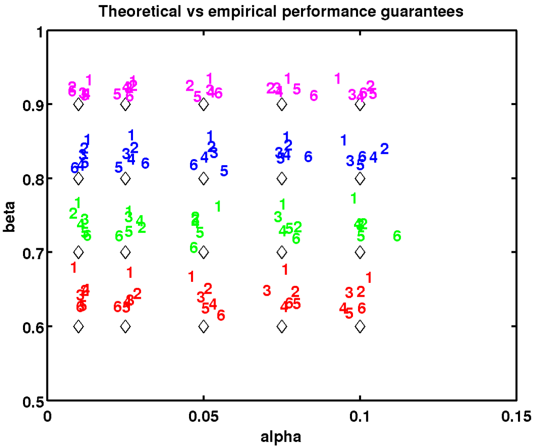


Figure: Targeted ROC operating points (α, β) (diamonds) and observed operating points (number pairs) of correlation screen designed from Experimental Design Table. Each observed operating point determined by the sample size n ranging over $n = 10, 15, 20, 25, 30, 35$.

From false positive rate for fixed ρ to p-values

Recall asymptotic false positive probability for fixed δ, n, ρ

$$P(N_{\delta,\rho} > 0) = 1 - \exp(-\lambda_{\delta,\rho,n})$$

Can relate false postive probability to maximal correlation:

$$P(N_{\delta,\rho} > 0) = P(\max_i |\rho_i(\delta)| > \rho)$$

with $\rho_i(k)$ the (partial) correlation between i and its k-NN.

$\Rightarrow$ p-value associated with vertex i having observed k-NN (partial) correlation $= \hat{\rho}_i(k)$.

$$pv_k(i) = 1 - \exp(-\lambda_{k,\hat{\rho}_i(k),n})$$

Outline

Example: 4-node-dependent Graphical Model

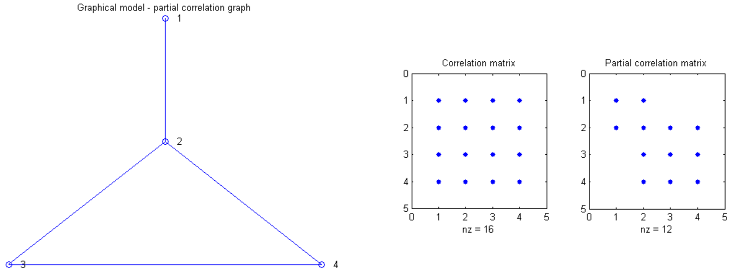
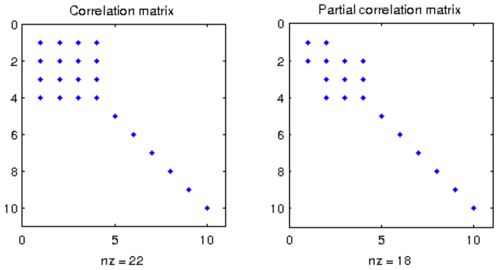


Figure: Graphical model with 4 nodes. Vertex degree distribution: 1 degree 1 node, 2 degree 2 nodes, 1 degree 3 node.

P =

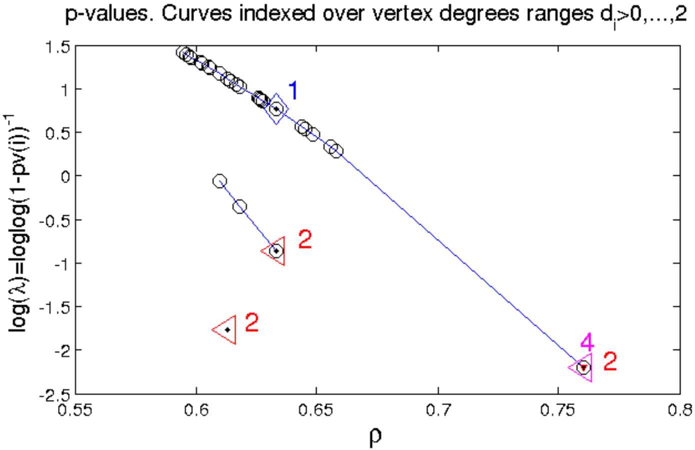
$$
\begin{matrix}
1.0000 & 0.4069 & 0 & 0 \\
0.4069 & 1.0000 & -0.5179 & -0.8138 \\
0 & -0.5179 & 1.0000 & 0.7071 \\
0 & -0.8138 & 0.7071 & 1.0000
\end{matrix}
$$

Example: First 10 nodes of 1000-node Graphical Model



- 4 node Gaussian graphical model embedded into 1000 node network with 996 i.i.d. "nuisance" nodes
- Simulate 40 observations from these 1000 variables.
- Critical threshold is $\rho_{c,1} = 0.593$. 10% level threshold is $\rho = 0.7156$.

