CORRELATION MINING IN LARGE NETWORKS WITH LIMITED SAMPLES

Alfred Hero

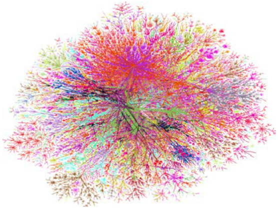
University of Michigan

Aug 20, 2014

Outline

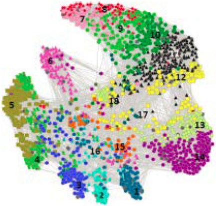
Network discovery from correlation

O/I correlation



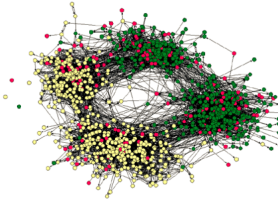
The Internet
(Burch and Cheswick, 1998)

gene correlation



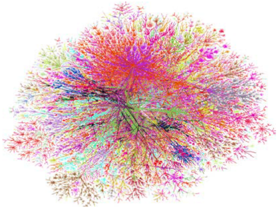
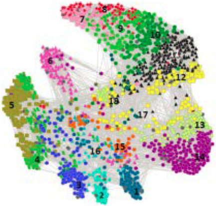
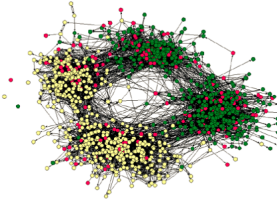
Gene pathways
(Huang, 2011)

mutual correlation



School friendships
(Moody, 2001)

Network discovery from correlation

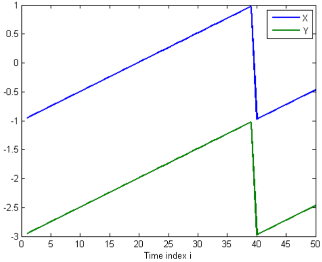
O/I correlation	*gene correlation*	*mutual correlation*
		
The Internet (Burch and Cheswick, 1998)	Gene pathways (Huang, 2011)	School friendships (Moody, 2001)

- "Big data" aspects
  - Large number of unknowns (hubs, edges, subgraphs)
  - Small number of samples for inference on unknowns
  - Crucial need to manage uncertainty (false positives)

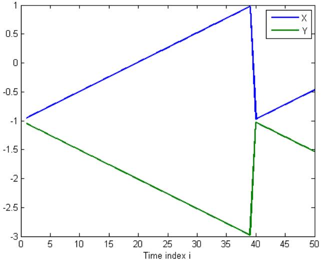
Sample correlation: $p = 2$ variables $n = 50$ samples

Sample correlation:

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



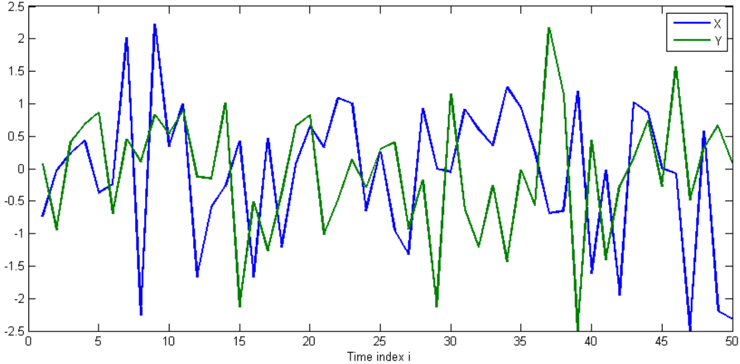
,



Positive correlation =1

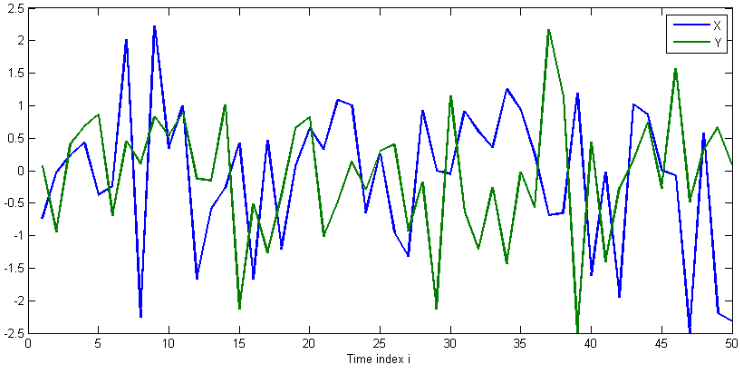
Negative correlation =-1

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



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

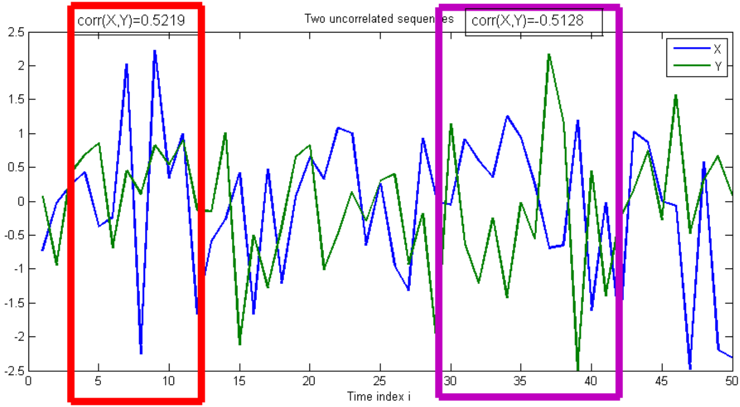
Sample correlation for two sequences: $p = 2$, $n = 50$



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 two sequences: $p = 2$, $n < 15$



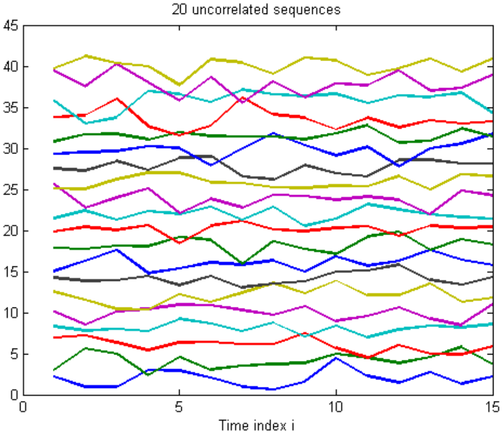
Q: Are the two time sequences X_i and Y_j correlated?

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

Real-world example: reported correlation divergence

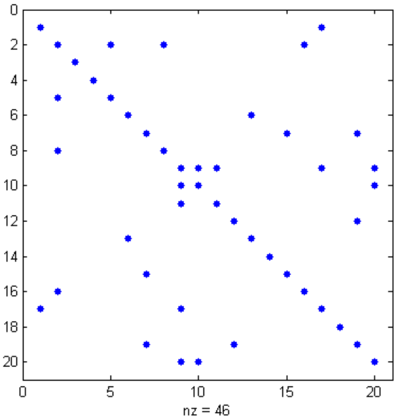
Source: FuturesMag.com `www.futuresmag.com/.../Dom%20FEB%2024.JPG`

Correlating a set of $p = 20$ sequences



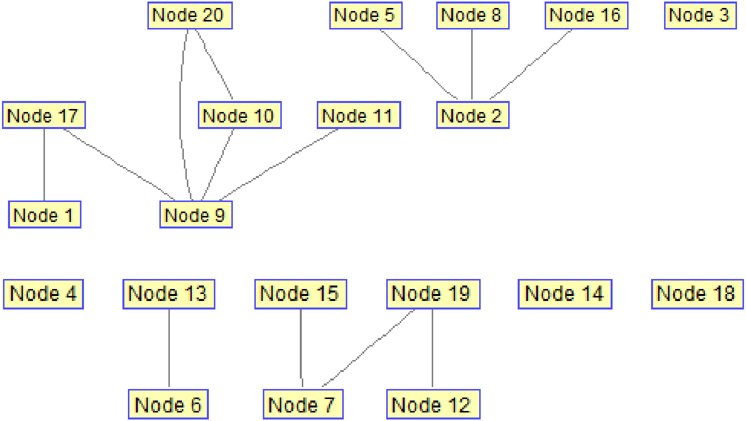
Q: Are any pairs of sequences correlated? Are there patterns of correlation?

