Large scale correlation mining: fundamental performance limits

Alfred Hero

University of Michigan

July 29, 2015

Acknowledgments

Students and collaborators

- Bala Rajaratnam (Stanford)
- Hamed Firouzi (Goldman)

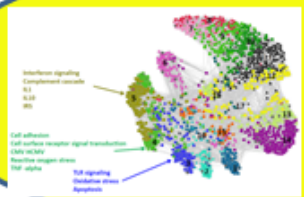
Sponsors

- Isaac Newton Institute, Cambridge UK
- AFOSR Complex Networks Program (FA9550-09-1-0471)
- ARO MURI Value of Information Program
- NSF: Theoretical Foundations Program

Outline

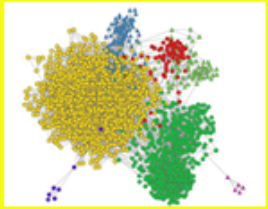
Correlation mining and network discovery

Molecular pathways



PLoS '11, Science TM '13, BMC Bioinfo '12,...

Spammer communities

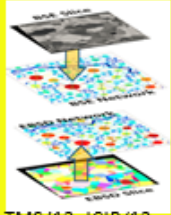


CEAS 10', DMKD '12,, JSTSP '14 ...

Social collaborative retrieval nets

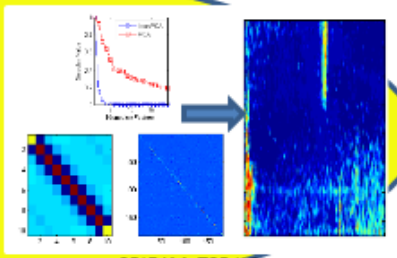
CAMSAP '13, WSDM '14, JSTSP '14 ...

Grain networks



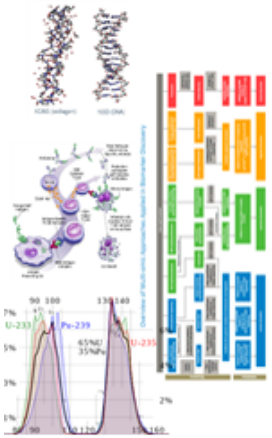
TMS '13, ICIP '13, ...

KPCA GMTI Tracks



SPIE '14, TSP '13, ...

COMPLEX NETWORK DISCOVERY



Domain info

TIT '13, PNAS 06, JCB '09 ...

Value of Info

TSP'11, Sensors '11,...

Correlation mining

Feature representation

Data acquisition and sampling

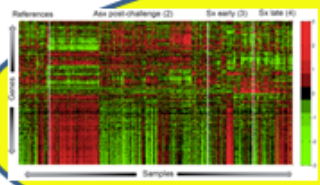
Error control α

JASA '11, TIT '12, NIPS '06,'11...

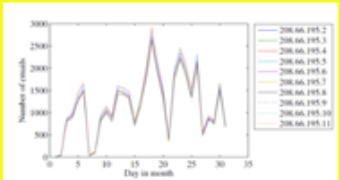
Budget

TAES '06, TSP'12, AISTATS '13...

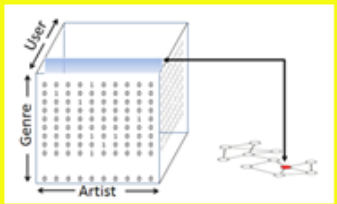
HIGH DIMENSIONAL DATA



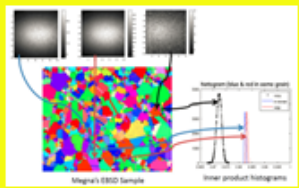
mRNA expression



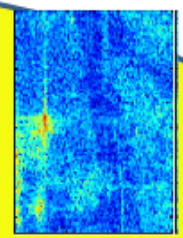
Email volume



Personal/Social data



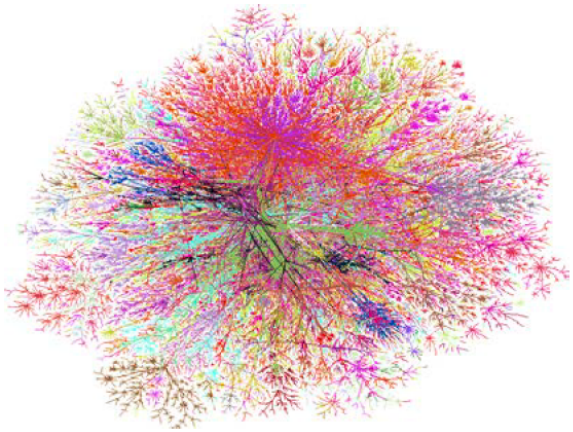
Microscopy data



Gotcha GMTI

Big Data aspects of correlation mining

O/I correlation

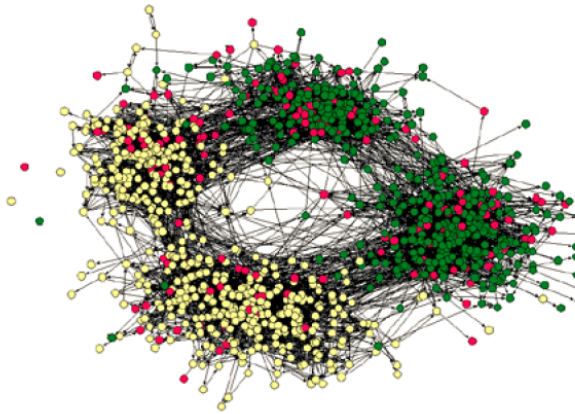


The Internet
(Burch and Cheswick, 1998)

gene correlation

Gene pathways
(Huang, 2011)

mutual correlation



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)
  - Scalability of methods to exascale is desired

Misreporting of correlations is a real problem

Table 1. We have found 12 papers in which claims coming from observational studies were tested in randomised clinical trials. Many of the trials are quite large. In most of the observational studies multiple claims were tested, often in factorial designs, e.g. vitamin D and calcium individually and together along with a placebo group. Note that none of the claims replicated in the direction claimed in the observational studies and that there was statistical significance in the opposite direction five times