Example: 1000-node Graphical Model



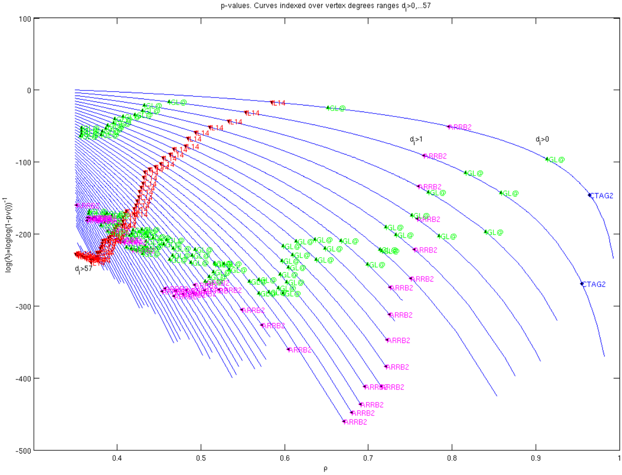
Note: $\log(\lambda) = -2$ is equivalent to pv$= 1 - e^{-e^{\log\lambda}} = 0.127$.

Example: NKI gene expression dataset

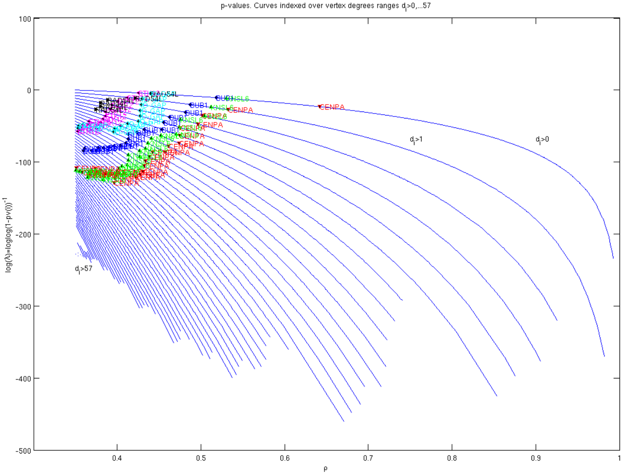
Netherlands Cancer Institute (NKI) early stage breast cancer

- $p = 24,481$ 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 $\rho_{c,1} = 0.296$

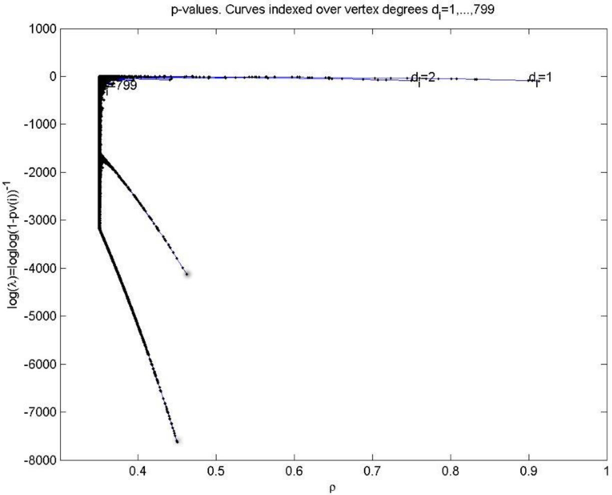
NKI p-value waterfall plot for partial correlation hubs: selected discoveries shown



NKI p-value waterfall plot for partial correlation hubs: Peng *et al* discoveries shown



NKI p-value waterfall plot for correlation hubs



Outline

Summary and perspectives

- Conclusions
  - For large p correlation mining hypersensitive to false positives
  - Theory of false positive phase transitions and significance has been developed in context of hub screening on $\mathbf{R}$, $\mathbf{R}^{\dagger}$, and $\mathbf{R}_{\mathbf{x}}^{\dagger}\mathbf{R}_{\mathbf{xy}}$.
- Other problems of interest
  - Higher order measures of dependence (information flow)
  - Time dependent samples of correlated multivariates
  - Missing data some components of multivariate are intermittent
  - Screening for other non-isomorphic sub-graphs
  - Vector valued node attributes: canonical correlations.
  - Misaligned signals: account for registration errors.

Y. Chen, A. Wiesel, Y.C. Eldar, and A.O. Hero. Shrinkage algorithms for mmse covariance estimation. *Signal Processing, IEEE Transactions on*, 58(10):5016–5029, 2010.

A. Hero and B. Rajaratnam. Hub discovery in partial correlation models. *IEEE Trans. on Inform. Theory*, 58(9):6064–6078, 2012. available as Arxiv preprint arXiv:1109.6846.

A.O. Hero and B. Rajaratnam. Large scale correlation screening. *Journ. of American Statistical Association*, 106(496):1540–1552, Dec 2011. Available as Arxiv preprint arXiv:1102.1204.

A. Wiesel, Y.C. Eldar, and A.O. Hero. Covariance estimation in decomposable gaussian graphical models. *Signal Processing, IEEE Transactions on*, 58(3):1482–1492, 2010.

K.S. Xu, M. Kliger, Y. Chen, P. Woolf, and A.O. Hero. Revealing social networks of spammers through spectral clustering. In *IEEE International COnference on Communications (ICC)*, june 2009.