Learning correlations in large scale data

Alfred Hero

University of Michigan

Aug 10, 2015

Acknowledgments

Students and collaborators on correlation learning

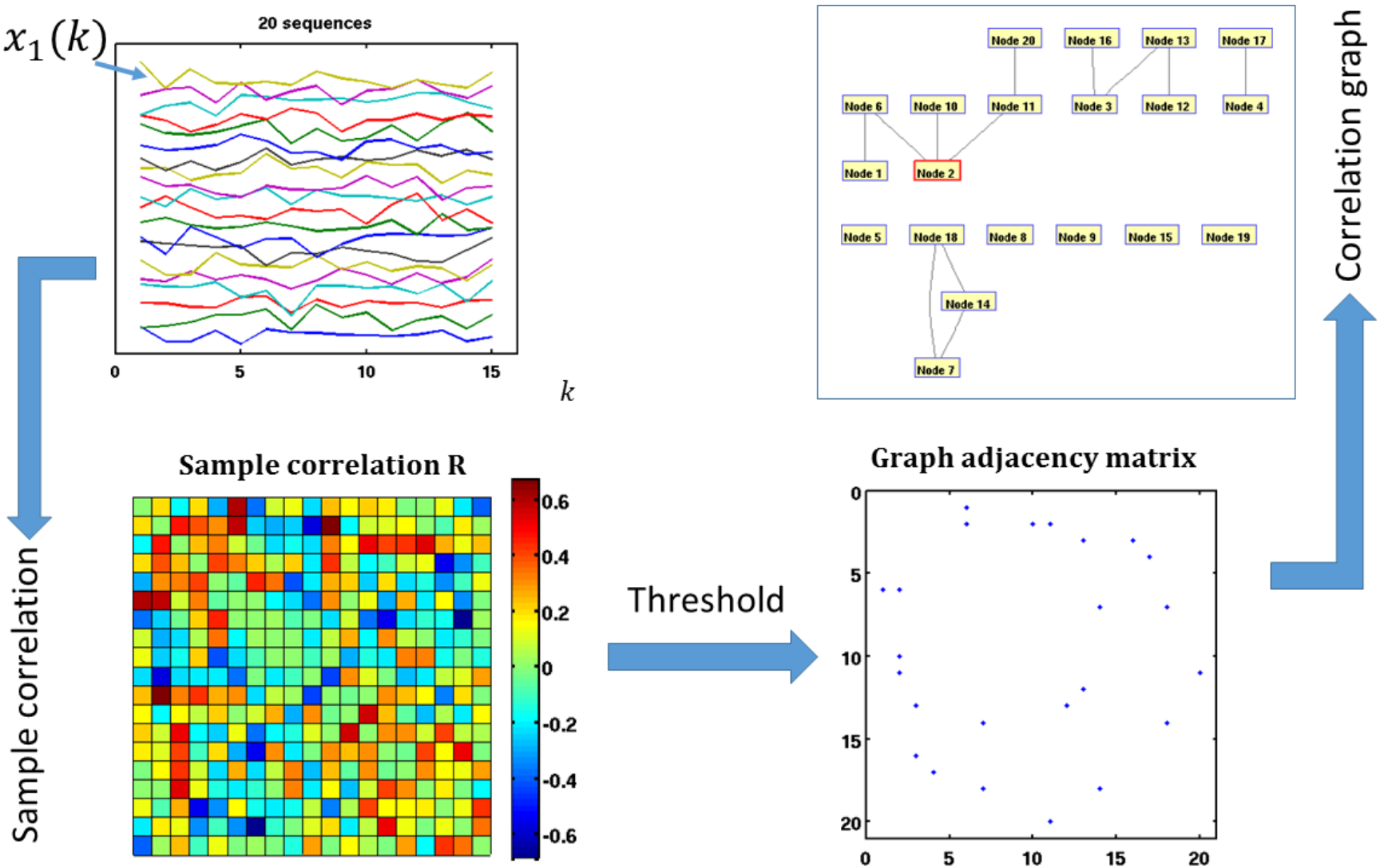
- Kevin Moon (UM ECE Student)
- Brandon Oselio (UM ECE Student)
- Kristjan Greenewald (UM ECE Student)
- Yaya Zai (UM Bioinformatics student)
- Elizabeth Hou (UM Statistics Student)
- Yun Wei (UM Math Student)
- Taposh Banerjee (UM Postdoc)
- Hamed Firouzi (Goldman)
- Ted Tsiligkaridis (MIT LL)
- Zaoshi Meng (Vicarious)
- Mark Hsiao (Whispers)
- Dennis Wei (IBM Watson)
- Ami Wiesel (Hebrew Univ.)
- Bala Rajaratnam (Stanford)

Sponsors

- AFOSR Complex Networks Program (FA9550-09-1-0471)
- ARO MURI Value of Information Program
- ARO: Social Informatics Program
- DARPA Predicting Health and Disease Program
- NSF: Theoretical Foundations Program
- NIH Biomedical Imaging Institute

Outline

Data ⟶ correlation ⟶ adjacency ⟶ network



Why is correlation important in SP/ML?

- 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

Why is correlation important in SP/ML?

- 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 cov 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 $H_1 : \mathbf{\Sigma}_X = \mathbf{\Sigma}_1$

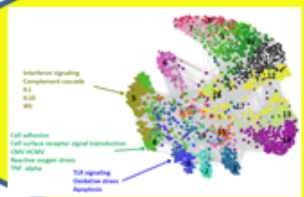
$$\overline{\mathbf{X}}^T(\mathbf{\Sigma}_0^{-1} - \mathbf{\Sigma}_1^{-1})\overline{\mathbf{X}} \underset{H_0}{\overset{H_1}{\gtrless}} \eta$$

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

$$\frac{\overline{\mathbf{X}}^T\mathbf{\Sigma}_0^{-1}\overline{\mathbf{X}}}{\overline{\mathbf{X}}^T\overline{\mathbf{X}}} \underset{H_0}{\overset{H_1}{\gtrless}} \eta$$

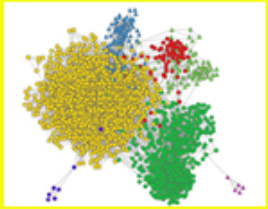
Learning correlations and complex network discovery

Molecular pathways



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

Spammer communities

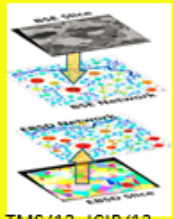


CEAS 10', DMKD '12,, JSTSP '14 ...

Social collaborative retrieval nets

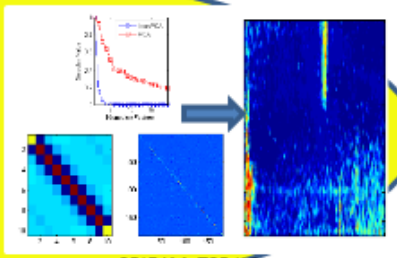
CAMSAP '13, WSDM '14, JSTSP '14 ...

Grain networks



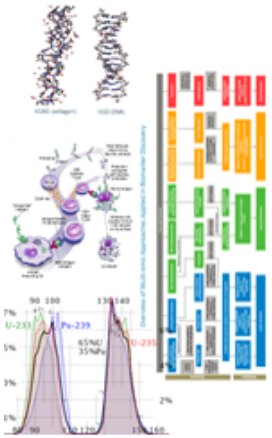
TMS '13, ICIP '13, ...

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

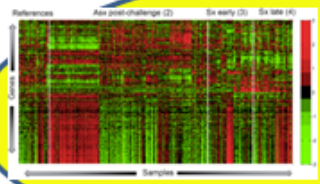
Error control α

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

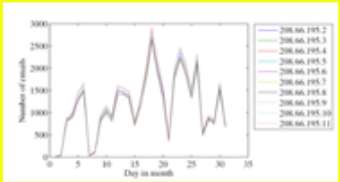
Budget $

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

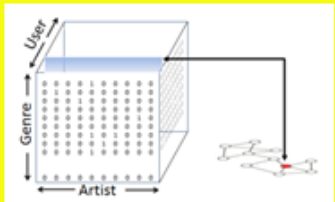
HIGH DIMENSIONAL DATA



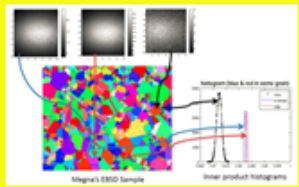
mRNA expression



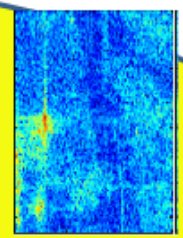
Email volume



Personal/Social data



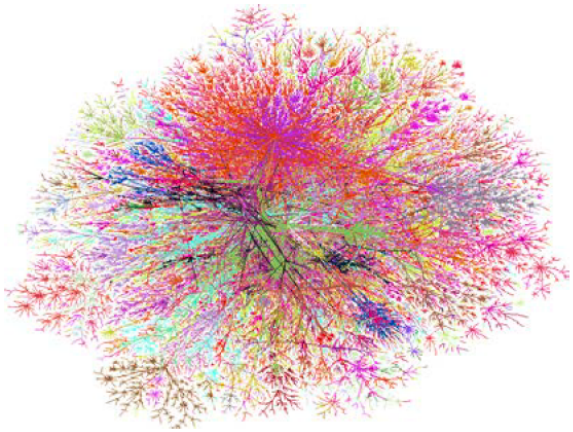
Microscopy data



Gotcha GMTI

Big Data aspects of correlation mining

O/I correlation

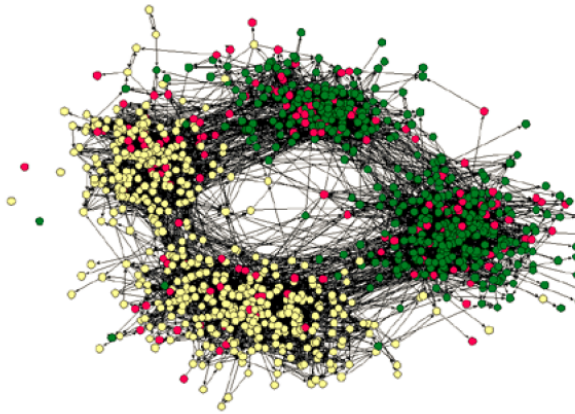


The Internet
(Burch and Cheswick, 1998)

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, precision)
  - Scalability of methods to exascale data 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

Outline

Tasks: estimation, identification, detection, screening

- Covariance estimation: sparse regularized l_2 or $l_{\mathcal{F}}$
  - Banded covariance estimation: Bickel-Levina (2008) Sparse eigendecomposition model: Johnstone-Lu (2007)
  - Stein shrinkage estimator: Ledoit-Wolf (2005), Chen-Weisel-Eldar-H (2010)
- Correlation identification 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)
- Correlation detection: 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.