Thresholded (0.5) sample correlation matrix **R**



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

Associated sample correlation graph

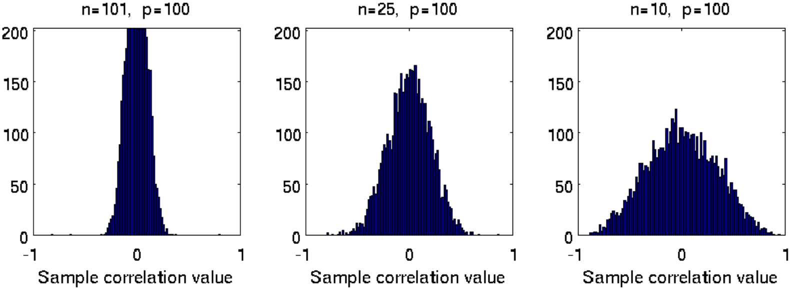


Graph has an edge between node (variable) *i* and *j* if *ij*-th entry of thresholded correlation is non-zero.

Sequences are actually uncorrelated Gaussian.

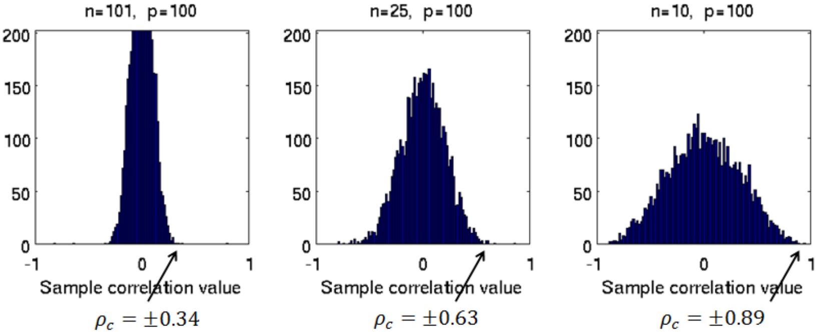
The problem of false discoveries: phase transitions

- 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



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

Objective of correlation mining

Objective: estimate or detect patterns of correlation in complex sample-poor environments

High level question being addressed

> *What are the fundamental properties of a network of p interacting variables that can be accurately estimated from a small number n of measurements?*

Regimes

- $n/p \to \infty$: sample rich regime (CLT, LLNs)
- $n/p \to c$: sample critical regime (Semi-circle, Marchenko-Pastur)
- $n/p \to 0$: sample starved regime (Chen-Stein)

Importance of correlation mining in SP applications

- Network modeling: learning/simulating descriptive models
- Empirical prediction: forecast a response variable Y
- Classification: estimate type of correlation from samples
- Anomaly detection: localize unusual activity in a sample

Importance of correlation mining in SP applications

- Network modeling: learning/simulating descriptive models
- Empirical prediction: forecast a response variable Y
- Classification: estimate type of correlation from samples
- Anomaly detection: localize unusual activity in a sample

Each application requires estimate of covariance matrix $\mathbf{\Sigma}_X$ or its inverse

Prediction: Linear minimum MSE predictor of q variables $\mathbf{Y}$ from $\mathbf{X}$

$$\hat{\mathbf{Y}} = \mathbf{\Sigma}_{YX}\mathbf{\Sigma}_X^{-1}\mathbf{X}$$

Covariance matrix related to inter-dependency structure.

Classification: QDA test $H_0 : \mathbf{\Sigma}_X = \mathbf{\Sigma}_0$ vs $H_1 : \mathbf{\Sigma}_X = \mathbf{\Sigma}_1$

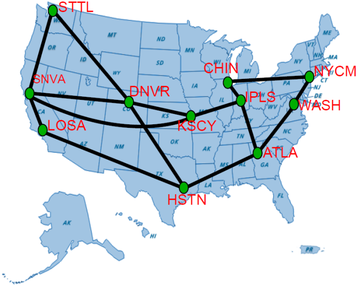
$$\overline{\mathbf{X}}^T(\mathbf{\Sigma}_0^{-1} - \mathbf{\Sigma}_1^{-1})\overline{\mathbf{X}} \quad \underset{H_0}{\overset{H_1}{\gtrless}} \quad \eta$$

Anomaly detection: Mahalanobis test $H_0 : \mathbf{\Sigma}_X = \mathbf{\Sigma}_0$ vs $H_1 : \mathbf{\Sigma}_X \neq \mathbf{\Sigma}_0$

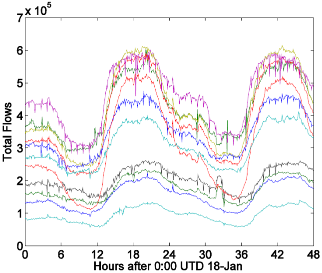
$$\frac{\overline{\mathbf{X}}^T\mathbf{\Sigma}_0^{-1}\overline{\mathbf{X}}}{\ldots} \quad \overset{H_1}{>} \quad \eta$$

Correlation mining on Abilene network traffic

11 node Abilene network



11 x 576 NetFlow measurements



Correlation mining: infer properties of correlation from small number of samples.

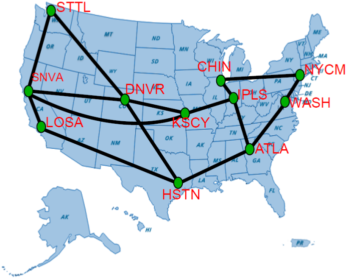
- p: number of variables
- P: number of unknown parameters
- n: number of independent samples

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

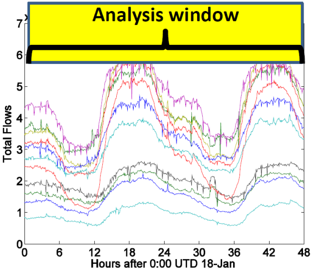
Abiline: Spatial-only correlation mining: i.i.d. over time

$$p = 11,\ P = \binom{11}{2},\ n = 576$$

11 node Abilene network



11 x 576 NetFlow measurements



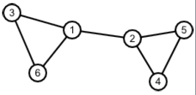
11x11 spatial correlation matrix

$$\begin{bmatrix} 1 & 0.5 & -0.5 & 0.1 & 0.2 & 0.7 & 0.1 & \dots & 0.2 \\ 0.5 & 1 & 0.1 & 0.5 & -0.7 & 0.1 & -0.3 & \dots & -0.1 \\ -0.5 & 0.1 & 1 & 0.4 & 0.1 & 0.6 & 0.2 & \dots & 0.1 \\ 0.1 & 0.5 & 0.4 & 1 & 0.8 & 0.2 & 0.1 & \dots & 0.2 \\ 0.2 & -0.7 & 0.1 & 0.8 & 1 & 0.1 & 0.3 & \dots & 0.1 \\ 0.7 & 0.1 & 0.6 & 0.2 & 0.1 & 1 & 0.1 & \dots & 0.1 \\ 0.1 & -0.3 & 0.2 & 0.1 & 0.3 & 0.1 & 1 & \dots & 0.2 \\ \dots & \dots & \dots & \dots & \dots & \dots & \dots & & 0.1 \\ 0.1 & -0.1 & 0.1 & 0.2 & 0.1 & 0.1 & 0.2 & 0.1 & 1 \end{bmatrix}$$