ID no.	*Pos.*	*Neg.*	*No. of claims*	*Treatment(s)*	*Reference*
1	0	1	3	Vit E, beta-carotene	*NEJM* 1994; **330**: 1029–1035
2	0	3	4	Hormone Replacement Ther.	*JAMA* 2003; **289**: 2651–2662, 2663–2672, 2673–2684
3	0	1	2	Vit E, beta-carotene	*JNCI* 2005; **97**: 481–488
4	0	0	3	Vit E	*JAMA* 2005; **293**: 1338–1347
5	0	0	3	Low Fat	*JAMA*. 2006; **295**: 655–666
6	0	0	3	Vit D, Calcium	*NEJM* 2006; **354**: 669–683
7	0	0	2	Folic acid, Vit B6, B12	*NEJM* 2006; **354**: 2764–2772
8	0	0	2	Low Fat	*JAMA* 2007; **298**: 289–298
9	0	0	12	Vit C, Vit E, beta-carotene	*Arch Intern Med* 2007; **167**: 1610–1618
10	0	0	12	Vit C, Vit E	*JAMA* 2008; **300**: 2123–2133
11	0	0	3	Vit E, Selenium	*JAMA* 2009; **301**: 39–51
12	0	0	3	HRT + Vitamins	*JAMA* 2002; **288**: 2431–2440
Totals	0	5	52		

Source: Young and Karr, Significance, Sept. 2011

Related work: estimation, selection, testing, 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).
  - Bayesian estimation: Rajaratnam-Massam-Carvalho (2008)
  - 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.

Outline

Correlation matrix and its support set

- $p \times n$: measurement matrix. $\mathbb{X} \sim \mathcal{N}(\mu, \mathbf{\Sigma} \otimes \mathbf{I}_n)$

$$\mathbb{X} = \begin{bmatrix} x_{11} & \cdots & x_{1n} \\ \vdots & \ddots & \vdots \\ x_{p1} & \cdots & x_{pn} \end{bmatrix} = [\mathbf{X}_1, \ldots, \mathbf{X}_n]$$

- $\mathbf{\Sigma} = E[(\mathbf{X}_1 - \mu)(\mathbf{X}_1 - \mu)^T]$ is $p \times p$ sparse covariance matrix
- $\mathbf{\Gamma}$ is $p \times p$ sparse correlation matrix

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

- Adjacency matrix: $\mathbf{A}_o = h_0(\mathbf{\Gamma})$,

$$h_\rho(u) = \frac{1}{2}\left(\mathrm{sgn}(|u| - \rho) + 1\right)$$

- Connectivity support set: $\mathbf{S}_o = \mathbf{S}_o^{(1)} = I(\mathtt{sum}(\mathbf{A}_o) > 1)$
- Hub degree $\geq \delta$ support set: $\mathbf{S}_o^{(\delta)} = I(\mathtt{sum}(\mathbf{A}_o) > \delta)$

Empirical estimation of correlation and support set

- $p \times p$ sample covariance matrix

$$\hat{\mathbf{\Sigma}} = \mathbb{X}(\mathbf{I} - \frac{1}{n}\mathbf{1}\mathbf{1}^T)\mathbb{X}^T \quad \frac{1}{n-1}$$

- $p \times p$ sample correlation matrix

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

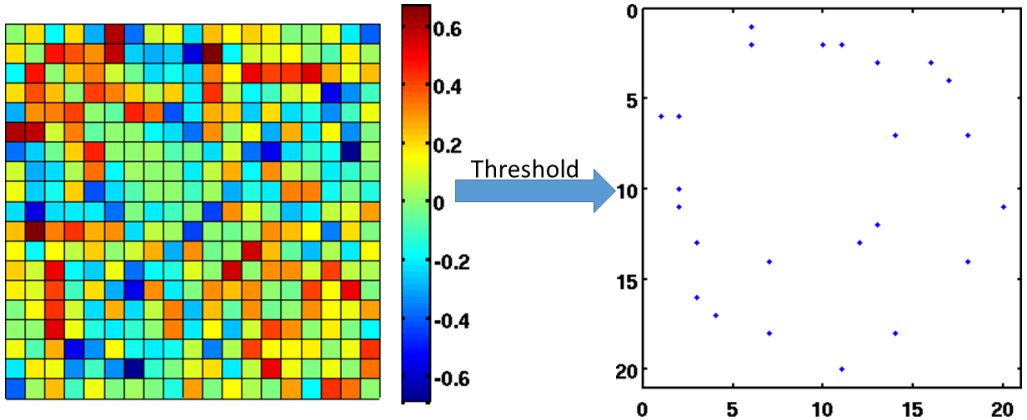
- Sample estimator of adjacency matrix at correlation level $\rho \in [0, 1]$:

$$\hat{\mathbf{A}}_o(\rho) = h_\rho(\mathbf{R})$$

- Sample estimator of connectivity support $\mathbf{S}_o(\rho)$ at level $\rho \in [0, 1]$:

$$\hat{\mathbf{S}}_o(\rho) = I(\mathtt{sum}(\hat{\mathbf{A}}_o(\rho)) > \delta)$$

Estimation vs support recovery vs screening for dependency



- **Correlation screening and detection**: false positive error

$$P_0(N_\rho > 0)$$

$N_\rho = \mathrm{card}\{\hat{\mathbf{S}}_o(\rho)\}$ is number of discoveries above threshold ρ.

- **Support recovery**: support misclassification error

$$P_\Sigma(\hat{\mathbf{S}}_o(\rho) \,\Delta\, \mathbf{S}_o \neq \phi)$$

- **Covariance estimation**: Frobenius norm error

$$\|\boldsymbol{\Sigma} - \hat{\boldsymbol{\Sigma}}\|_F$$

- **Uncertainty quantification**: estimation of estimator tail probabilities

Asymptotic regimes (H-R 2011, 2012, 2014, 2015)