Learning a 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

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

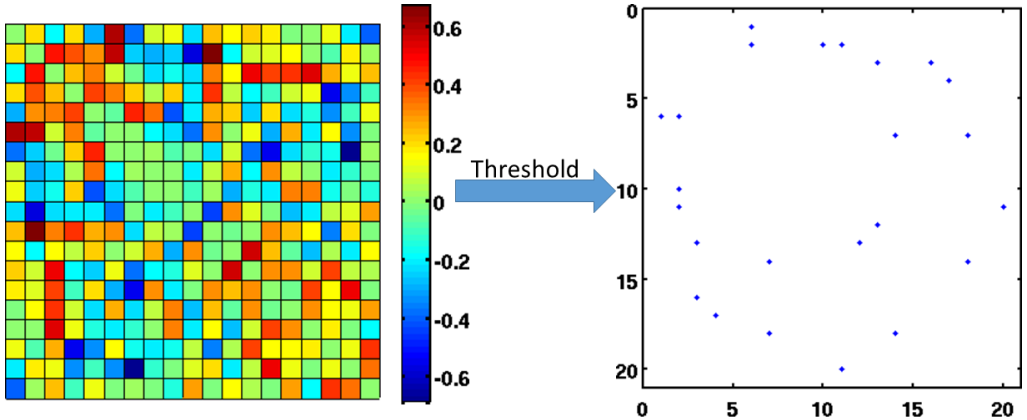
- Sample estimator of adjacency matrix at correlation level $\rho \in [0, 1]$:

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

- 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

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

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

Outline

Learning rates: sample-size vs. camplexity regimes

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 ('Low complexity')
- Mixed high D asymptotics: $n \to \infty$, $p \to \infty$ ('Medium complexity')
- Purely high D asymptotics: n fixed, $p \to \infty$ ('High complexity')

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.

Outline

Example: covariance estimation in high dimension

- n: # of available samples (sample size)
- $P = p(p+1)/2 = O(p^2)$: # of unknown model params (complexity)

Standard covariance matrix (SCM) $\hat{\mathbf{\Sigma}} = \mathbf{S}_n$ requires

1. $n > p$ for $\hat{\mathbf{\Sigma}}^{-1}$ to exist
2. $n > p^2$ for accurate estimates in Frobenius norm

$$\|\mathbf{\Sigma} - \hat{\mathbf{\Sigma}}\|_F^2 = O\left(p^2/n\right), \qquad (p^2/n = \text{sample-complexity ratio})$$

Example: covariance estimation in high dimension

- n: # of available samples (sample size)
- $P = p(p+1)/2 = O(p^2)$: # of unknown model params (complexity)

Standard covariance matrix (SCM) $\hat{\mathbf{\Sigma}} = \mathbf{S}_n$ requires

1. $n > p$ for $\hat{\mathbf{\Sigma}}^{-1}$ to exist
2. $n > p^2$ for accurate estimates in Frobenius norm

$$\|\mathbf{\Sigma} - \hat{\mathbf{\Sigma}}\|_F^2 = O\left(p^2/n\right), \qquad (p^2/n = \text{sample-complexity ratio})$$

Structure reduces sample-complexity ratio: $\|\mathbf{\Sigma} - \hat{\mathbf{\Sigma}}\|_F^2 < O\left(p^2/n\right)$

Structure	**Error** $\|\mathbf{\Sigma} - \hat{\mathbf{\Sigma}}\|_F^2 =$
Toeplitz $\mathbf{\Sigma} = \mathrm{toeplitz}(r_1, \ldots, r_p)$	$O\left(p/n\right)$
Sparse: $\mathbf{\Sigma} = \mathbf{\Omega}_{sp}$, $\mathrm{nnz}(\mathbf{\Omega}_{sp}) = O(p)$	$O\left(p\log(p)/n\right)$
Low Kron rank $\mathbf{\Sigma} = \sum_{i=1}^{r} \mathbf{A}_i \bigotimes \mathbf{B}_i$	$O\left(r(S^2 + T^2)/n\right)$

- Bühlmann and van de Geer (2011)
- Tsiligkaridis and H (2014), Greenewald and H (2014)

Kronecker product of matrices

Let $\mathbf{A}$ be a $T \times T$ matrix and $\mathbf{B}$ be a $S \times S$ matrix. For $p = ST$ define the $p \times p$ matrix $\mathbf{C}$ by the Kronecker product factorization $\mathbf{C} = \mathbf{A} \bigotimes \mathbf{B}$ where

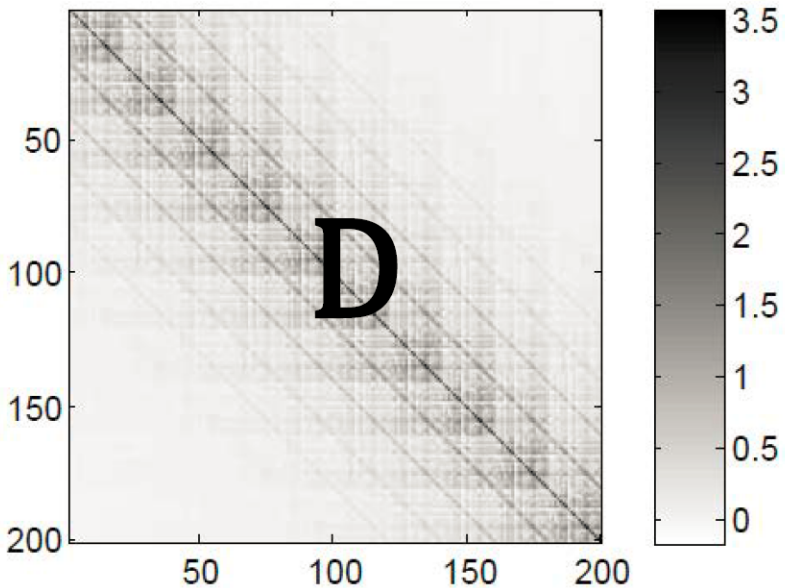
$$\mathbf{A} \bigotimes \mathbf{B} = \begin{bmatrix} a_{11}\mathbf{B} & \cdots & a_{1p}\mathbf{B} \\ \vdots & \ddots & \vdots \\ a_{p1}\mathbf{B} & \cdots & a_{pp}\mathbf{B} \end{bmatrix}$$

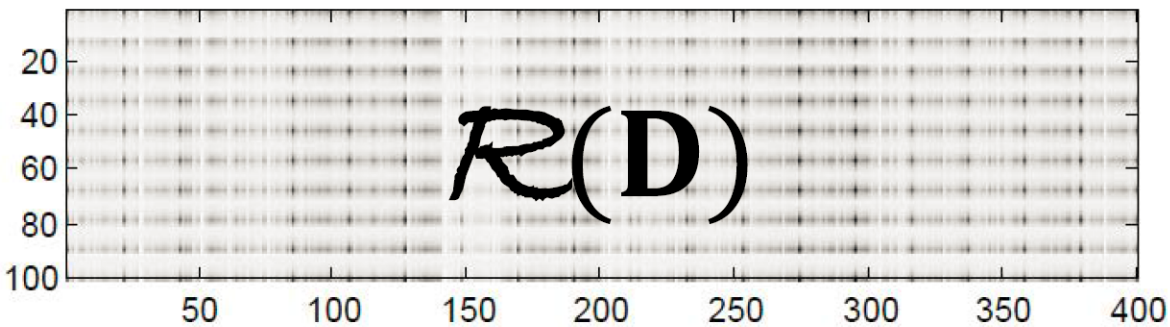
Kronecker product properties (VanLoan-Pitsianis 1992):

- $\mathbf{C}$ is p.d. if $\mathbf{A}$ and $\mathbf{B}$ are p.d.
- $\mathbf{C}^{-1} = \mathbf{A}^{-1} \bigotimes \mathbf{B}^{-1}$ if $\mathbf{A}$ and $\mathbf{B}$ are invertible.
- $|\mathbf{C}| = |\mathbf{A}|\ |\mathbf{B}|$
- For any $pq \times pq$ matrix $\mathbf{D}$

$$\|\mathbf{D} - \mathbf{A} \bigotimes \mathbf{B}\|_F^2 = \|\mathcal{R}(\mathbf{D}) - \mathrm{vec}(\mathbf{A})\mathrm{vec}(\mathbf{B})^T\|_F^2$$

$\mathcal{R}$ is permutation operator mapping $\mathbb{R}^{ST \times ST}$ to $\mathbb{R}^{T^2 \times S^2}$





Kronecker product model for covariance matrix

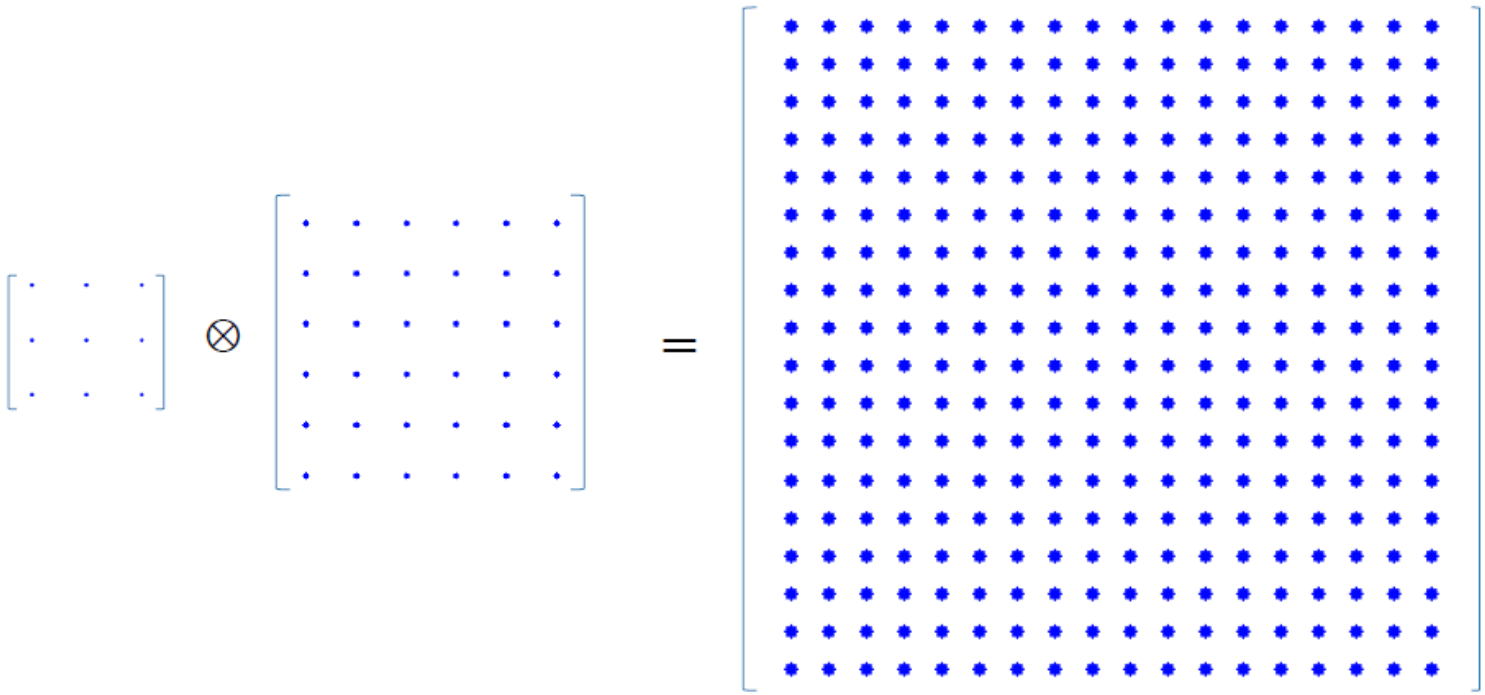


Figure: 18×18 covariance matrix has 18*17/2=153 unknown cross-correlation parameters. Kronecker product covariance model reduces this to $3 + 15 = 18$ parameters.