11x11 adjacency matrix

$$\begin{bmatrix} 0 & 1 & 1 & 0 & 0 & 1 & 0 & \dots & 0 \\ 1 & 0 & 0 & 1 & 1 & 0 & 0 & \dots & 0 \\ 1 & 0 & 0 & 0 & 0 & 1 & 0 & \dots & 0 \\ 0 & 1 & 0 & 0 & 1 & 0 & 0 & \dots & 0 \\ 0 & 1 & 0 & 1 & 0 & 0 & 0 & \dots & 0 \\ 1 & 0 & 1 & 0 & 0 & 0 & 0 & \dots & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & \dots & 0 \\ \dots & \dots & \dots & \dots & \dots & \dots & \dots & & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \end{bmatrix}$$

Correlation graph

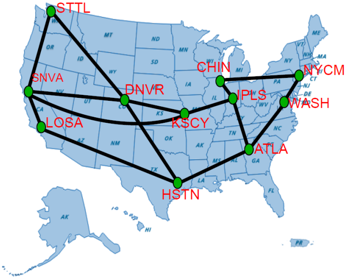


H and Rajaratnam, “Large scale correlation screening," J. Amer Statistical Association, 2011.

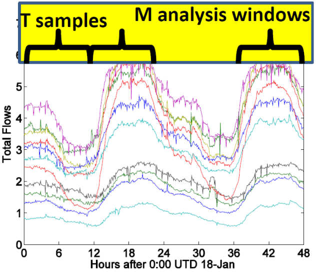
Abilene: Spatio-temp correlation mining

$$p = 11,\ P = \binom{11}{2} M,\ n = T \text{ (per window)}$$

11 node Abilene network



11 x 576 NetFlow measurements



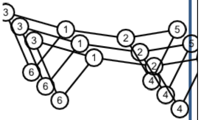
11x11 spatial correlation matrices

$$\begin{bmatrix} 1 & 0.5 & -0.5 & 0.1 & 0.2 & 0.7 & 0.1 & \ldots & 0.2 \\ 0.5 & 1 & 0.1 & 0.5 & -0.7 & 0.1 & -0.3 & \ldots & -0.1 \\ -0.5 & 0.1 & 1 & 0.4 & 0.1 & 0.6 & 0.2 & \ldots & 0.1 \\ 0.1 & 0.5 & 0.4 & 1 & 0.8 & 0.2 & 0.1 & \ldots & 0.2 \\ 0.2 & -0.7 & 0.1 & 0.8 & 1 & 0.1 & 0.3 & \ldots & 0.1 \\ 0.7 & 0.1 & 0.6 & 0.2 & 0.1 & 1 & 0.1 & \ldots & 0.1 \\ 0.1 & -0.3 & 0.2 & 0.1 & 0.3 & 0.1 & 1 & \ldots & 0.2 \\ \ldots & \ldots & \ldots & \ldots & \ldots & \ldots & \ldots & & 0.1 \\ 0.1 & -0.1 & 0.1 & 0.2 & 0.1 & 0.1 & 0.2 & 0.1 & 1 \end{bmatrix}$$

11x11 adjacency matrices

$$\begin{bmatrix} 0 & 1 & 1 & 0 & 0 & 1 & 0 & \ldots & 0 \\ 1 & 0 & 0 & 1 & 1 & 0 & 0 & \ldots & 0 \\ 1 & 0 & 0 & 0 & 0 & 1 & 0 & \ldots & 0 \\ 0 & 1 & 0 & 0 & 1 & 0 & 0 & \ldots & 0 \\ 0 & 1 & 0 & 1 & 0 & 0 & 0 & \ldots & 0 \\ 1 & 0 & 1 & 0 & 0 & 0 & 0 & \ldots & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & \ldots & 0 \\ \ldots & \ldots & \ldots & \ldots & \ldots & \ldots & \ldots & & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \end{bmatrix}$$

Correlation graphs

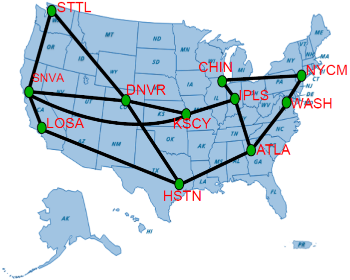


H and Rajaratnam, “Large scale correlation screening," J. Amer Statistical Association, 2011.

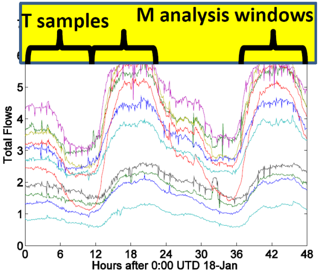
Spatio-temp correlation mining: stationary over time

$$p = 11T,\ P = \binom{11}{2}T,\ n = M = 576/T$$

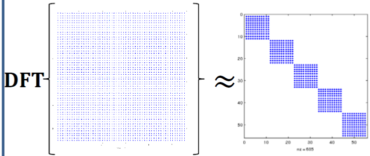
11 node Abilene network



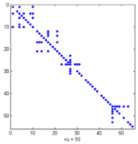
11 x 576 NetFlow measurements



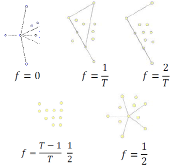
(11 T)x(11 T) spatio-temp correlation



(11 T)x(11 T) adjacency



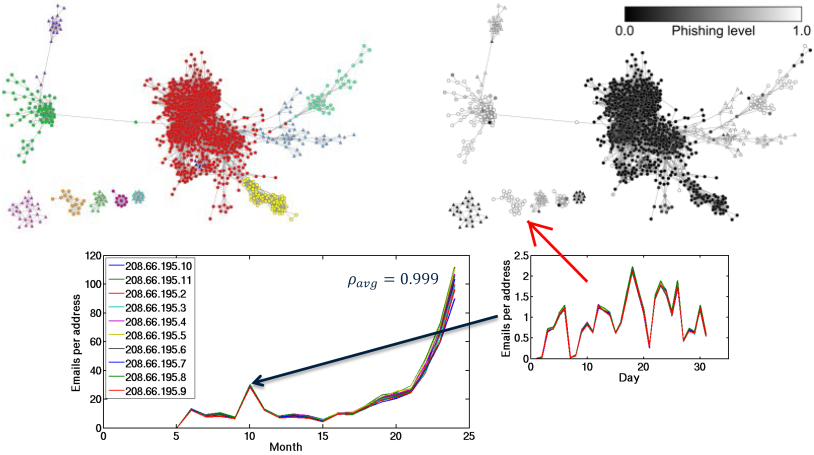
Correlation graphs



Firouzi, Wei and H, “Spectral correlation mining of multivariate time series,” Excursions in Harmonic Analysis, 2014.