Asymptotic framework	Terminology	Sample size	Dimension	Application setting	References
		n	p		
Classical (or sample increasing)	small dimensional	$\longrightarrow \infty$	fixed	"small data"	Fisher [28, 29], Rao [68, 69], Neyman and Pearson [61], Wilks [84], Wald [79, 80, 81, 82], Cramér [16, 15], Le Cam [51, 52], Chernoff [13], Kiefer and Wolfowitz[46], Bahadur [3], Efron [22]
Mixed asymptotics	high dimensional	$\longrightarrow \infty$	$\longrightarrow \infty$	"medium sized" data (mega or giga scales)	Donoho [20], Zhao and Yu [87], Meinshausen and Bühlmann [58], Candès and Tao [10], Bickel, Ritov, and Tsybakov[6], Peng, Wang, Zhou, and Zhu [64], Wainwright [77, 78], Khare, Oh, and Rajaratnam, [44]
	very high dimensional	$\longrightarrow \infty$	$\longrightarrow \infty$		
	ultra high dimensional	$\longrightarrow \infty$	$\longrightarrow \infty$		
Purely high dimensional	purely high dimensional	fixed	$\longrightarrow \infty$	"Big Data" (tera, peta and exascales)	Hero and Rajaratnam [35] Hero and Rajaratnam [36] Firouzi, Hero and Rajaratnam [25]

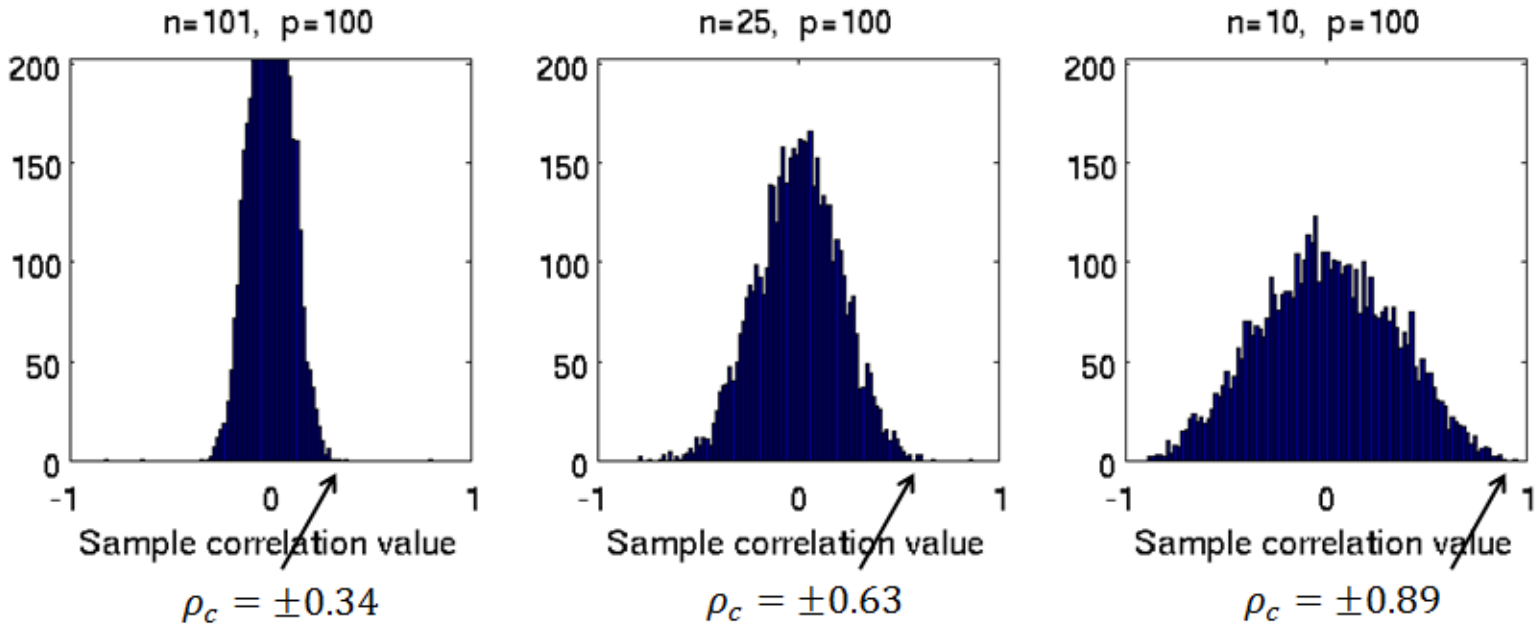
- Classical asymptotics: $n \to \infty$, p fixed ('small data')
- Mixed high D asymptotics: $n \to \infty$, $p \to \infty$ ('Medium data')
- Purely high D asymptotics: n fixed, $p \to \infty$ ('Big data')

It is important to design the procedure for the prevailing sampling regime

• H and Rajaratnam, "Large scale correlation mining for biomolecular network discovery," in Big data over networks, Cambridge 2015.

• H and Rajaratnam, "Foundational principles for large scale inference," IEEE Proceedings 2015.

Purely high D: phase transitions (H-R 2011, 2012, 2014)



- Impossible to reliably detect small correlations with finite n
- Possible to reliably detect large correlations even when $n \ll p$
- Critical threshold ρ_c on mean number of spurious discoveries

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

- $c_n = O(n^{-3/2})$ is only weakly dependent on $\mathbf{\Sigma}$ if block sparse

Purely high D convergence theorem (H-R 2012)

Asymptotics of hub screening[1]: (H and Rajaratnam 2012): Assume that columns of $\mathbb{X}$ are i.i.d. with bounded elliptically contoured density and row sparse covariance $\mathbf{\Sigma}$.