Leads to Kronecker MLE (matrix normal): Dawid (1981), Werner-Jansson-Stoica (2008), Tsiligkaridis-H-Zhou (2013)

Sparse Kronecker product model for covariance matrix

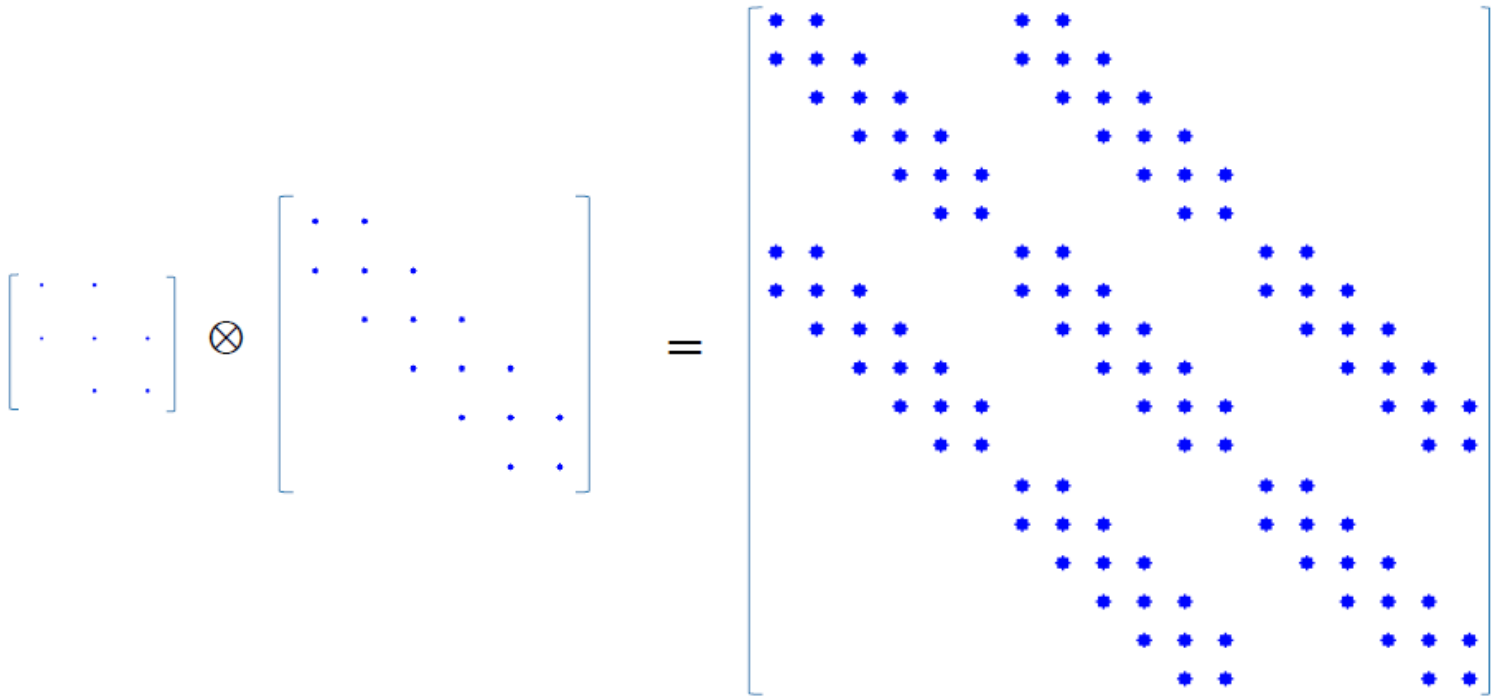


Figure: A sparse Kronecker product covariance model reduces number of parameters from 153 to 7 unknown correlation parameters.

Leads to KGlasso (sparse matrix normal): Allen-Tibshirani (2010), Yin-Li (2012), Tsiligkaridis-H-Zhou (2013)

Kronecker covariance matrix decomposition

Approximate $\mathbf{S}_n$ using Kronecker sum with r terms

$$\mathbf{S}(\mathbf{A}, \mathbf{B}) = \sum_{i=1}^{r} \mathbf{A}_i \bigotimes \mathbf{B}_i$$

Constraints: $\mathbf{A}_i$ and $\mathbf{B}_i$ such that $\mathbf{S}(\mathbf{A}, \mathbf{B})$ is n.n.d.

Many possible approximations (Kolda and Bader 2009)

- CANDECOMP/PARAFAC (CP) models (Carrol&Chang 1970, Harshman 1970)
- Tucker models (Tucker 1966, DeLauthauwer *et al* 2000)
- Other variants: INDSCAL, PARAFAC2, PARATUCK2 ...

Kronecker covariance matrix decomposition

Approximate $\mathbf{S}_n$ using Kronecker sum with r terms

$$\mathbf{S}(\mathbf{A}, \mathbf{B}) = \sum_{i=1}^{r} \mathbf{A}_i \bigotimes \mathbf{B}_i$$

Constraints: $\mathbf{A}_i$ and $\mathbf{B}_i$ such that $\mathbf{S}(\mathbf{A}, \mathbf{B})$ is n.n.d.

Many possible approximations (Kolda and Bader 2009)

- CANDECOMP/PARAFAC (CP) models (Carrol&Chang 1970, Harshman 1970)
- Tucker models (Tucker 1966, DeLauthauwer *et al* 2000)
- Other variants: INDSCAL, PARAFAC2, PARATUCK2 ...

Estimation procedure: minimize Frobenious norm error
$\Rightarrow \min \|\mathbf{S}_n - \mathbf{S}(\mathbf{A}, \mathbf{B})\|_F^2 = \min \|\mathcal{R}(\mathbf{S}_n) - \mathbf{R})\|_F^2$, $\mathbf{R} \in \mathbb{R}^{T \times S}$

$\Rightarrow$ can apply nuclear norm relaxation (Fazel 2002, Recht-Fazel-Parillo 2007, Hiriart-Urruty and Le 2011)

Estimation of Kronecker factors: Kronecker PCA

Relaxed Kronecker PCA on permuted $\mathbf{S}_n$

$$\hat{\mathbf{R}} = \text{amin}_{\mathbf{R}}\{\|\mathcal{R}(\mathbf{S}_n) - \mathbf{R}\|_F^2 + \lambda\|\mathbf{R}\|_*\}$$

Solution is explicit (Kronecker sum approximation):

Theorem (Tsiligkaridis-H 2013)

The solution to the relaxed Kronecker PCA minimization:

$$\hat{\mathbf{\Sigma}} = \mathcal{R}^{-1}\left(\hat{\mathbf{R}}\right), \quad \hat{\mathbf{R}} = \sum_{i=1}^{\min(S^2,T^2)} \left(\sigma_i - \frac{\lambda}{2}\right)_+ \mathbf{u}_i\mathbf{v}_i^T$$

- $(\sigma_k, \mathbf{u}_k, \mathbf{v}_k)$ *is the k-th component of the SVD of* $\mathcal{R}(\mathbf{S}_n)$

Tsiligkaridis and Hero, "Covariance Estimation in High Dimensions via Kronecker Product Expansions,", IEEE Trans. on TSP, 2013.

K-PCA estimator's MSE convergence rates

Theorem (Tsiligkaridis-H 2013)

Assume $\mathbf{S}_n \in \mathbb{R}^{ST \times ST}$ *is p.d and let* $M = \max(S, T, n)$. *Let* λ *satisfy*

$$\lambda = C(S^2 + T^2 + \log(M))/n$$

Then, with probability at least $1 - 2M^{-1/4C}$ *the K-PCA estimator* $\hat{\mathbf{\Sigma}}_{S,T,r}$ *of* $\mathbf{\Sigma}$ *satisfies:*

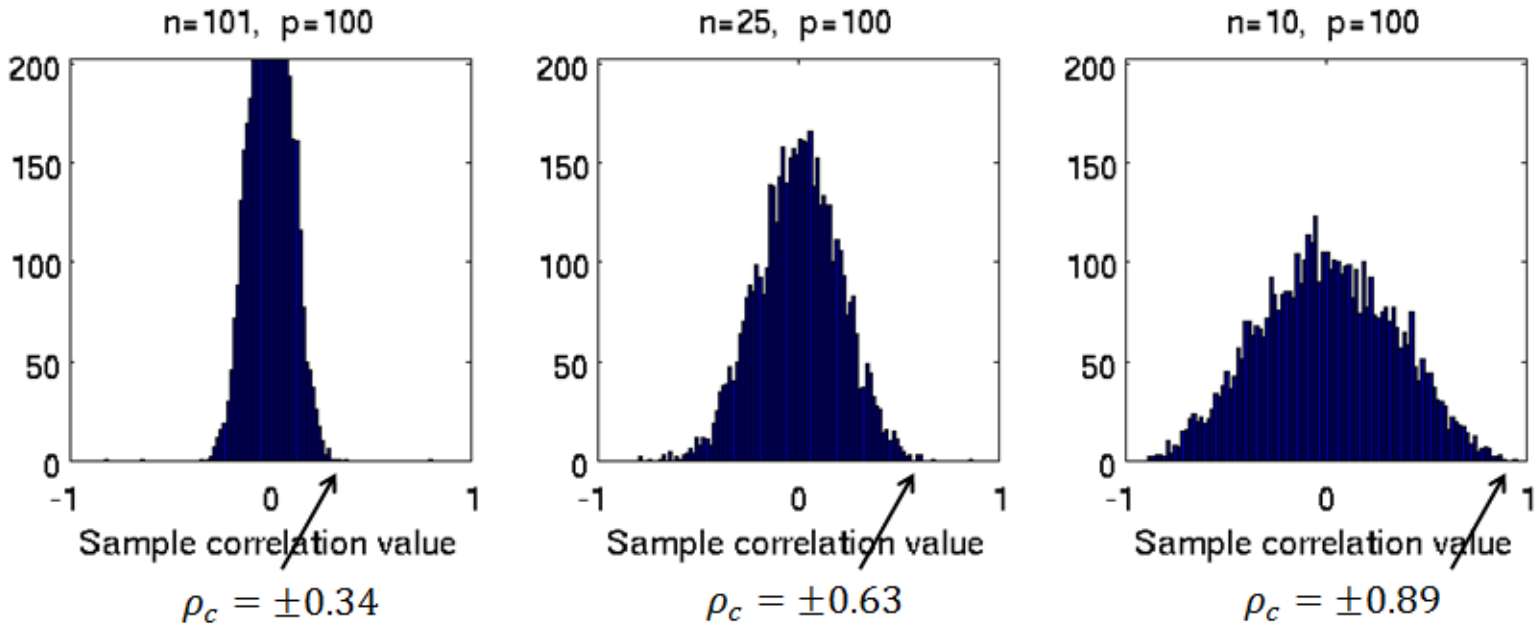
$$\begin{aligned} \|\hat{\mathbf{\Sigma}}_{p.q.r} - \mathbf{\Sigma}\|_F^2 \quad \leq \quad & \min_{\mathbf{R}:\mathrm{rank}(\mathbf{R}) \leq r} \|\mathbf{R} - \mathcal{R}(\mathbf{\Sigma})\|_F^2 \\ & + C^{`} \left(r(S^2 + T^2 + \log(M))/n \right) \end{aligned}$$

where $C^{`} = (1.5(1 + \sqrt{2})C)^2$.

Tsiligkaridis and Hero, "Covariance Estimation in High Dimensions via Kronecker Product Expansions,", IEEE Trans. on TSP, 2013.