Correlation mining for community detection

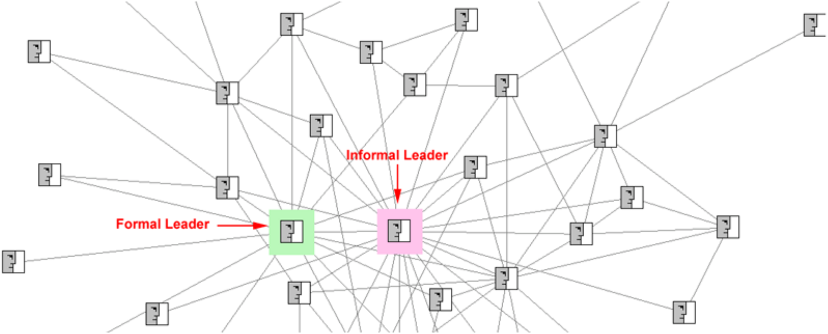
$p = 100,000, \; n = 30$



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

Correlation mining for detecting hubs of dependency

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

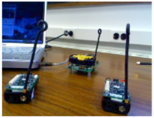


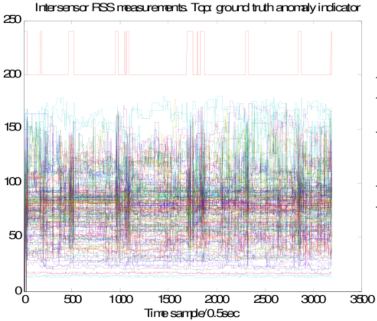
Source: orgnet.com

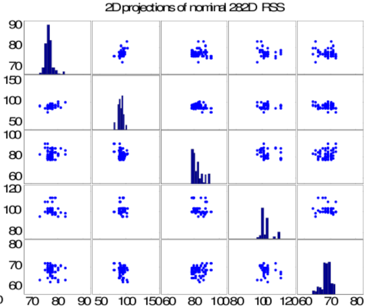
Informal leader has higher hub degree δ than formal leader

Correlation mining for intrusion detection

$p = 182, n = 20$



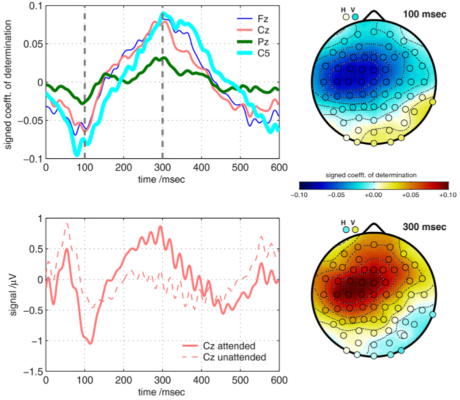



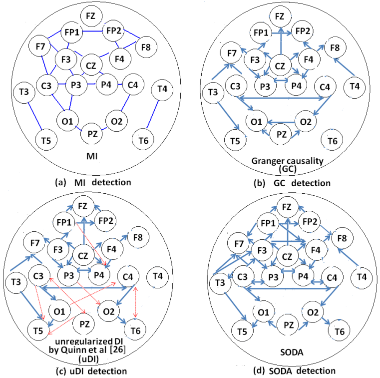


Chen, Wiesel and H, "Robust shrinkage estimation of high dimensional covariance matrices," IEEE TSP 2011

Correlation mining for neuroscience

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



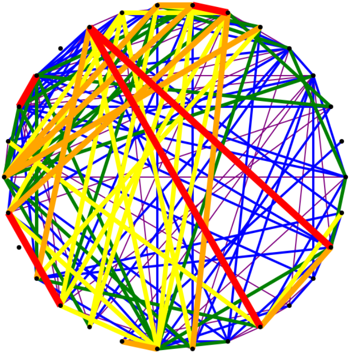


Xu, Syed and H, "EEG spatial decoding with shrinkage optimized directed information assessment," ICASSP 2012

Correlation mining for musicology: Mazurka Project

$p = 3134$, $n = 15$

One of 49 Chopin Mazurkas

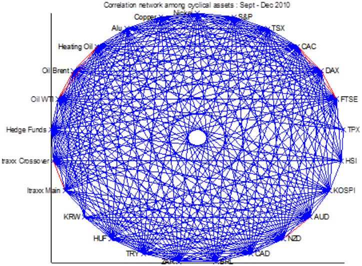


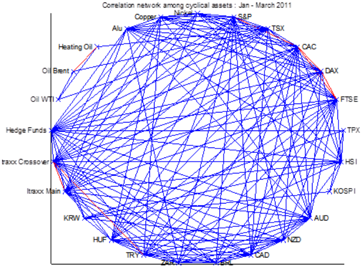
Correlation of 30 performers

(Center for History and Analysis of Recorded Music (CHARM) `http://www.charm.rhul.ac.uk`)

Correlation mining for finance

$p = 2000, n_1 = 60, n_2 = 80$





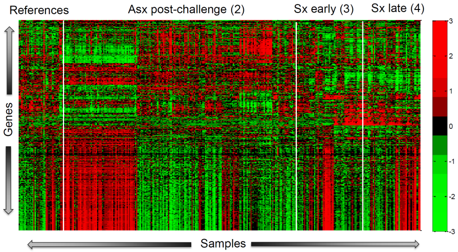
Source: "What is behind the fall in cross assets correlation?" J-J Ohana, 30 mars 2011, Riskelia's blog.

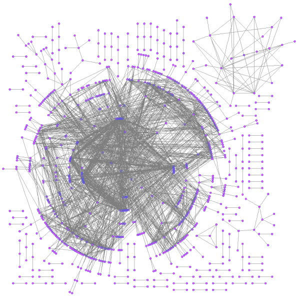
- Left: Average correlation: 0.42, percent of strong relations 33%
- Right: Average correlation: 0.3, percent of strong relations 20%

Firouzi, Wei and H, Spatio-Temporal Analysis of Gaussian WSS Processes, IEEE GlobalSIP, 2013

Correlation mining for biology: gene-gene network

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



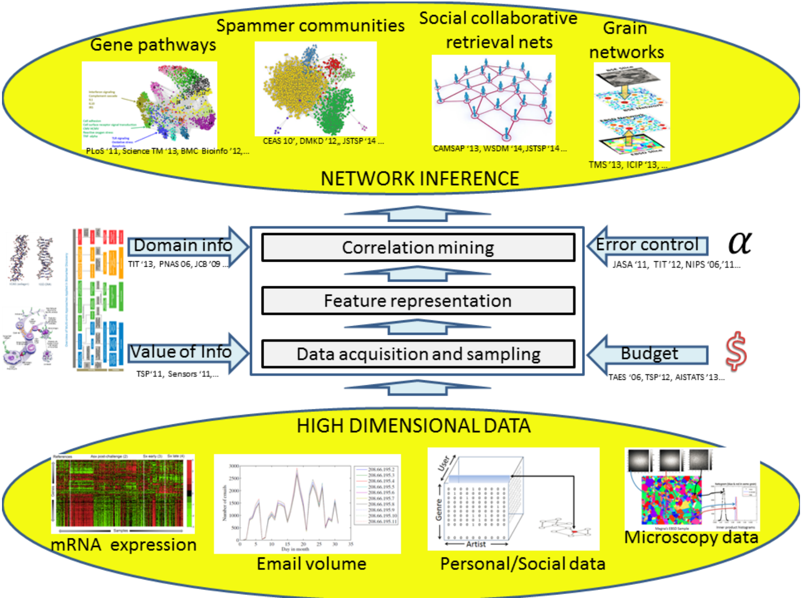


Gene expression correlation graph

Q: What genes are hubs in this correlation graph?

Huang, . . ., and H, Temporal Dynamics of Host Molecular Responses Differentiate. . . , PLoS Genetics, 2011

Correlation mining pipeline



Outline

Measurement matrix, correlation and partial correlation

	Variable 1	Variable 2	...	Variable d
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. rows $\mathbf{X}^i$ with $\mathbf{\Sigma} = \mathrm{cov}(\mathbf{X}^i)$

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

Measurement matrix, correlation and partial correlation

	Variable 1	Variable 2	...	Variable d
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. rows $\mathbf{X}^i$ with $\boldsymbol{\Sigma} = \mathrm{cov}(\mathbf{X}^i)$

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

- $p \times p$ *correlation* matrix:

$$\boxed{\boldsymbol{\Gamma} = \mathrm{diag}(\boldsymbol{\Sigma})^{-1/2}\, \boldsymbol{\Sigma}\, \mathrm{diag}(\boldsymbol{\Sigma})^{-1/2}}$$

Measurement matrix, correlation and partial correlation

	Variable 1	Variable 2	...	Variable d
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. rows $\mathbf{X}^i$ with $\mathbf{\Sigma} = \mathrm{cov}(\mathbf{X}^i)$

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

- $p \times p$ *correlation* matrix:

$$\boxed{\mathbf{\Gamma} = \mathrm{diag}(\mathbf{\Sigma})^{-1/2}\, \mathbf{\Sigma}\, \mathrm{diag}(\mathbf{\Sigma})^{-1/2}}$$

- $p \times p$ *partial correlation* matrix:

$$\boxed{\mathbf{\Omega} = \mathrm{diag}(\mathbf{\Sigma}^{-1})^{-1/2}\, \mathbf{\Sigma}^{-1}\, \mathrm{diag}(\mathbf{\Sigma}^{-1})^{-1/2}}$$

Correlation vs Partial Correlation

Sparsity is a key property since leads to fewer unknown parameters

- 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) .