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

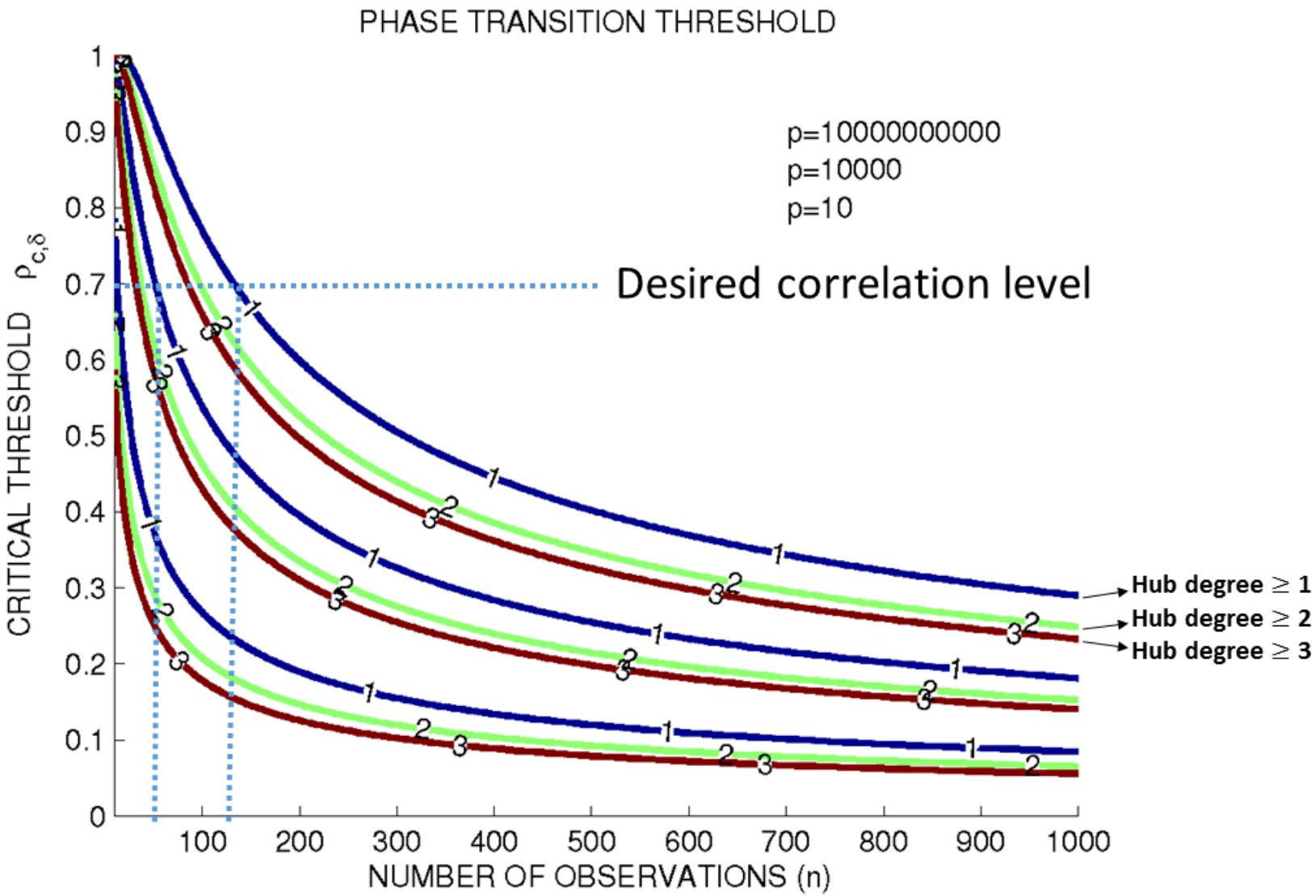
$$P(N_{\delta,\rho} > 0) \to \begin{cases} 1 - \exp(-\lambda_{\delta,\rho,n}/2), & \delta = 1 \\ 1 - \exp(-\lambda_{\delta,\rho,n}), & \delta > 1 \end{cases}.$$

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

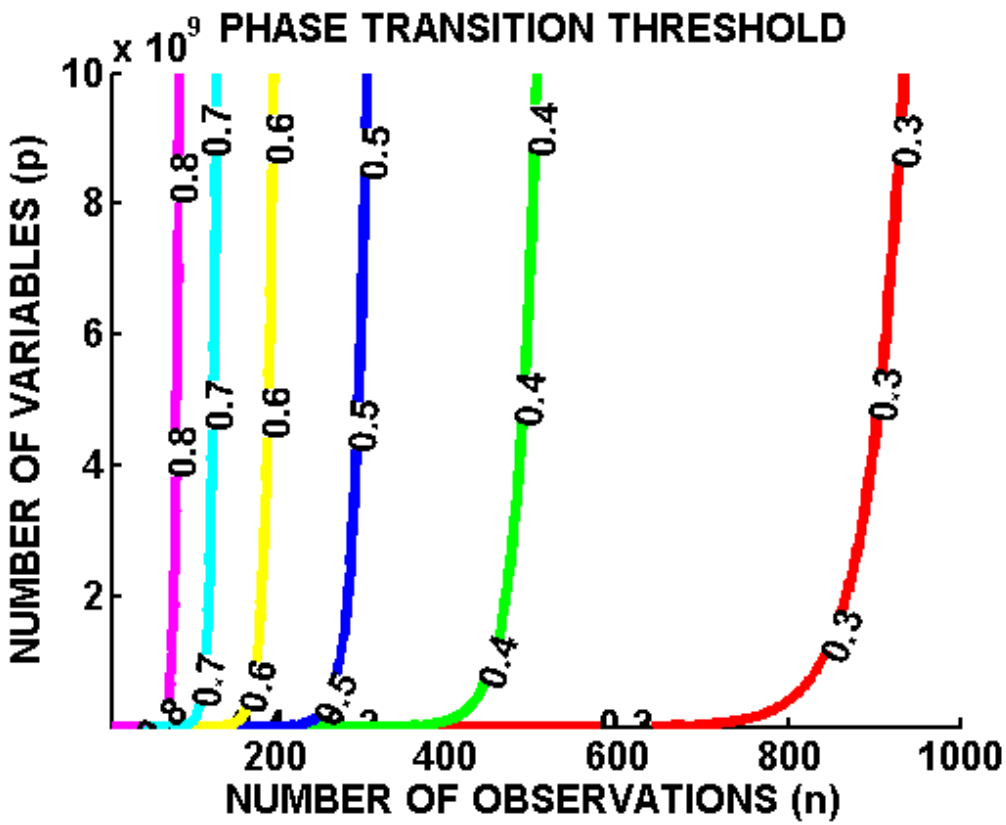
$$P_0(\rho,n) = 2B((n-2)/2,1/2)\int_{\rho}^{1}(1-u^2)^{\frac{n-4}{2}}\,du$$

[1] Generalized to local screening in (Firouzi-H 2013) and complex valued screening in (Firouzi-W-H 2014)

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



Critical phase transition threshold in n and p ($\delta = 1$)



- H and Rajaratnam, "Foundational principles for large scale inference," IEEE Proceedings 2015.
- H and Rajaratnam, "Large scale correlation mining for biomolecular network discovery," in Big data over networks, Cambridge 2015.