Outline

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

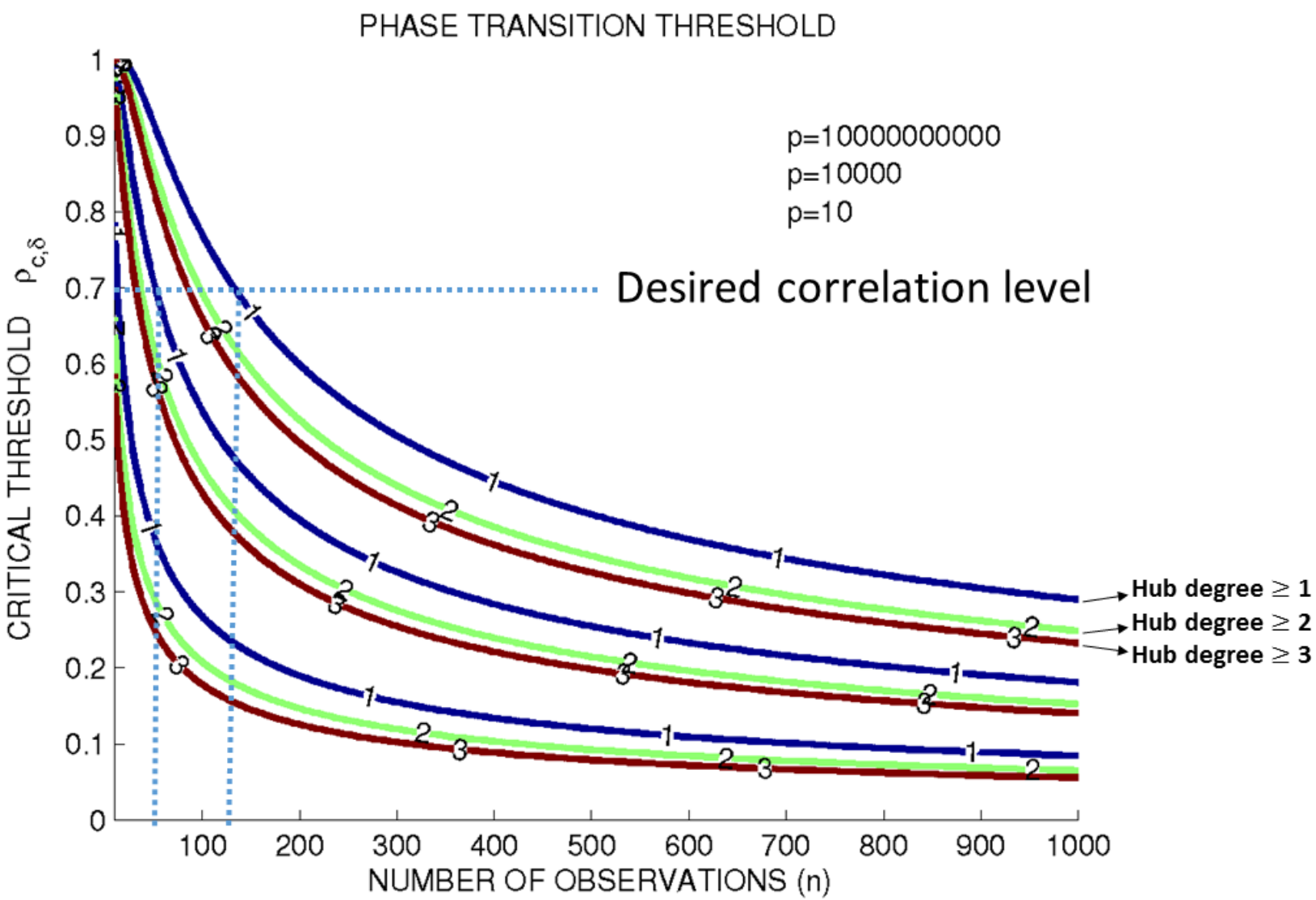
$$P(N_{\delta,\rho} > 0) \to \begin{cases} 1 - \exp(-\lambda_{\delta,\rho,n}/2), & \delta = 1 \\ 1 - \exp(-\lambda_{\delta,\rho,n}), & \delta > 1 \end{cases}.$$

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

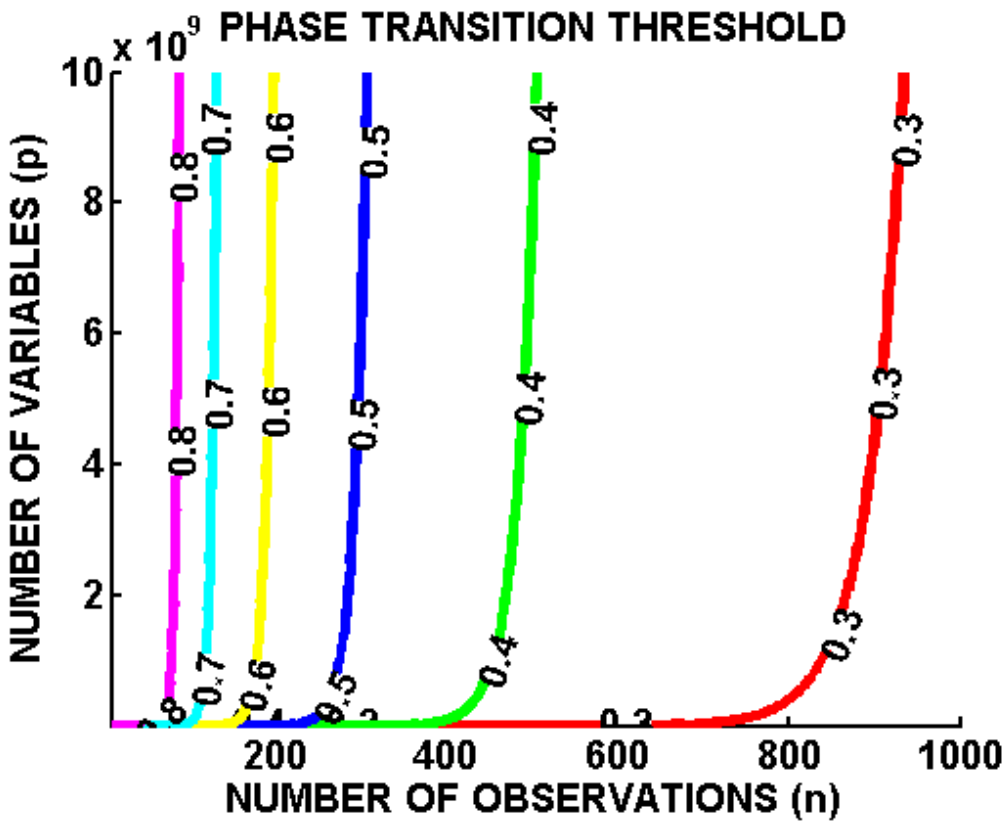
$$P_0(\rho,n) = 2B((n-2)/2,1/2)\int_{\rho}^{1}(1-u^2)^{\frac{n-4}{2}}\,du$$

[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.

Outline

Sample complexity regimes for different tasks

Sample complexity: How fast n must increase in p to maintain constant error:

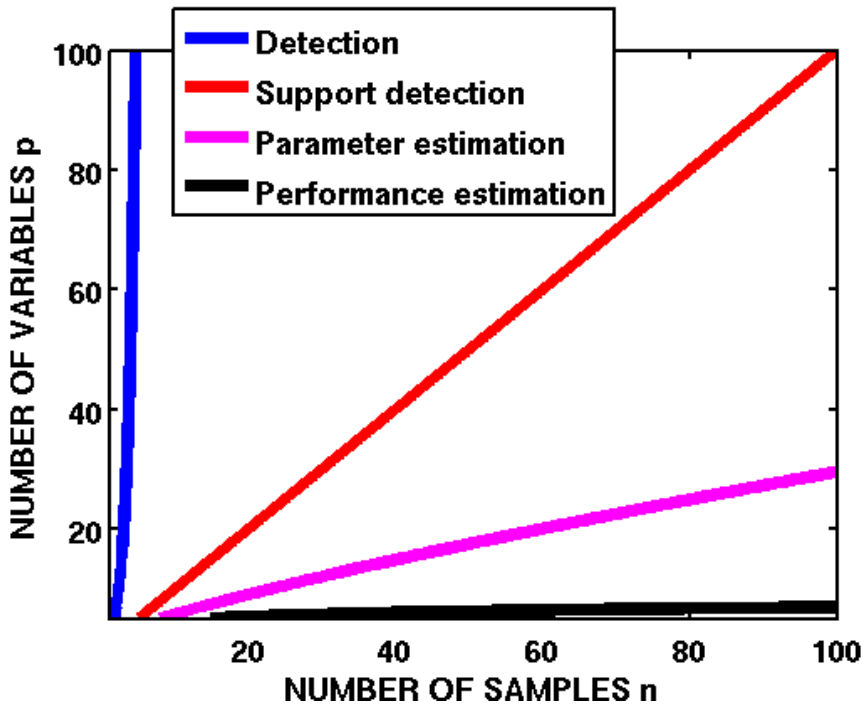
$$\lim_{\substack{n,p\to\infty \\ n=g(p)}} \text{error}(n,p) = c$$

Task	Screening	Detection	Identification	Estimation	Confidence
Error	$P(N_e > 0)$	$P(N_e > 0)$	$P(\{\mathcal{S}\Delta\hat{\mathcal{S}}\} = \phi)$	$E[\|\|\mathbf{\Omega} - \hat{\mathbf{\Omega}}\|\|_F^2]$	$E[(f_\Omega - \hat{f})^2]$
Bound	$1 - e^{-\kappa_n}$	$pe^{-n\beta}$	$2^{p^\nu} e^{-n\beta}$	$\frac{p\log p}{n}\beta$	$n^{-2/(1+p)}\beta$
Regimes	$\frac{\log p}{n} \to \infty$	$\frac{\log p}{n} \to \alpha$	$\frac{p^\nu}{n} \to \alpha$	$\frac{p\log p}{n} \to \alpha$	$\frac{p}{\log n} \to \alpha$
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

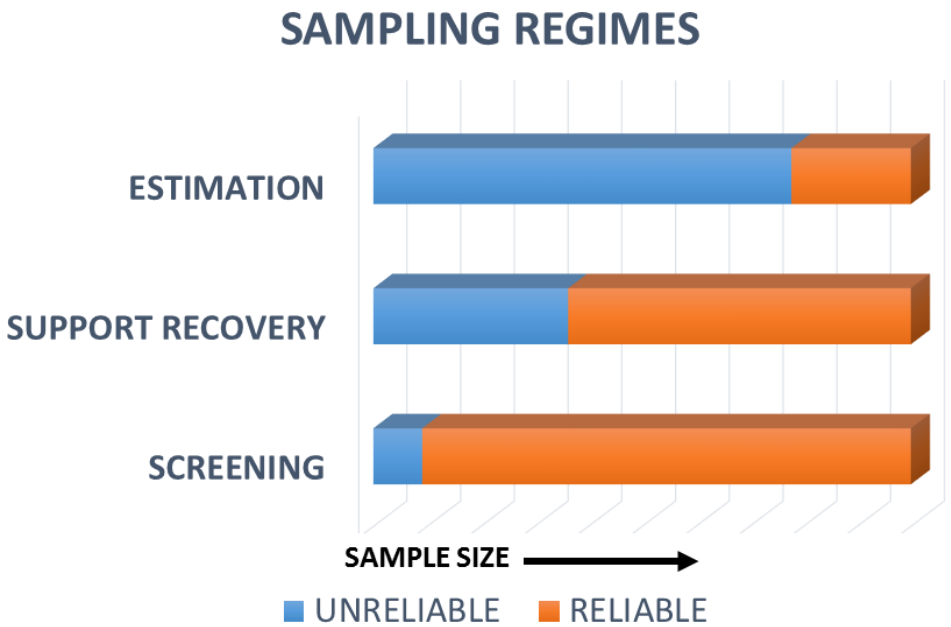
Sample complexity regimes for different tasks