Example: Gaussian graphical models (GGM)

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

- GGM specifies a graph associated with $p(\mathbf{x})$ (Lauritzen 1996)
- $\mathcal{G}$ has an edge e_{ij} iff $[\mathbf{K}]_{ij} \neq 0$
- Adjacency matrix $\mathbf{B}$ of $\mathcal{G}$ obtained by thresholding $\mathbf{K}$

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

To discover $\mathbf{K}_{ij} = 0$, ρ can be arbitrary positive threshold

Example: Gaussian graphical models (GGM)

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

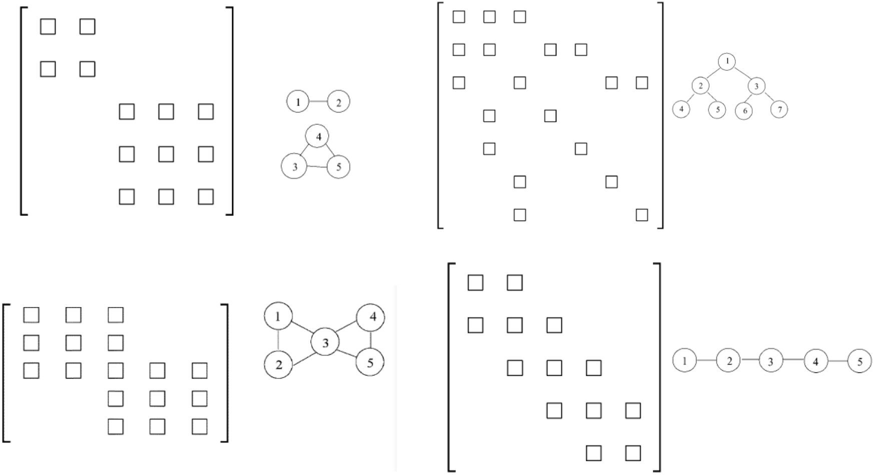
- GGM specifies a graph associated with $p(\mathbf{x})$ (Lauritzen 1996)
- $\mathcal{G}$ has an edge e_{ij} iff $[\mathbf{K}]_{ij} \neq 0$
- Adjacency matrix $\mathbf{B}$ of $\mathcal{G}$ obtained by thresholding $\mathbf{K}$

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

To discover $\mathbf{K}_{ij} = 0$, ρ can be arbitrary positive threshold

In practice: $\hat{\mathbf{K}}_{ij}$ is never zero $\Rightarrow$ ρ must be carefully chosen

Example: GGM - $\boldsymbol{\Sigma}$ or $\boldsymbol{\Sigma}^{-1}$ and $G = (V, E)$



Wiesel, Eldar and H IEEE TSP 2010

Concrete example: spatial Gauss Markov 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:

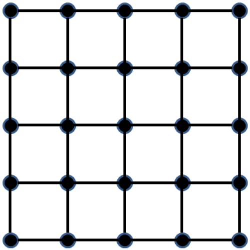
$$X^t_{i,j} = \frac{(X^t_{i+1,j} + X^t_{i-1,j})\Delta^2 y + (X^t_{i,j+1} + X^t_{i,j-1})\Delta^2 x - W^t_{i,j}\Delta^2 x \Delta^2 y}{2(\Delta^2 x + \Delta^2 y)}$$

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

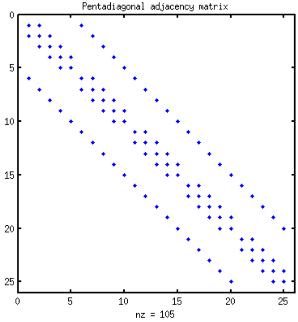
$$\mathbf{K} = \mathrm{cov}^{-1}(\mathbf{X}^t) = \sigma^2_W [\mathbf{I} - \mathbf{A}][\mathbf{I} - \mathbf{A}]^T$$

A is sparse "pentadiagonal" matrix.

Example: 5×5 Poisson random field graphical model



Graph G_K on $\mathbb{R}^2$



corresp. **K** adjacency matrix

Example: Gauss random field from Poisson equation

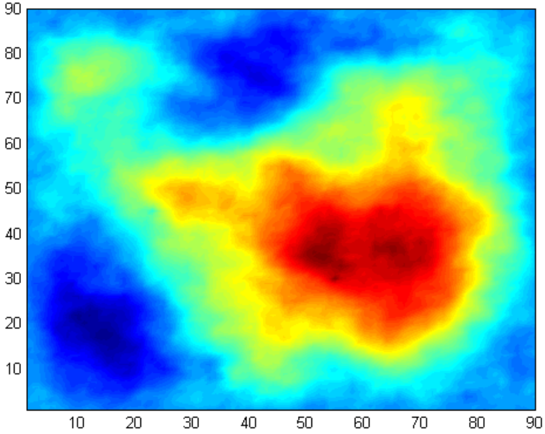


Figure: Poisson random field. $\mathbf{W}^t = \mathbf{N}_{iso} + sin(\omega_1 t)\mathbf{e}_1 + sin(\omega_2 t)\mathbf{e}_2$ ($\omega_1 = 0.025$, $\omega_2 = 0.02599$, SNR=0dB).

Empirical correlation graph for Gauss random field

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

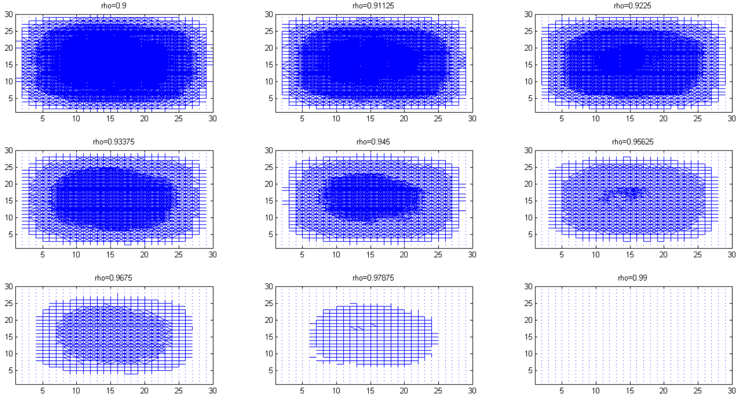


Figure: Empirical corr at various threshold levels. p=900, n=1500

Empirical partial correlation graph for Gauss random field

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

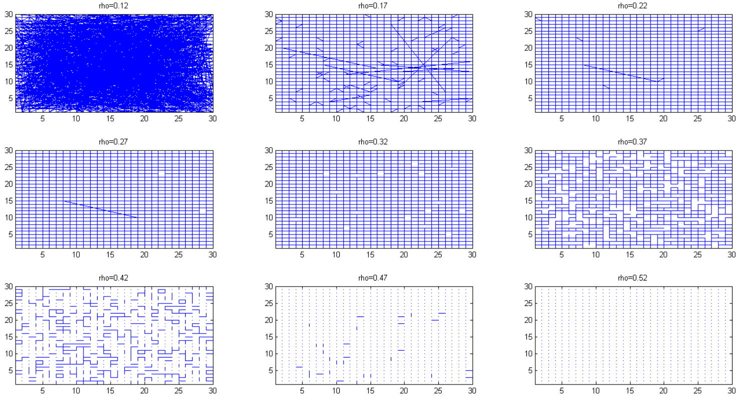


Figure: Empirical parcorr at various threshold levels. p=900, n=1500

Outline

Prior work: cov estimation, selection, screening

- 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).
  - Sparse Kronecker GGM (Matrix Normal):Allen-Tibshirani (2010), Tsiligkaridis-Zhou-H (2012)
- 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 screening (H, Rajaratnam 2011, 2012)
  - Find variables having high correlation wrt other variables
  - Find hubs of degree $\geq k \equiv$ test maximal k-NN.