Outline

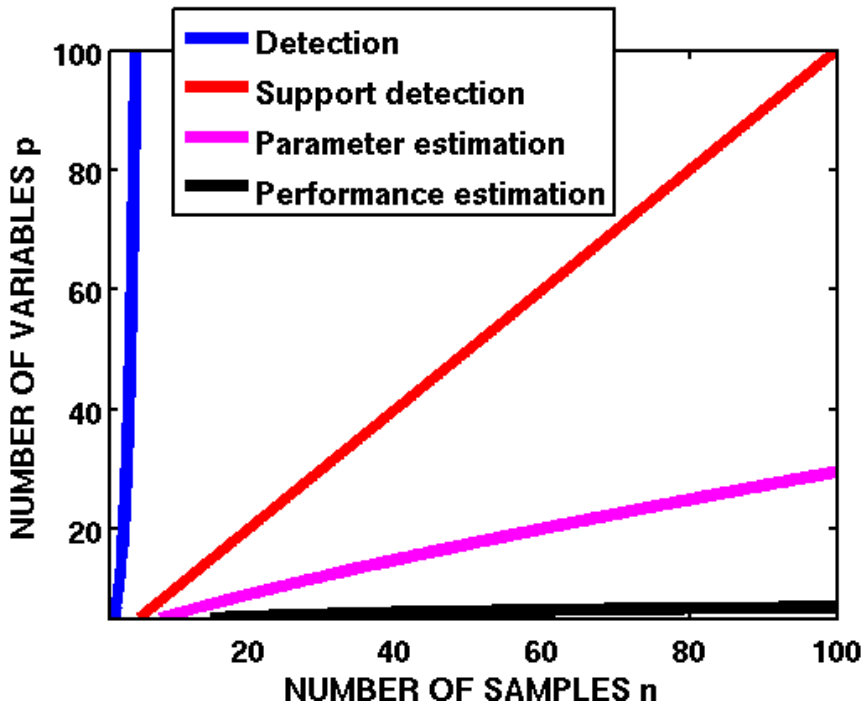
Sample complexity regimes for different tasks

Task	Screening	Detection	Support detection	Param. estimation	Perform. estimation
Risk	$P(N_e > 0)$	$P(N_e > 0)$	$P(\text{card}\{\mathcal{S}\Delta\hat{\mathcal{S}}\} = \phi)$	$E[\|\|\mathbf{\Omega} - \hat{\mathbf{\Omega}}\|\|_F^2]$	$\int E[(f_\Omega(\mathrm{x}) - \hat{f}(\mathrm{x}))^2]d\mathrm{x}$
Bound	$1 - e^{-\kappa_n}$	$pe^{-n\alpha}$	$2^p e^{-n\alpha}$	$\frac{p \log p}{n}\alpha$	$n^{-2/(1+p)}\alpha$
Regimes	$\frac{\log p}{n} \to \infty$	$\frac{\log p}{n} \to \alpha$	$\frac{p}{n} \to \alpha$	$\frac{p \log p}{n} \to \alpha$	$\frac{p}{\log n} \to \alpha$
Threshold	$\rho_c \to 1$	$\rho_c \to \rho^*$	$\rho_c \to 0$	$\rho_c \to 0$	$\rho_c \to 0$

H and Rajaratnam, "Foundational principles for large scale inference," IEEE Proceedings 2015

- Unifying framework: value-of-information for specific tasks
- Sample complexity regime specified by # available samples
- Some of these regimes require knowledge of sparsity factor
- From L to R, regimes require progressively larger sample size

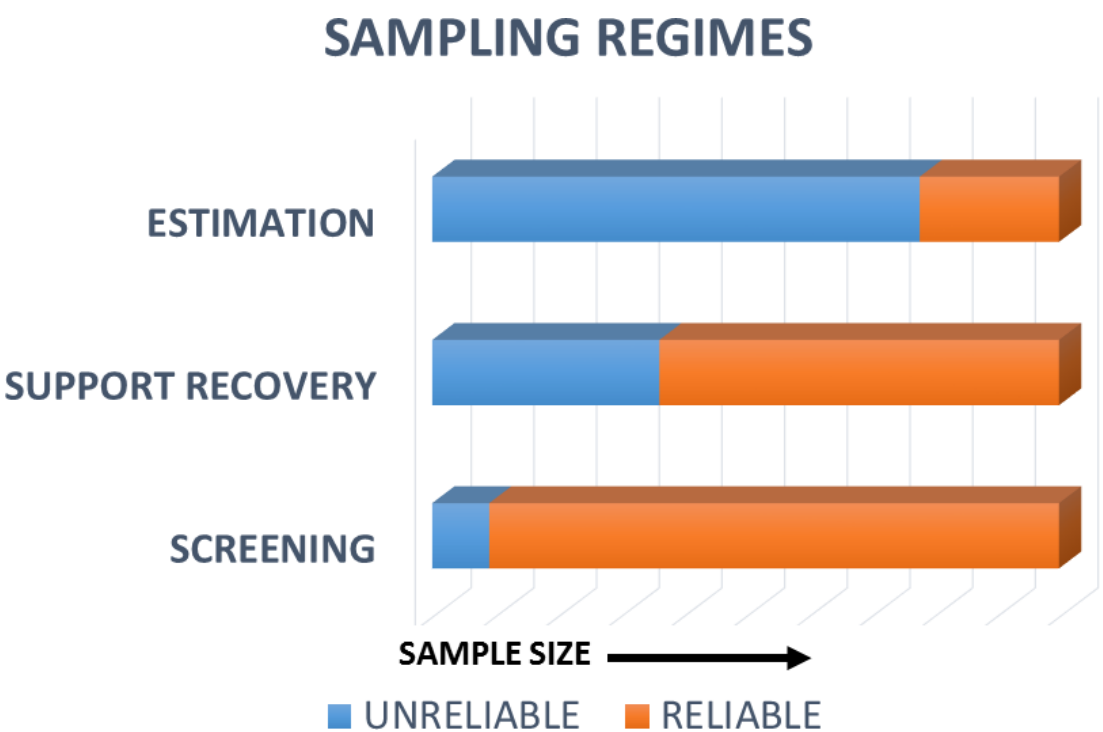
Sample complexity regimes for different tasks



H and Rajaratnam, "Foundational principles for large scale inference," IEEE Proceedings 2015

- There are niche regimes for reliable screening, detection, . . . , performance estimation
- Smallest amount of data needed to screen for high correlations
- Largest amount of data needed to quantify uncertainty