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



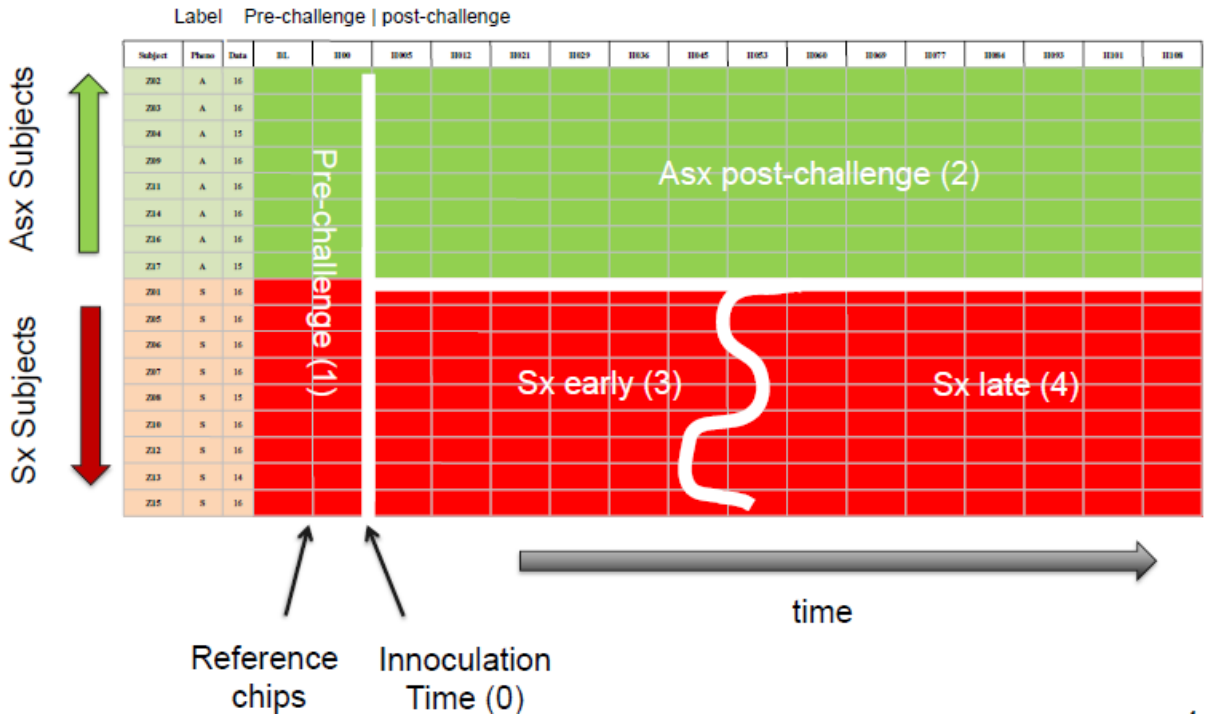
Dichotomous sampling regimes has motivated (Firouzi-H-R 2013, 2015):

- 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

- Firouzi, H and Rajaratnam, "Predictive correlation screening," AISTATS 2013
- Firouzi, H and Rajaratnam, "Two-stage sampling, prediction and adaptive regression via correlation screening (SPARCS)," arxiv vol. 1502:06189, 2015

Outline

Application: Screening for hub genes in flu trials



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

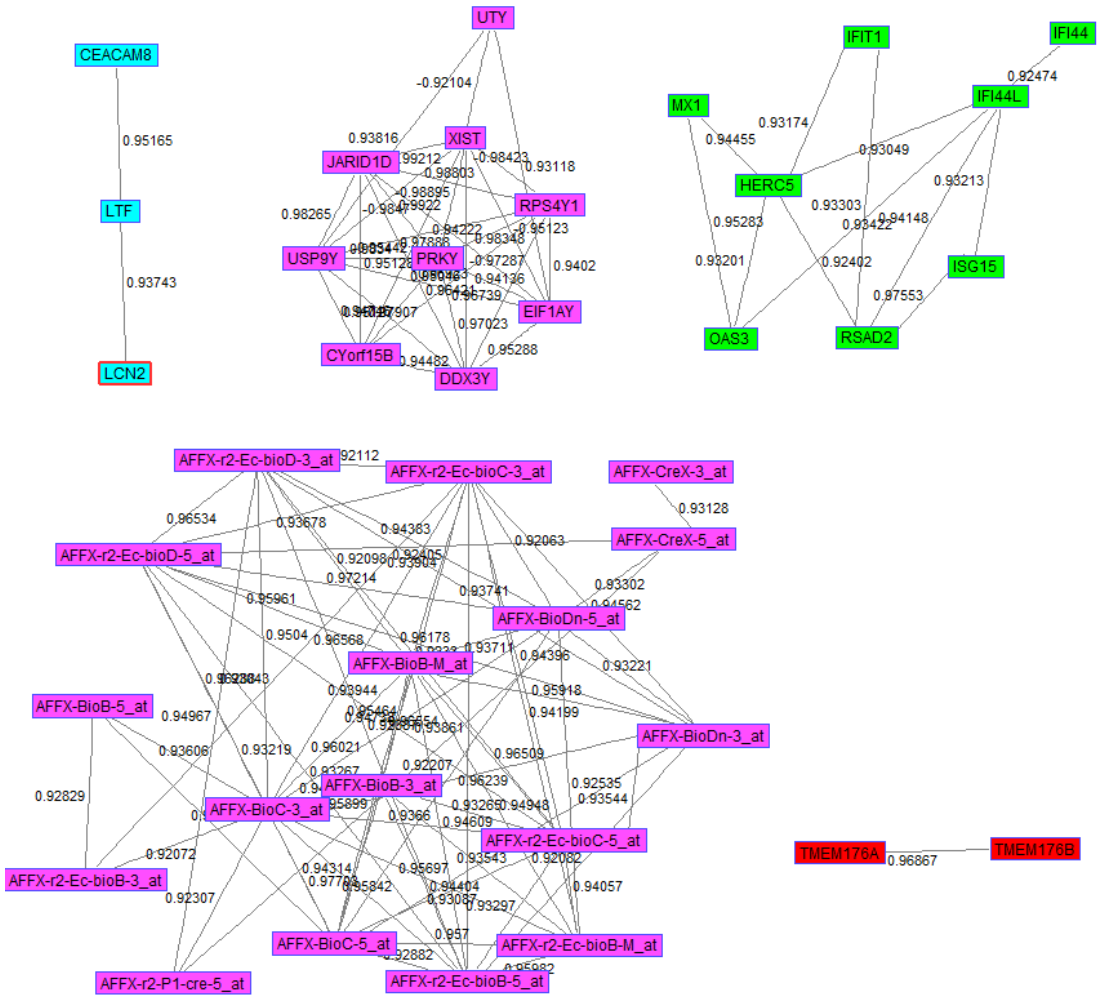
Samples fall into 3 categories

- 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: $\rho = 0.57$

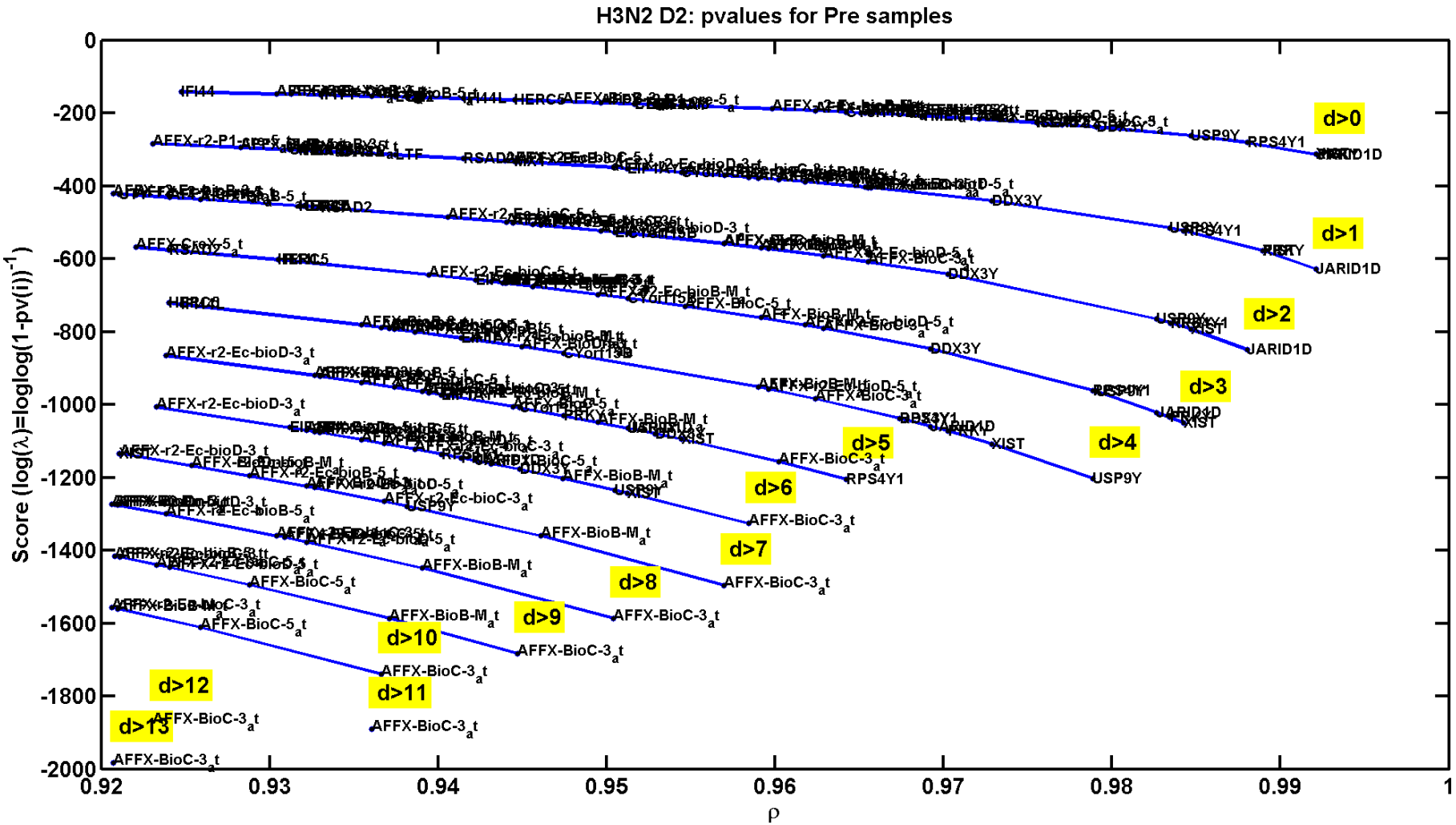
• H and Rajaratnam, "Large scale correlation mining for biomolecular network discovery," in Big data over networks, Cambridge 2015.