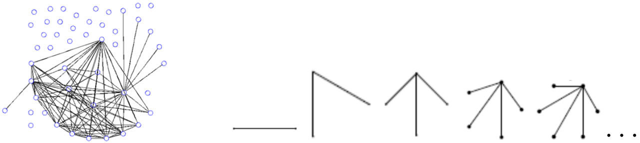
Prior work: cov estimation, selection, screening

- 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).
  - Sparse Kronecker GGM (Matrix Normal):Allen-Tibshirani (2010), Tsiligkaridis-Zhou-H (2012)
- 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 screening (H, Rajaratnam 2011, 2012)
  - Find variables having high correlation wrt other variables
  - Find hubs of degree $\geq k$ $\equiv$ test maximal k-NN.

Here we focus on the hub screening problem

Screening for hubs (H-Rajaratnam 2011, 2012)

After applying threshold ρ obtain a graph G having edges E



- Number of hub nodes in G: $N_{\delta,\rho} = \sum_{i=1}^{p} I(d_i \geq \delta)$

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

$\mathbf{C}$ is either sample correlation matrix

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

or sample partial correlation matrix

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

Asymptotics for fixed sample size n, $p \to \infty$, and $\rho \to 1$

Asymptotics of hub screening: (Rajaratnam and H 2011, 2012))
Assume that rows of $n \times p$ matrix $\mathbb{X}$ are i.i.d. circular complex random variables with bounded elliptically contoured density and block sparse covariance.

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*

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

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

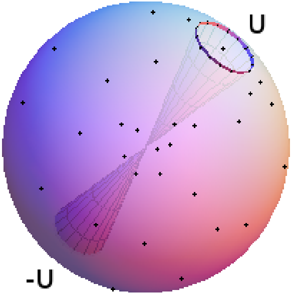
$$P_0(\rho,n) = 2B((n-2)/2,1/2)\int_\rho^1 (1-u^2)^{\frac{n-4}{2}}du$$

Elements of proof (Hero&Rajaratnam 2012)

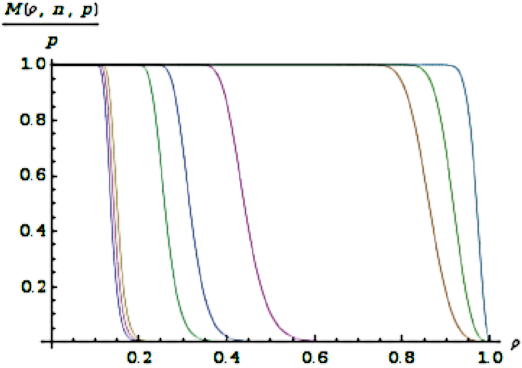
- Z-score representations for sample correlation

$$\mathbf{R} = \mathbb{U}^H \mathbb{U}, \quad \mathbb{U} = [\mathbf{U}_1, \ldots, \mathbf{U}_p], \qquad \mathbf{U}_i \in S_{n-2}$$

  S_{n-2} is sphere of dimension $n-2$ in $\mathbb{R}^{n-1}$.
- $P_0(\rho, n)$: 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: $P(N_{\delta,\rho} = 0) \to e^{-\lambda_{\delta,\rho,n}}$



$P(N_{\delta,\rho} > 0)$ as function of ρ ($\delta = 1$)

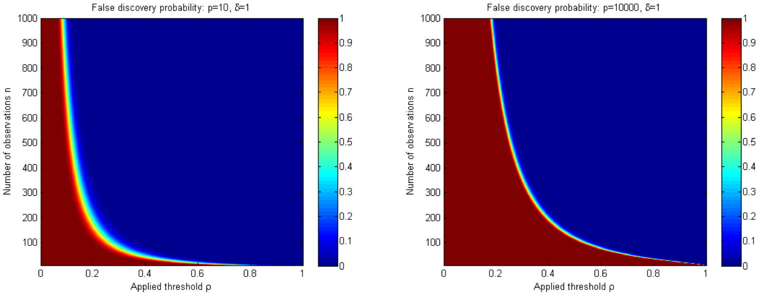


n	550	500	450	150	100	50	10	8	6
ρ_c	0.188	0.197	0.207	0.344	0.413	0.559	0.961	0.988	0.9997

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

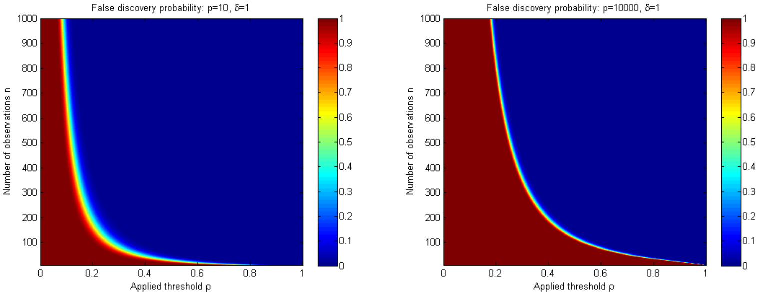
$$\rho_c = \sqrt{1 - c_n(p-1)^{-2/(n-4)}}$$

$P(N_{\delta,\rho} > 0)$ as function of ρ and n ($\delta = 1$)



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

$P(N_{\delta,\rho} > 0)$ as function of ρ and n ($\delta = 1$)

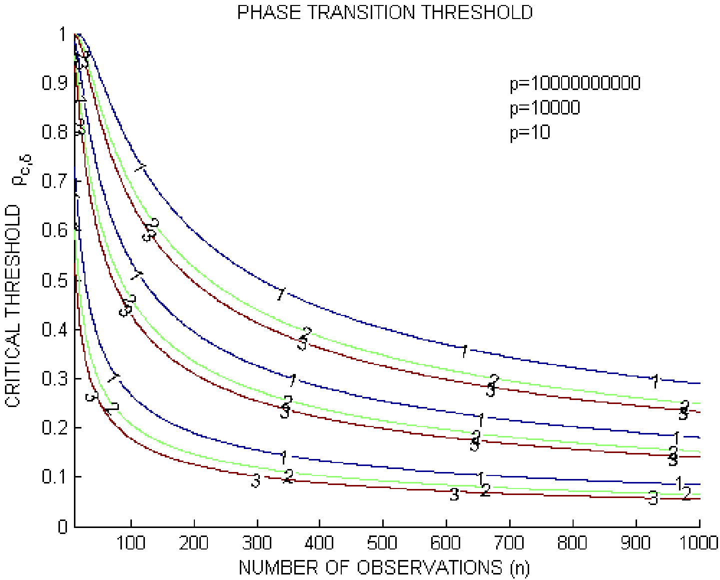


p=10	($\delta = 1$)	p=10000

Critical threshold for any $\delta > 0$:

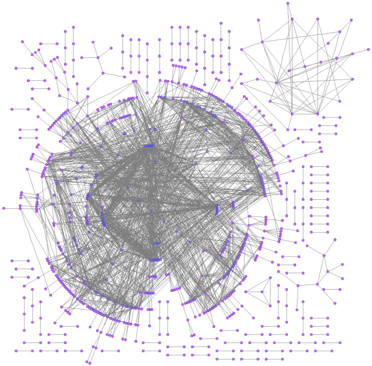
$$\rho_c = \sqrt{1 - c_{\delta,n}(p-1)^{-2\delta/\delta(n-2)-2}}$$

Critical threshold ρ_c as function of n (H-Rajaratnam 2012)



Outline

Hub mining of large correlation networks put to practice



- Partial correlation graph with $24,000$ nodes
- 14 Billion potential edges
- Phase transition threshold depends on node degree
- How to visualize the "highly significant" nodes?