Implication: adapt inference task to sample size

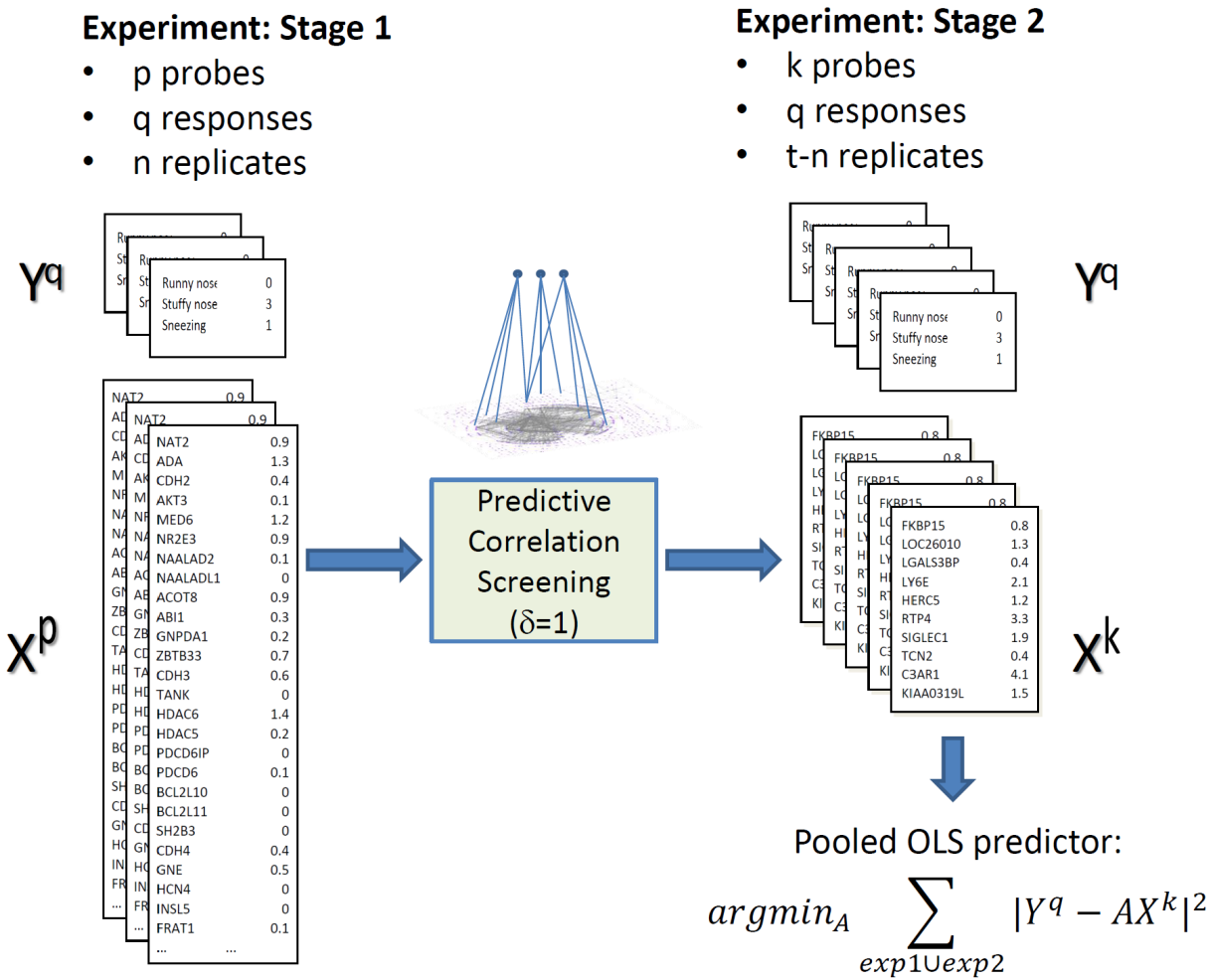


Dichotomous sampling regimes has motivated (Firouzi-H-R 2014):

- Progressive correlation mining
  $\Rightarrow$ match the mining task to the available sample size.
- Multistage correlation mining for budget limited applications
  $\Rightarrow$ Screen small exploratory sample prior to big collection

Outline

Sampling, Prediction and Adaptive Regression via Correlation Screening



• Firouzi, H and Rajaratnam, "Two-stage sampling, prediction and adaptive regression via correlation screening (SPARCS)," arxiv vol. 1502:06189, 2015.

SPARCS recovery of support of active variables

Theorem (Firouzi, H, Rajaratnam, 2013, 2015)

Assume that the response Y satisfies the following noiseless ground truth model:

$$Y = a_{i_1} X_{i_1} + a_{i_2} X_{i_2} + \cdots + a_{i_k} X_{i_k}$$

If $n \geq \Theta(\log p)$ then, with probability at least $1 - 1/p$, PCS recovers support of active variables π_0.

- Analogous to condition for LASSO support recovery (Obozinski, Wainright, Jordan 2008).
- The constant in $\Theta(\log p)$ is increasing in dynamic range coefficient

$$\frac{|\pi_0|^{-1} \sum_{l \in \pi_0} |a_l|}{\min_{j \in \pi_0} |a_j|} \in [1, \infty)$$

  - Worst case: high dynamic range in active regression coefficients.

Optimal pre-screening allocation under budget μ

Assume that: cost(acquisition of 1 sample of 1 variable)=1. Define

- Total budget for two-stage experiment: μ.
- Number of selected variables k. Total number of samples t.