Susceptibility: Correlation screening the pre-inoc. samples

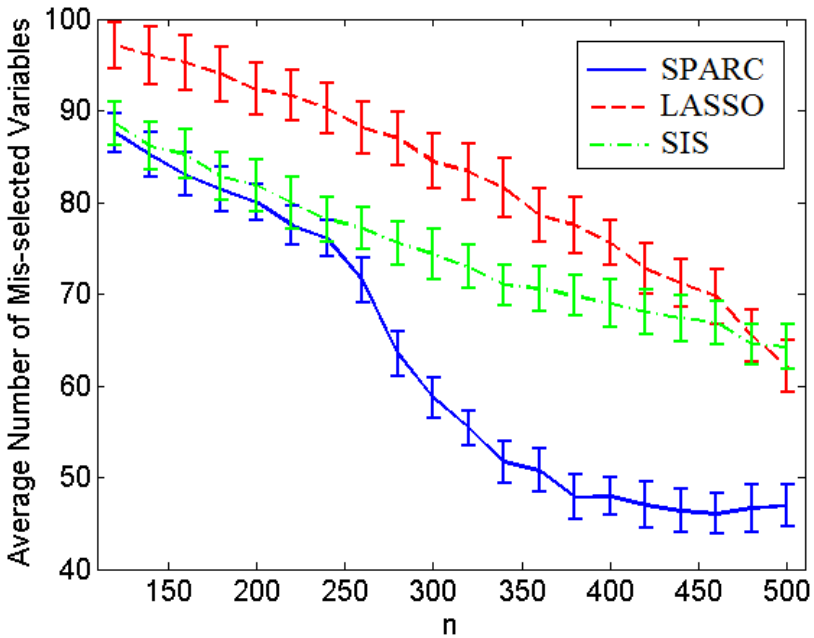
- Screen correlation at FWER 10^{-6}: 1658 genes, 8718 edges
- Screen partial correlation at FWER 10^{-6}: 39 genes, 111 edges



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



Multi-stage predictor design: SPARCS comparisons



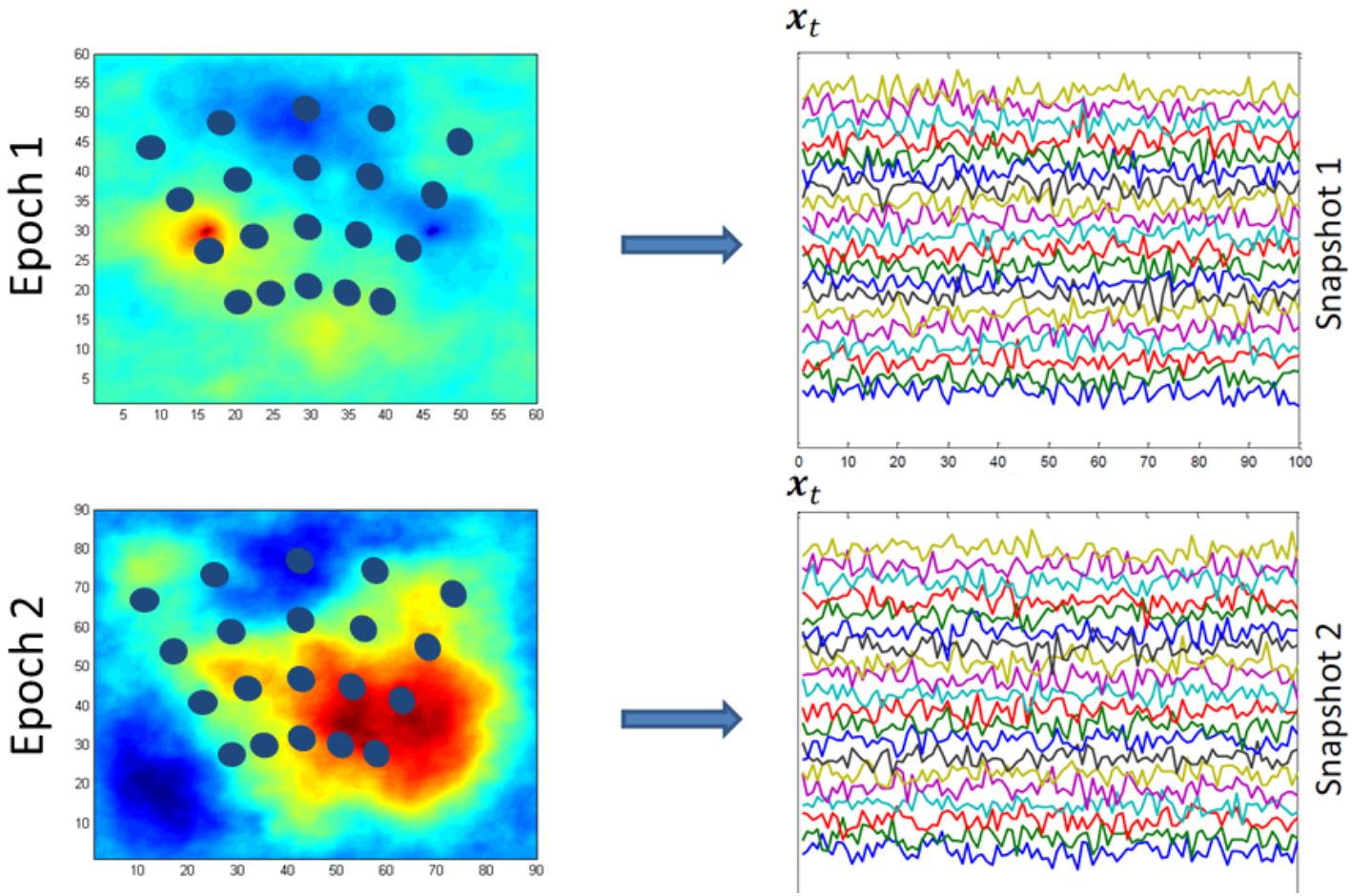
Symptom	RMSE: LASSO	RMSE: SIS	RMSE: SPARC
Runny Nose	0.7182	0.6896	**0.6559**
Stuffy Nose	0.9242	**0.7787**	0.8383
Sneezing	0.7453	0.6201	**0.6037**
Sore Throat	0.8235	0.7202	**0.5965**
Earache	**0.2896**	0.3226	0.3226
Malaise	1.0009	**0.7566**	0.9125
Cough	0.5879	0.7505	**0.5564**
Shortness of Breath	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," AISTATS 2013
- Firouzi, H and Rajaratnam, "Two-stage sampling, prediction and adaptive regression via correlation screening (SPARCS)," arxiv vol. 1502:06189, 2015.