[H and Rajaratnam, JASA 2011, IEEE IT 2012]

Experimental Design Table (EDT): mining connected nodes

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

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

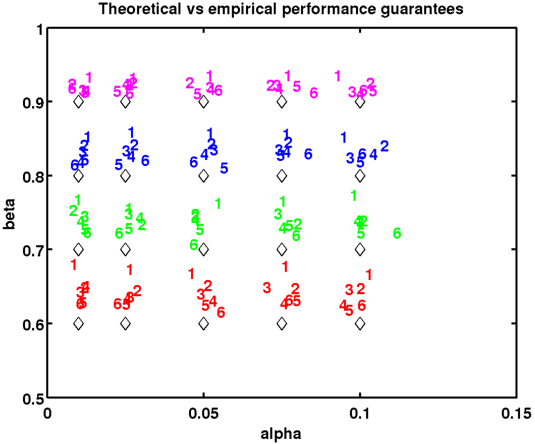


Figure: Targeted ROC operating points (α, β) (diamonds) and observed operating points (number pairs) of correlation screen designed from Experimental Design Table. Each observed operating point determined by the sample size n ranging over $n = 10, 15, 20, 25, 30, 35$.

From false positive rates to p-values

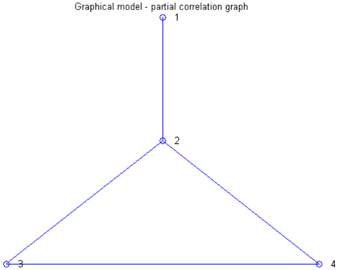
- Hub screening p-value algorithm:
  - Step 1: Compute critical phase transition threshold $\rho_{c,1}$ for discovery of connected vertices ($\delta = 1$).
  - Step 2: Generate partial correlation graph with threshold $\rho^* > \rho_{c,1}$.
  - Step 3: Compute p-values for each vertex of degree $\delta = k$ found

$$pv_k(i) = P(N_{k,\rho(i)} > 0) = 1 - \exp(-\lambda_{k,\rho(i,k)})$$

    where $\rho(i,k)$ is sample correlation between $\mathbf{X}_i$ and its k-th NN.
  - Step 4: Render these p-value trajectories as a "waterfallplot'.

$$\log(\lambda)_{k,\rho(i,k)} \text{ vs. } \rho(i,k) \text{ for } k = 1, 2, \ldots$$

Example: 4-node-dependent Graphical Model



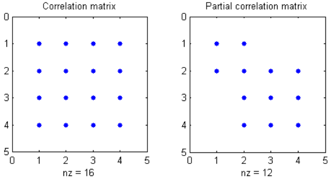
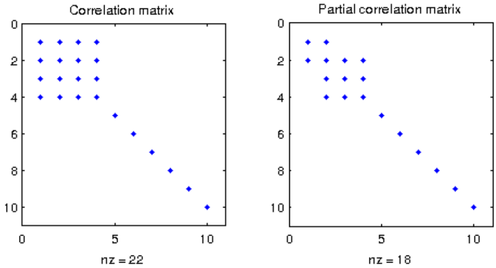


Figure: Graphical model with 4 nodes. Vertex degree distribution: 1 degree 1 node, 2 degree 2 nodes, 1 degree 3 node.

P =

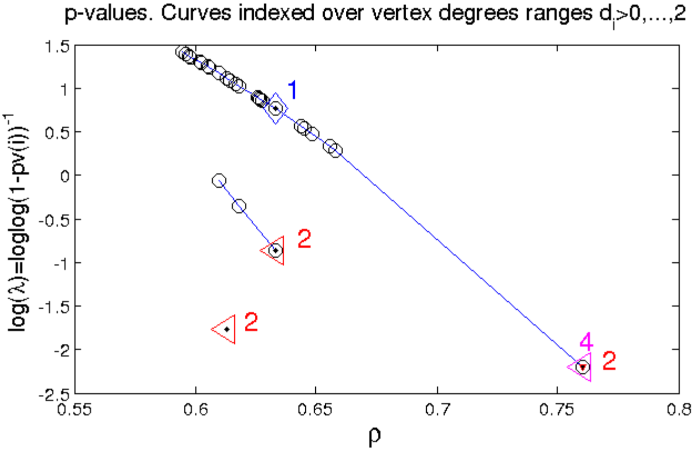
$$
\begin{array}{rrrr}
1.0000 & 0.4069 & 0 & 0 \\
0.4069 & 1.0000 & -0.5179 & -0.8138 \\
0 & -0.5179 & 1.0000 & 0.7071 \\
0 & -0.8138 & 0.7071 & 1.0000
\end{array}
$$

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



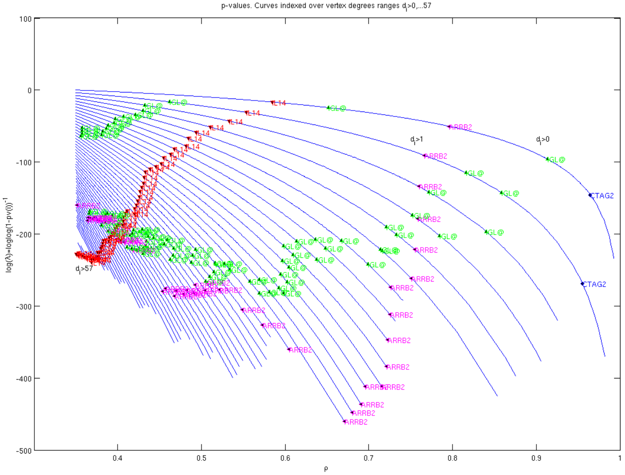
Note: $\log(\lambda) = -2$ is equivalent to pv$= 1 - e^{-e^{\log\lambda}} = 0.127$.

Example: NKI gene expression dataset

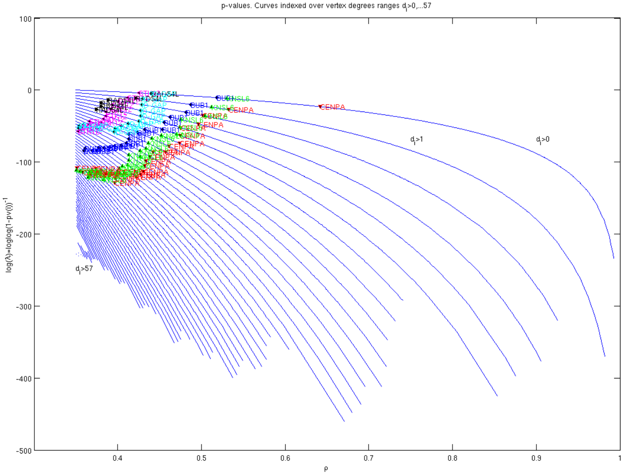
Netherlands Cancer Institute (NKI) early stage breast cancer