To meet budget t, n, k, p must satisfy:

$$np + (t - n)k \leq \mu$$

Theorem

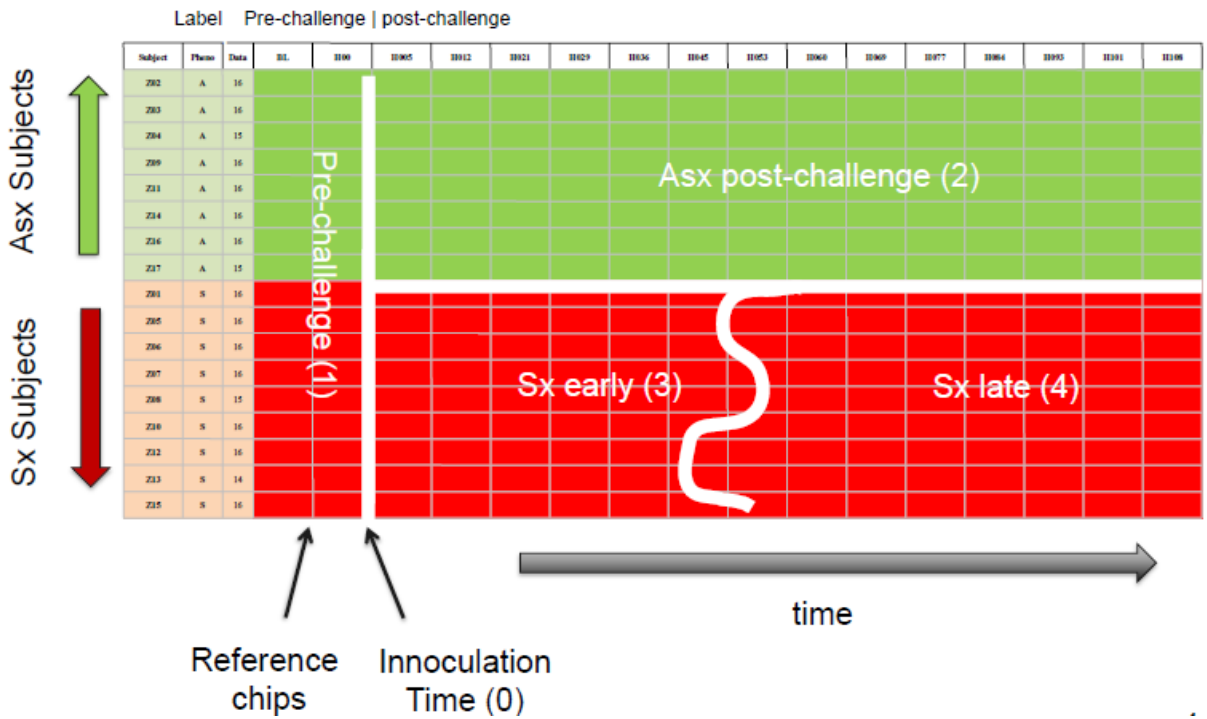
MSE optimal pre-screening allocation rule for two-stage predictor

$$n = \begin{cases} O(\log t), & c(p-k)\log t + kt \leq \mu \\ 0, & o.w. \end{cases}$$

When budget is tight skip stage 1 ($n = 0$).