Application: Kronecker PCA to spatio-temporal data

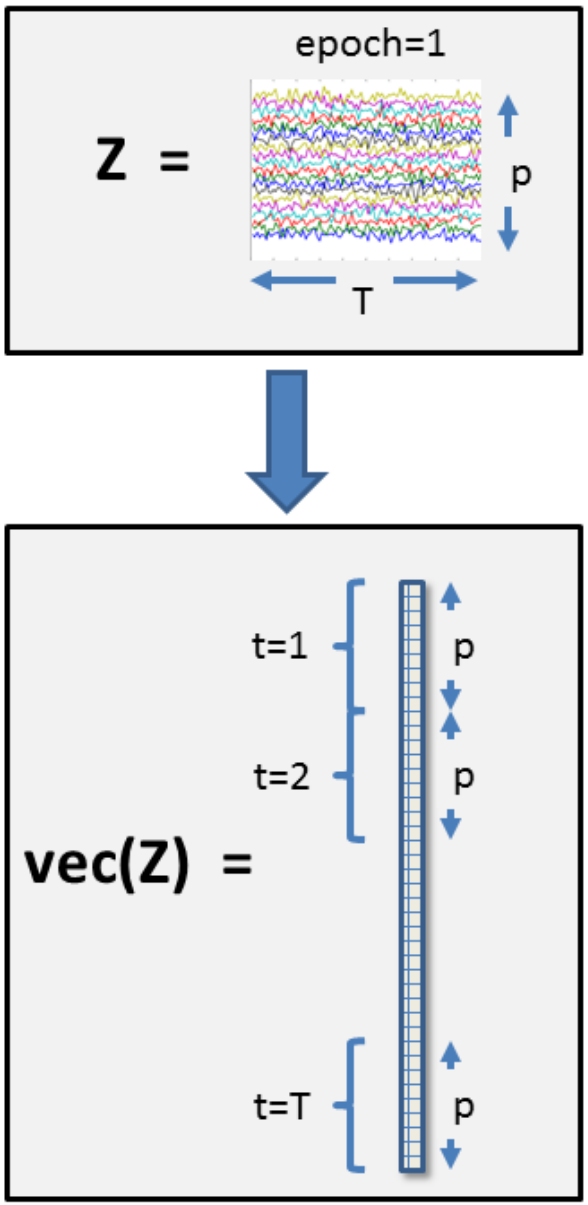


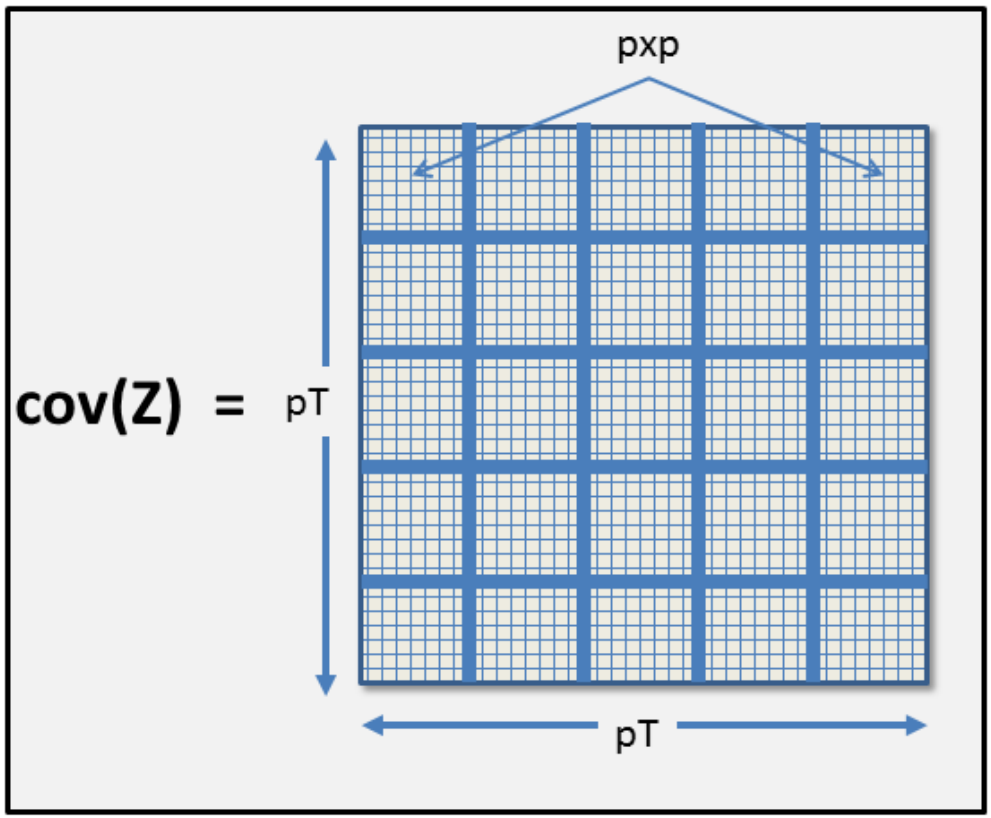
$T = 100,\ S = 20$

T. Tsiligkaridis and A.O. Hero, "Covariance Estimation in High Dimensions via Kronecker Product Expansions,"

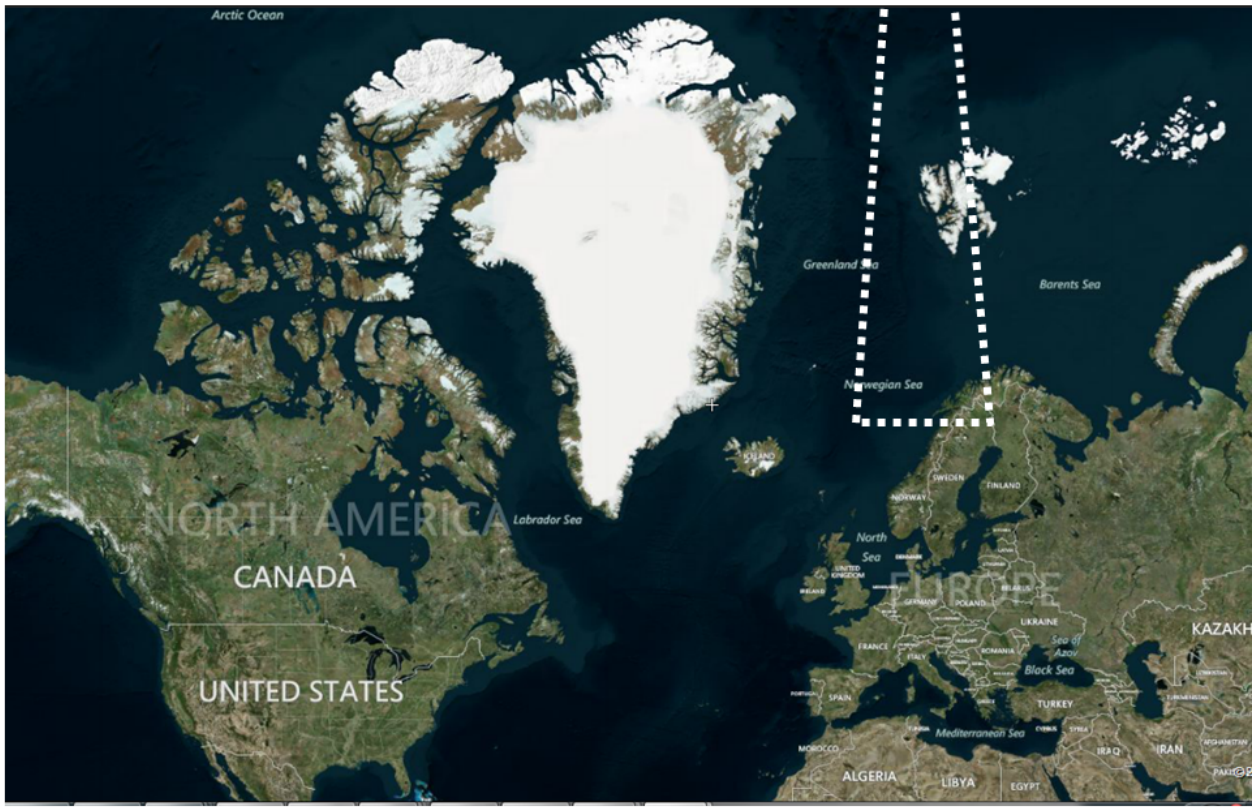
IEEE Trans on Signal Processing, Vol 61, No. 21, pp. 5347 - 5360, Nov 2013.

Spatio-temporal covariance has row dimension $M = pT$





Application: NCEP 10×10 over Arctic to Norwegian sea

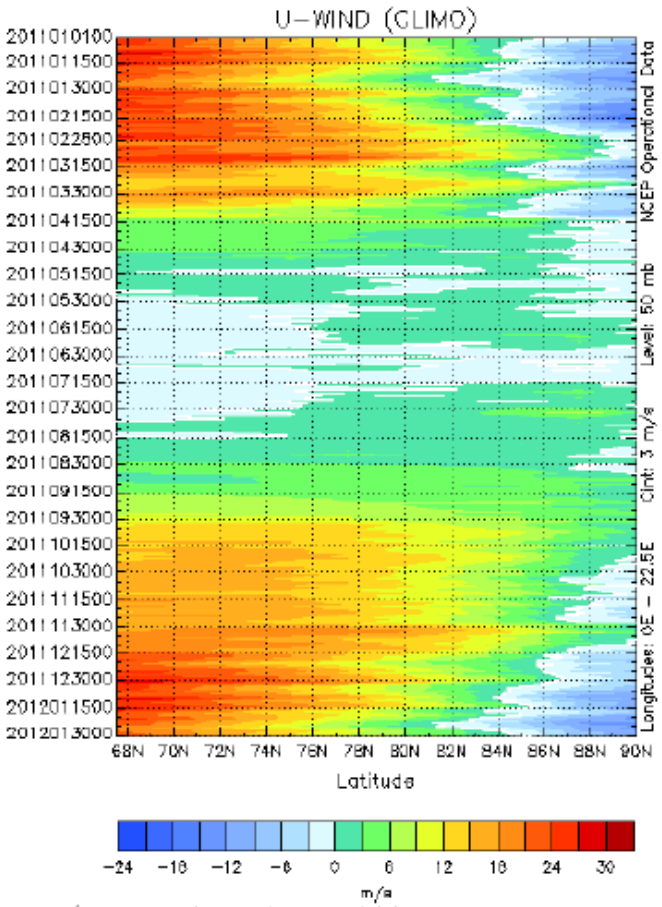


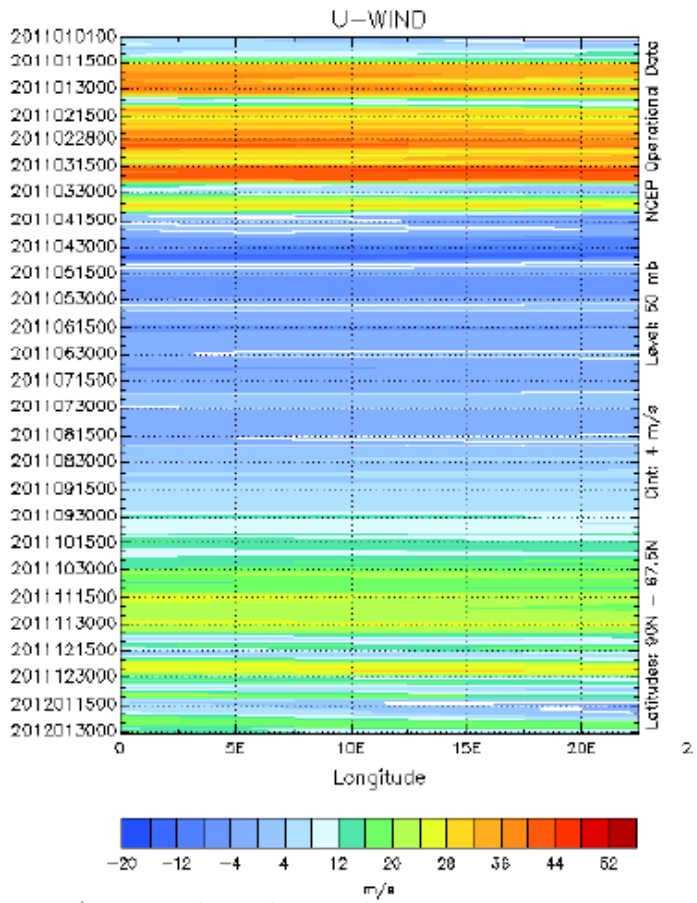
Region extending over latitudes 90-67.5 degrees N and longitudes 0-22.5 degrees E

- $S = 100$ spatial locations (10×10 spatial grid)
- $T = 8$ time points (2 day time window)
- $n = 224$ epochs (over period 2003-2007)
- Phase transition threshold: $\rho_c = 0.27$, 10% FA threshold is 0.33.