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

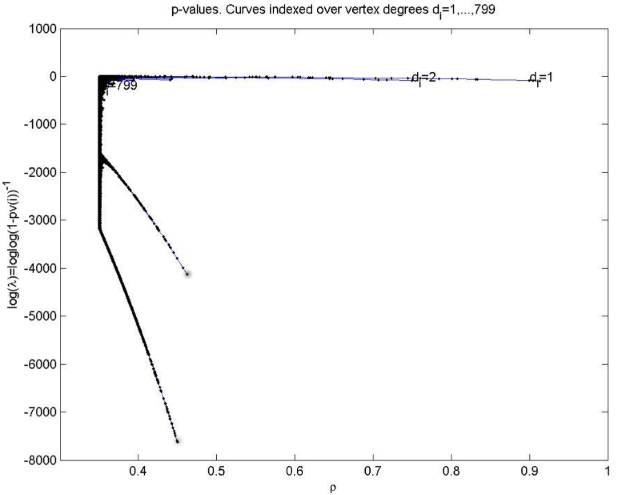
NKI p-value waterfall plot for partial correlation hubs: selected discoveries shown



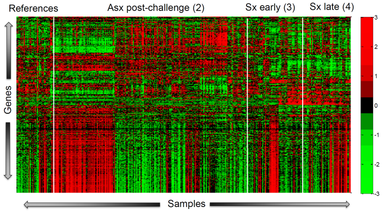
NKI p-value waterfall plot for partial correlation hubs: Peng *et al* discoveries shown



NKI p-value waterfall plot for correlation hubs



Application: correlation-mining a flu challenge study



- 17 subjects inoculated and sampled over 7 days
- 373 samples collected
- 21 Affymetrix gene chips assayed for each subject
- $p = 12023$ genes recorded for each sample
- 10 symptom scored from $\{0, 1, 2, 3\}$ for each sample

[Huang *et al*, PLoS Genetics, 2011]

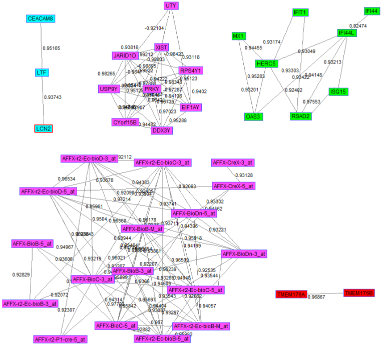
Application: correlation-mining a flu challenge study

Samples fall into 3 categories

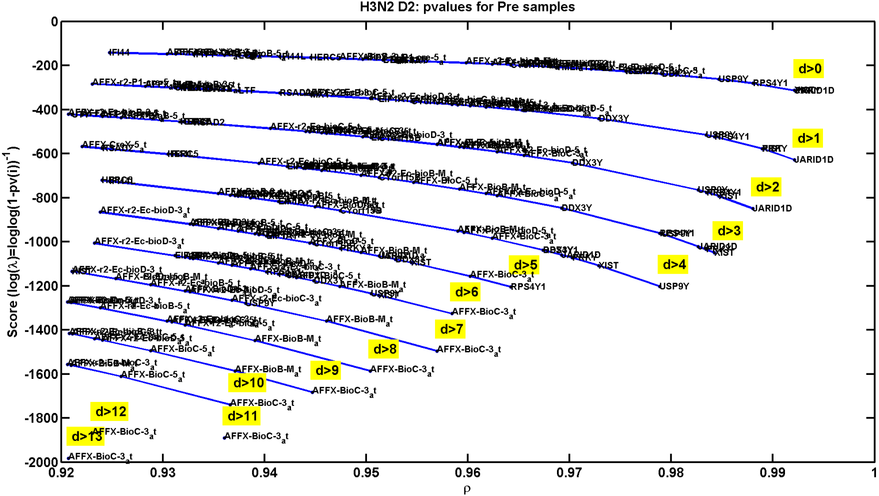
- Pre-inoculation samples
  - Number of Pre-inoc. samples: $n = 34$
  - Critical threshold: $\rho_c = 0.70$
  - 10^{-6} FWER threshold: $\rho = 0.92$
- Post-inoculation symptomatic samples
  - Number of Post-inoc. Sx samples: $n = 170$
  - Critical threshold: $\rho_c = 0.36$
  - 10^{-6} FWER threshold: $\rho = 0.55$
- Post-inoculation asymptomatic samples
  - Number of Pre-inoc. samples: $n = 152$
  - Critical threshold: $\rho_c = 0.37$
  - 10^{-6} FWER threshold: $\rho = 0.57$

Application: correlation-mining Pre-inoc. samples

- Correlation screening at FWER 10^{-6} found 1658 genes, 8718 edges
- Parcorr screening at FWER 10^{-6} found 39 genes, 111 edges



P-value waterfall analysis (Pre-inoc. parcor)



Outline

Conclusions

What we covered

- Asymptotic correlation mining theory developed for "hyper-high" dimensional setting:

$$n \textit{ fixed while } p \to \infty$$

- Universal phase transition thresholds under block sparsity
- Phase transitions useful for properly sample-sizing experiments

Conclusions

What we covered

- Asymptotic correlation mining theory developed for "hyper-high" dimensional setting:

$$n \text{ fixed while } p \to \infty$$

- Universal phase transition thresholds under block sparsity
- Phase transitions useful for properly sample-sizing experiments

Not covered here

- Linear predictor application: Prescreened OLS outperforms lasso for small n large p (Firouzi, Rajaratnam, H, 2013)
- Structured covariance: Kronecker, Toeplitz, low rank+sparse, etc (Tsiligkaridis and H 2013), (Greenewald and H 2014) ,,
- Non-linear correlation mining (Todros and H, 2011, 2012)
- Spectral correlation mining: bandpass measurements, stationary time series (Firouzi and H, 2014)

H. Firouzi, A.O. Hero, and B. Rajaratnam. Predictive correlation screening: Application to two-stage predictor design in high dimension. In *Proceedings of AISTATS. Also available as arxiv:1303.2378*, 2013a.

H. Firouzi, D. Wei, and A.O. Hero. Spatio-temporal analysis of gaussian wss processes via complex correlation and partial correlation screening. In *Proceedings of IEEE GlobalSIP Conference. Also available as arxiv:1303.2378*, 2013b.

H. Firouzi, D. Wei, and AO Hero. Spectral correlation hub screening of multivariate time series. In R. Balan, M. Begué, J. J. Benedetto, W. Czaja, and K. Okoudjou, editors, *Excursions in Harmonic Analysis: The February Fourier Talks at the Norbert Wiener Center*. Springer, 2014.

Hamed Firouzi and Alfred O Hero. Local hub screening in sparse correlation graphs. In *SPIE Optical Engineering+ Applications*, pages 88581H–88581H. International Society for Optics and Photonics, 2013.

K. Greenewald, T. Tsiligkaridis, and A.O. Hero. Kronecker sum decompositions of space-time data. *arXiv preprint arXiv:1307.7306*, 2013.

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.

N. Patwari, A.O. Hero III, and A. Pacholski. Manifold learning visualization of network traffic data. In *Proceedings of the 2005 ACM SIGCOMM workshop on Mining network data*, pages 191–196. ACM New York, NY, USA, 2005.

Koby Todros and Alfred O Hero. On measure transformed canonical correlation analysis. *Signal Processing, IEEE Transactions on*, 60(9):4570–4585, 2012a.

Koby Todros and AO Hero. Measure transformed canonical correlation analysis with application to financial data. In *Sensor Array and Multichannel Signal Processing Workshop (SAM), 2012 IEEE 7th*, pages 361–364. IEEE, 2012b.

T. Tsiligkaridis, A Hero, and S. Zhou. On convergence of kronecker graphical lasso algorithms. *IEEE Trans. on Signal Processing*, 61(9):1743 –1755, 2013a.

T. Tsiligkaridis, A. Hero, and S. Zhou. Convergence properties of Kronecker Graphical Lasso algorithms. *IEEE Trans on Signal Processing (also available as arXiv:1204.0585)*, 61(7):1743–1755, 2013b.

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.