Outline

Flu challenge experiment



Zaas *et al*, Cell, Host and Microbe, 2009

Chen *et al*, IEEE Trans. Biomedical Eng, 2010

Chen *et al* BMC Bioinformatics, 2011

Puig *et al* IEEE Trans. Signal Processing, 2011

Huang *et al*, PLoS Genetics, 2011

Woods *et al*, PLoS One, 2012

Bazot *et al*, BMC Bioinformatics, 2013

Zaas *et al*, Science Translation Medicine, 2014

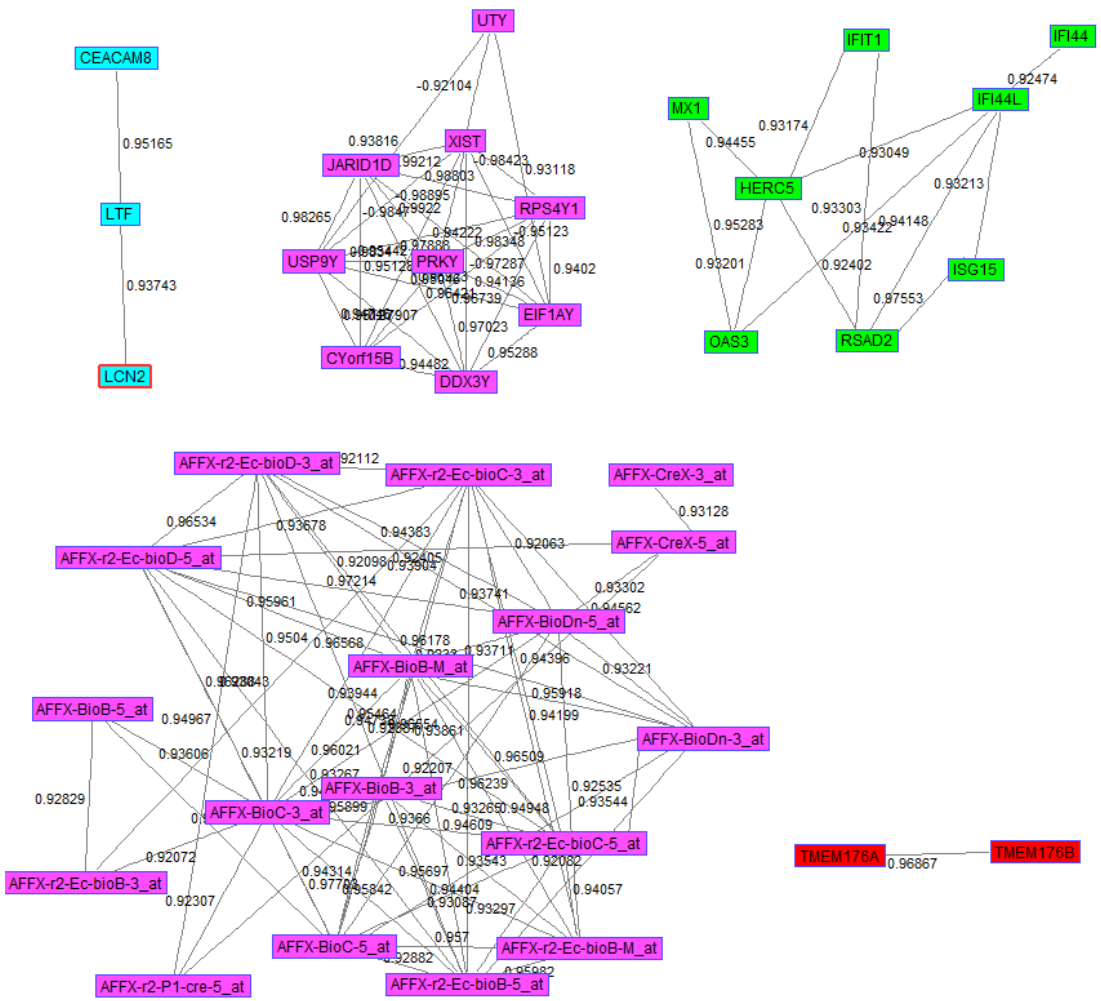
Critical threshold ρ_c for H3N2 DEE2

Samples fall into 3 categories

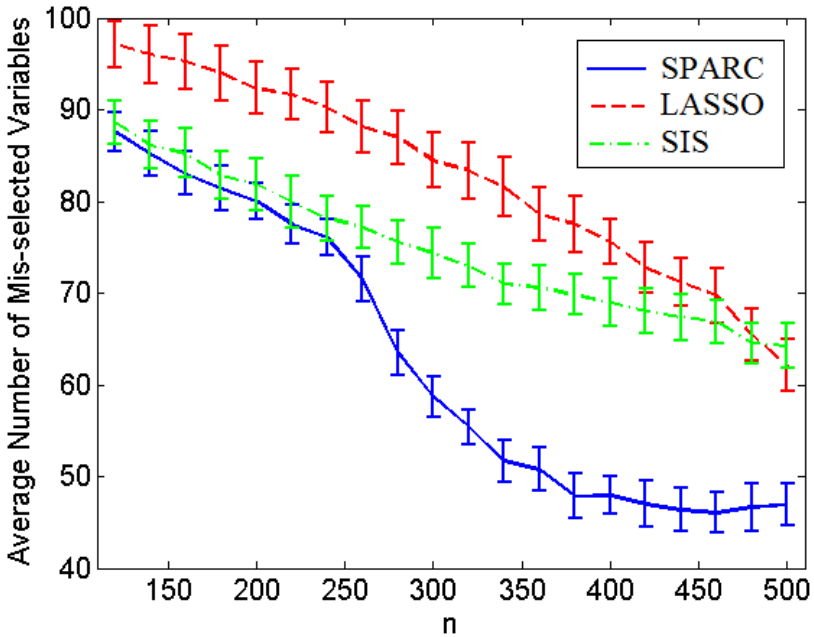
- Pre-inoculation samples
  - Number of Pre-inoc. samples: $n = 34$
  - Critical threshold: $\rho_c = 0.70$
  - 10^{-6} FWER threshold: $\rho = 0.92$
- Post-inoculation symptomatic samples
  - Number of Post-inoc. Sx samples: $n = 170$
  - Critical threshold: $\rho_c = 0.36$
  - 10^{-6} FWER threshold: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

Susceptibility: Correlation-mining the pre-inoc. samples

- Screen correlation at FWER 10^{-6}: 1658 genes, 8718 edges
- Screen partial correlation at FWER 10^{-6}: 39 genes, 111 edges



Prediction: SPARCS comparisons to LASSO and SIS



Symptom	RMSE: LASSO	RMSE: SIS	RMSE: SPARC
Runny Nose	0.7182	0.6896	**0.6559**
Stuffy Nose	0.9242	**0.7787**	0.8383
Sneezing	0.7453	0.6201	**0.6037**
Sore Throat	0.8235	0.7202	**0.5965**
Earache	**0.2896**	0.3226	0.3226
Malaise	1.0009	**0.7566**	0.9125
Cough	0.5879	0.7505	**0.5564**
Shortness of Breath	0.4361	0.5206	**0.4022**
Headache	0.7896	0.7500	**0.6671**
Myalgia	0.6372	0.5539	**0.4610**
Average for all symptoms	0.6953	0.6463	**0.6016**

Support recovery (simu)

Prediction (real data)

- Firouzi, H and Rajaratnam, "Predictive correlation screening: Application to two-stage predictor design in high dimension," AISTATS 2013
- Firouzi, H and Rajaratnam, "Two-stage sampling, prediction and adaptive regression via correlation screening (SPARCS)," arxiv vol. 1502:06189, 2015.

Outline

Conclusions

What we covered

- Asymptotic correlation mining theory developed for "Purely high" dimensional ("big data") setting:

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

- 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 "Purely high" dimensional ("big data") 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

- 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)
- Quickest change detection and correlation mining (Banerjee and H, 2015)


T. Banerjee and A. Hero, "Non-parametric quickest change detection for large scale random matrices," in *IEEE Intl Symposium on Information Theory*, 2015.


H. Firouzi, A. 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*, 2013.


H. Firouzi, B. Rajaratnam, and A. Hero, "Two-stage sampling, prediction and adaptive regression via correlation screening (SPARCS)," *arxiv*, vol. 1502:06189, , Feb 2015.


H. Firouzi, D. Wei, and A. 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*, 2013.


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


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


K. Greenewald, T. Tsiligkaridis, and A. 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*, vol. 58, no. 9, pp. 6064–6078, 2012.
available as Arxiv preprint arXiv:1109.6846.


A. Hero and B. Rajaratnam, "Large scale correlation screening," *Journ. of American Statistical Association*, vol. 106, no. 496, pp. 1540–1552, Dec 2011.
Available as Arxiv preprint arXiv:1102.1204.


A. Hero and B. Rajaratnam, "Large scale correlation mining for biomolecular network discovery," in *Big data over networks*, S. Cui, A. Hero, Z. Luo, and J. Moura, editors. Cambridge Univ Press, 2015. Preprint available in Stanford University Dept. of Statistics Report series.


A. Hero and B. Rajaratnam, "Foundational principles for large scale inference: Illustrations through correlation mining," *Proceedings of the IEEE (also available as arxiv 1502:06189)*, to appear, 2015.


K. Todros and A. O. Hero, "On measure transformed canonical correlation analysis," *Signal Processing, IEEE Transactions on*, vol. 60, no. 9, pp. 4570–4585, 2012.


K. Todros and A. Hero, "Measure transformed canonical correlation analysis with application to financial data," in *Sensor Array and Multichannel Signal Processing Workshop (SAM), 2012 IEEE 7th*, pp. 361–364. IEEE, 2012.


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)*, vol. 61, no. 7, pp. 1743–1755, 2013.


T. Tsiligkaridis, A. Hero, and S. Zhou, "On convergence of Kronecker graphical lasso algorithms," *IEEE Trans. on Signal Processing*, vol. 61, no. 9, pp. 1743 –1755, 2013.