Application: NCEP 10×10 over Arctic to Norwegian sea

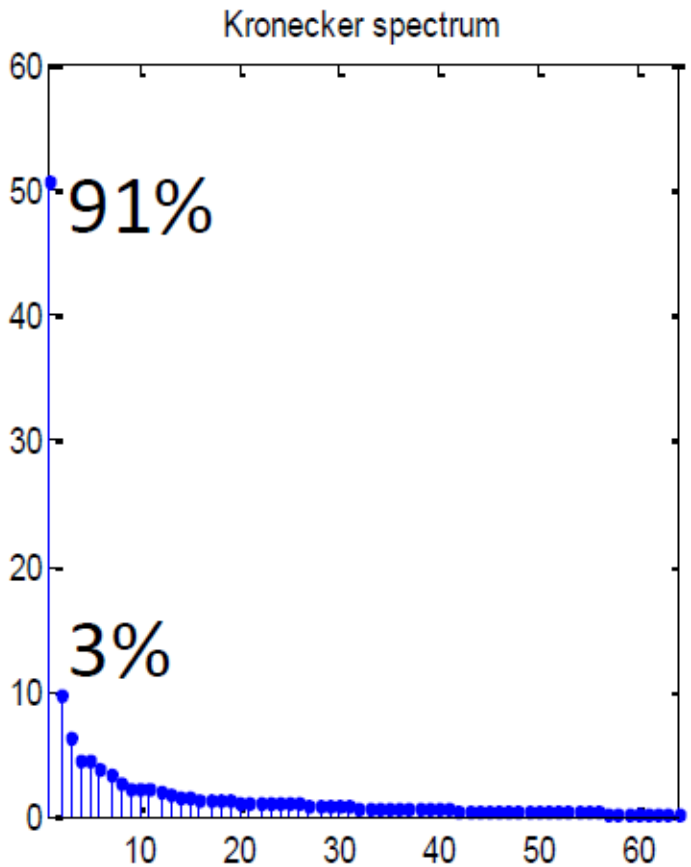
U component of windspeed

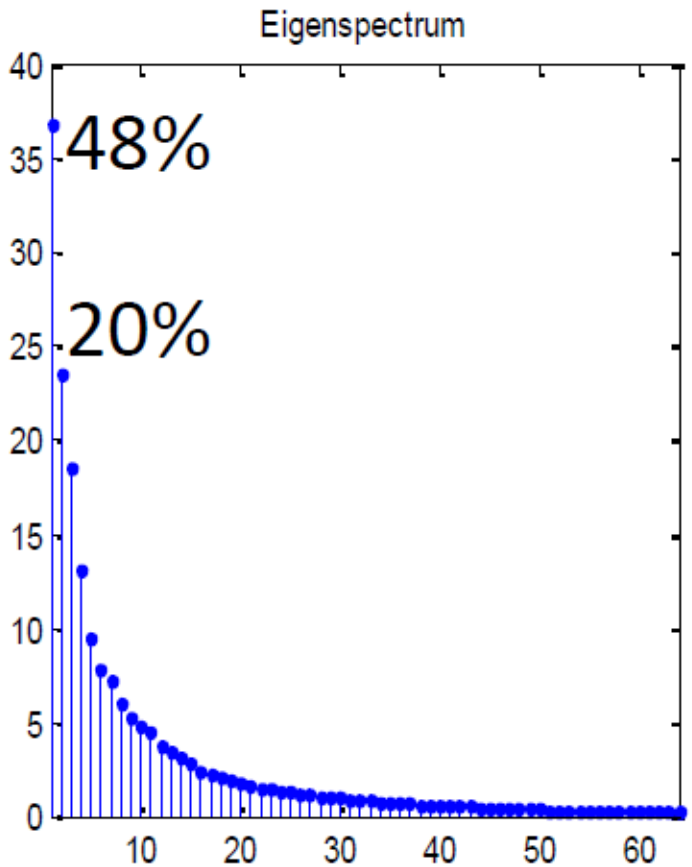




Lat (90-67.5) and long (0-22.5) is over 2.5 degree increments

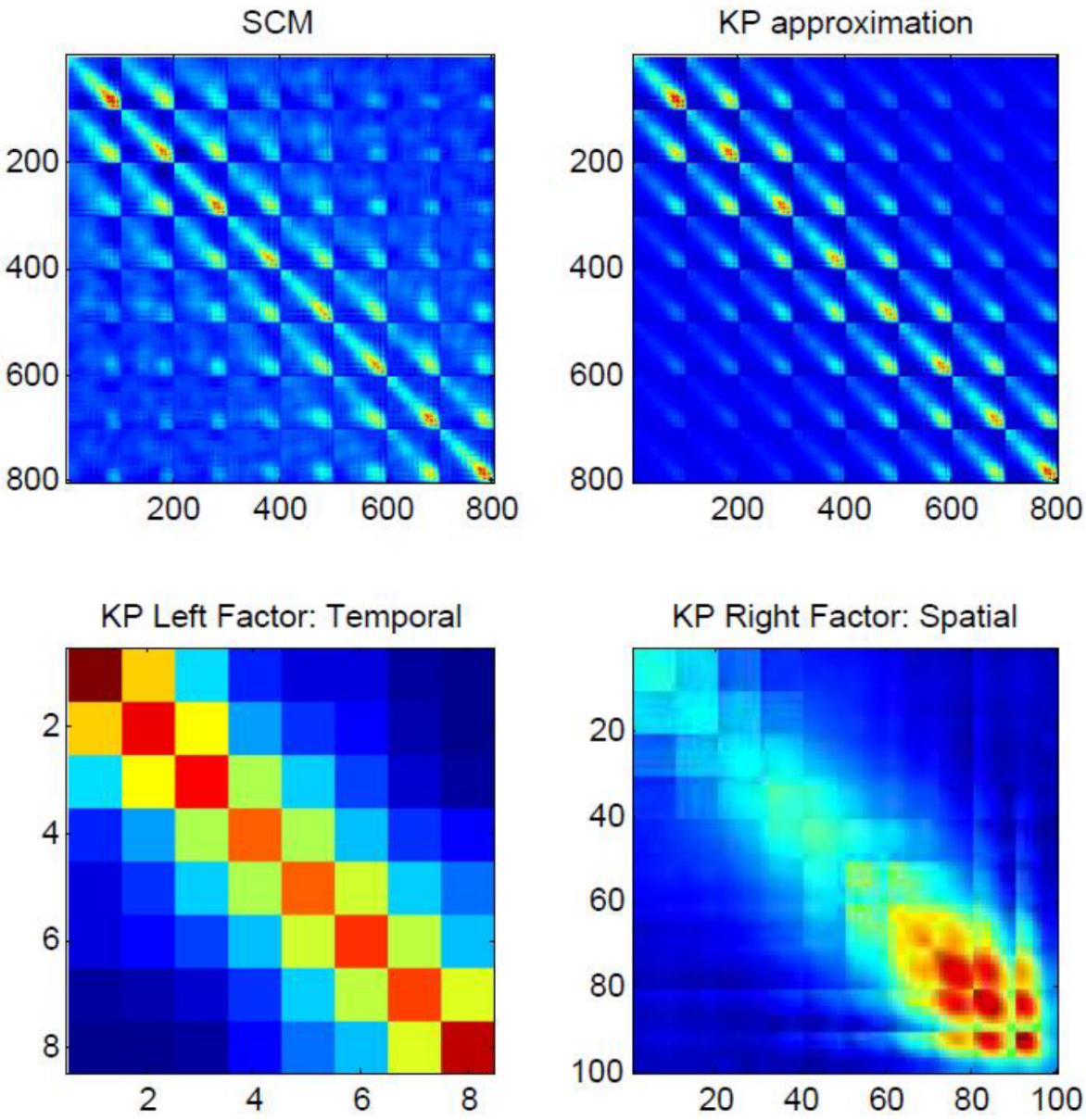
Application: NCEP 10×10 over Arctic to Norwegian sea





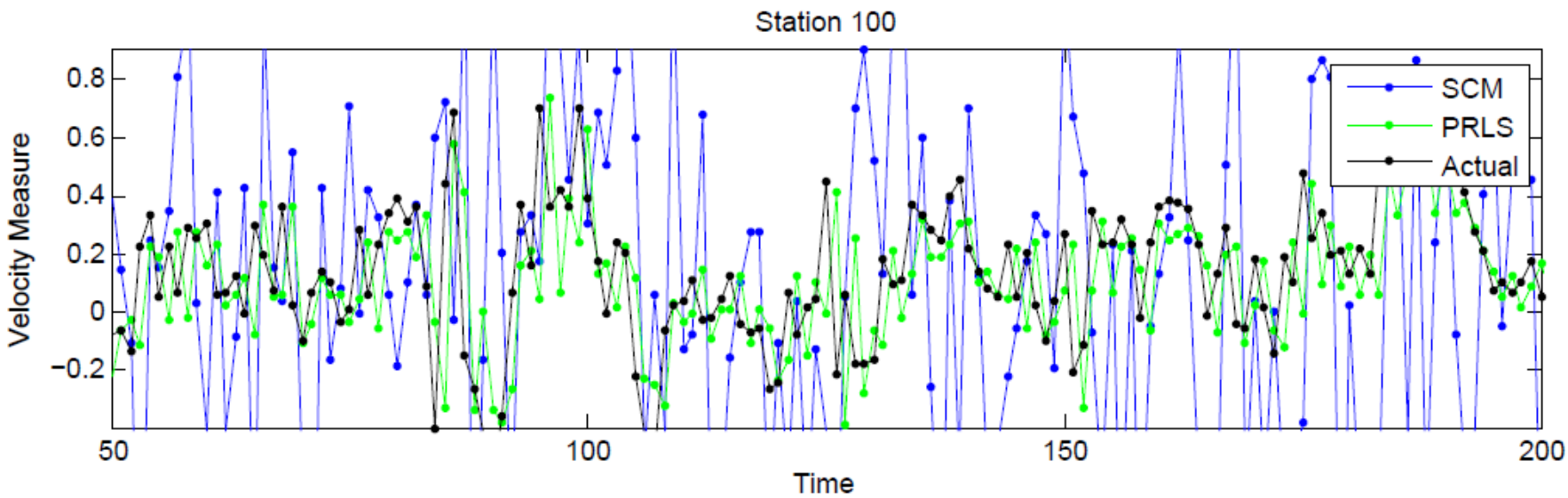
• Kronecker spectrum (left) significantly more concentrated than eigenspectrum (right)

Application: NCEP 10×10 Arctic to Norwegian sea



Application: NCEP 10×10 over Arctic to Norwegian sea

K-PCA in LS predictor yields higher prediction accuracy



- SCM covariance estimate $\hat{\mathbf{\Sigma}} = \mathbf{S}_n$
  - estimator is rank deficient
  - Prediction by min-norm (Moore-Penrose inverse) linear regression
- K-PCA covariance estimate $\hat{\mathbf{\Sigma}} = \sum_{i=1}^{r} \hat{\mathbf{A}}_i \bigotimes \hat{\mathbf{B}}_i$
  - estimator is full rank
  - Prediction by standard linear regression

Outline

Conclusions

- Correlation learning governed by phase transition thresholds that determine sample complexity
- Sample complexity depends on both model and learning task
  - Classical low dimensional: fixed p large n
  - Mixed high dimensional: large p and large n
  - Purely high dimensional: large p fixed n
- Applications:
  - Discovering immune hub genes by correlation screening
  - Training a Sx predictor using multi-stage data collection
  - Training a spatio-temporal predictor w/ Kronecker PCA

Conclusions

- Correlation learning governed by phase transition thresholds that determine sample complexity
- Sample complexity depends on both model and learning task
  - Classical low dimensional: fixed p large n
  - Mixed high dimensional: large p and large n
  - Purely high dimensional: large p fixed n
- Applications:
  - Discovering immune hub genes by correlation screening
  - Training a Sx predictor using multi-stage data collection
  - Training a spatio-temporal predictor w/ Kronecker PCA

Not covered here

- Learning spectral correlation (Firouzi and H, 2014)
- Screening for quickest change detection (Banerjee and H, 2015)
- Learning non-linear correlation (Todros and H, 2011, 2012)
- Meta learning of f-divergences (Moon 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.


C. Bazot, N. Dobigeon, J.-Y. Tourneret, A. K. Zaas, G. S. Ginsburg, and A. O. Hero, "Unsupervised bayesian linear unmixing of gene expression microarrays.," *BMC Bioinformatics*, vol. 14, pp. 99, 2013.


B. Chen, M. Chen, J. Paisley, A. Zaas, C. Woods, G. Ginsburg, A. O. H. III, J. Lucas, D. Dunson, and L. Carin, "Bayesian inference of the number of factors in gene-expression analysis: application to human virus challenge studies," *BMC BIoinformatics*, vol. 11, pp. 552, 2010.


B. Chen, M. Chen, J. Paisley, A. Zaas, C. Woods, G. Ginsburg, A. O. H. III, J. Lucas, D. Dunson, and L. Carin, "Detection of viruses via statistical gene expression analysis," *IEEE Trans. on Biomedical Engineering*, vol. 58, no. 3, pp. 468–479, 2011.


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.


Y. Huang, A. Zaas, A. Rao, N. Dobigeon, P. Woolf, T. Veldman, N. ien, M. McClain, J. Varkey, B. Nicholson, L. Carin, S. Kingsmore, C. Woods, G. Ginsburg, and A. Hero, III, "Temporal dynamics of host molecular responses differentiate symptomatic and asymptomatic influenza a infection," *PLoS Genet*, vol. 7, no. 8, pp. e1002234, 08 2011.


K. R. Moon, V. Delouille, and A. O. Hero III, "Meta learning of bounds on the bayes classifier error," *IEEE Signal Processing Workshop, Snowbird UT. Also on arXiv preprint arXiv:1504.07116*, 2015.


K. R. Moon and A. O. Hero III, "Ensemble estimation of multivariate f-divergence," *arXiv preprint arXiv:1404.6230*, 2014.


K. R. Moon and A. O. Hero III, "Multivariate f-divergence estimation with confidence," *Proc of Neural Information Systems (NIPS)*, 2014.


A. Puig, A. Wiesel, A. Zaas, C. Woods, and A. Ginsburg, G.and Hero III, "Order preserving factor analysis," *IEEE Trans. on Signal Processing*, p. To appear, June, 2011.


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 and A. Hero, "Covariance estimation in high dimensions via kronecker product expansions," *IEEE Trans. on Signal Processing (also available as arXiv:1302.2686)*, vol. 61, no. 21, pp. 5347–5360, 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.


C. W. Woods, M. T. McClain, M. Chen, A. K. Zaas, B. P. Nicholson, J. Varkey, T. Veldman, S. F. Kingsmore, Y. Huang, R. Lambkin-Williams, A. G. Gilbert, A. O. Hero, E. Ramsburg, J. E. Lucas, L. Carin, and G. S. Ginsburg, "A host transcriptional signature for presymptomatic detection of infection in humans exposed to influenza h1n1 or h3n2," *PloS one*, vol. 8, no. 1, pp. e52198, 2013.


A. Zaas, M. Chen, J. Varkey, T. Veldman, A. Hero, J. Lucas, Y. Huang, R. Turner, A. Gilbert, R. Lambkin-Williams, et al., "Gene Expression Signatures Diagnose Influenza and Other Symptomatic Respiratory Viral Infections in Humans," *Cell Host & Microbe*, vol. 6, no. 3, pp. 207–217, 2009.