Large scale correlation mining: fundamental performance limits

Alfred Hero

University of Michigan

July 29, 2015
1. Correlation mining

2. High dimensional analysis

3. Sample complexity

4. Two-stage Sampling, Prediction and Adaptive Regression via Correlation Screening (SPARCS)

5. Application to predicting health and disease

6. Conclusions
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

1. Correlation mining
2. High dimensional analysis
3. Sample complexity
4. Two-stage Sampling, Prediction and Adaptive Regression via Correlation Screening (SPARCS)
5. Application to predicting health and disease
6. Conclusions
Correlation mining and network discovery

- Molecular pathways
  - PLoS '11, Science '13, BMC Bioinfo '12...
- Spammer communities
  - CEAS '10, DNIK '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, ICSE '09...

**Value of Info**
- TSP '11, Sensors '11...

**Correlation mining**

**Feature representation**

**Data acquisition and sampling**

**Error control**
- 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**
- **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)
  - 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

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

Source: Young and Karr, Significance, Sept. 2011
Related work: estimation, selection, testing, screening

- Regularized $l_2$ or $l_F$ covariance estimation

- Gaussian graphical model selection
  - Bayesian estimation: Rajaratnam-Massam-Carvalho (2008)

- Independence testing
  - Sphericity test for multivariate Gaussian: Wilks (1935)

- 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

1. Correlation mining
2. High dimensional analysis
3. Sample complexity
4. Two-stage Sampling, Prediction and Adaptive Regression via Correlation Screening (SPARCS)
5. Application to predicting health and disease
6. Conclusions
Correlation matrix and its support set

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

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

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

$$
\Gamma = \text{diag}(\Sigma)^{-1/2} \Sigma \text{ diag}(\Sigma)^{-1/2}
$$

- Adjacency matrix: $A_o = h_0(\Gamma)$,

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

- Connectivity support set: $S_o = S_o^{(1)} = I(\text{sum}(A_o) > 1)$
- Hub degree $\geq \delta$ support set: $S_o^{(\delta)} = I(\text{sum}(A_o) > \delta)$
Empirical estimation of correlation and support set

- $p \times p$ sample covariance matrix
  \[
  \hat{\Sigma} = X(\mathbf{I} - \frac{1}{n} \mathbf{1} \mathbf{1}^T)X^T \frac{1}{n - 1}
  \]

- $p \times p$ sample correlation matrix
  \[
  R = \text{diag}(\hat{\Sigma})^{-1/2} \hat{\Sigma} \text{diag}(\hat{\Sigma})^{-1/2}
  \]

- Sample estimator of adjacency matrix at correlation level $\rho \in [0, 1]$
  \[
  \hat{A}_o(\rho) = h_\rho(R)
  \]

- Sample estimator of connectivity support $S_o(\rho)$ at level $\rho \in [0, 1]$
  \[
  \hat{S}_o(\rho) = I(\text{sum}(\hat{A}_o(\rho)) > \delta)
  \]
Estimation vs support recovery vs screening for dependency

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

\[ P_0(N_\rho > 0) \]

\[ N_\rho = \text{card}\{\hat{S}_o(\rho)\} \] is number of discoveries above threshold \( \rho \).

- **Support recovery**: support misclassification error

\[ P_\Sigma(\hat{S}_o(\rho) \Delta S_o \neq \phi) \]

- **Covariance estimation**: Frobenius norm error

\[ \| \Sigma - \hat{\Sigma} \|_F \]

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

<table>
<thead>
<tr>
<th>Asymptotic framework</th>
<th>Terminology</th>
<th>Sample size</th>
<th>Dimension</th>
<th>Application setting</th>
<th>References</th>
</tr>
</thead>
<tbody>
<tr>
<td>Classical (or sample increasing)</td>
<td>small dimensional</td>
<td>$\rightarrow \infty$</td>
<td>fixed</td>
<td>“small data”</td>
<td>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]</td>
</tr>
<tr>
<td>Mixed asymptotics</td>
<td>high dimensional</td>
<td>$\rightarrow \infty$</td>
<td>$\rightarrow \infty$</td>
<td>“medium sized” data (mega or giga scales)</td>
<td>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]</td>
</tr>
<tr>
<td></td>
<td>very high dimensional</td>
<td>$\rightarrow \infty$</td>
<td>$\rightarrow \infty$</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>ultra high dimensional</td>
<td>$\rightarrow \infty$</td>
<td>$\rightarrow \infty$</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Purely high dimensional</td>
<td>purely high dimensional</td>
<td>fixed</td>
<td>$\rightarrow \infty$</td>
<td>“Big Data” (tera, peta and exascales)</td>
<td>Hero and Rajaratnam [35]</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Hero and Rajaratnam [36]</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Firouzi, Hero and Rajaratnam [25]</td>
</tr>
</tbody>
</table>

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

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


- Impossible to reliably detect small correlations with finite $n$
- Possible to reliably detect large correlations even when $n \ll p$
- Critical threshold $\rho_c$ 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 $\Sigma$ if block sparse
Purely high D convergence theorem (H-R 2012)

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

**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

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

$$\lambda_{\delta,\rho,n} = p \left( \frac{p - 1}{\delta} \right) (P_0(\rho, n))^\delta J(\Sigma)$$

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

---

1 Generalized to local screening in (Firouzi-H 2013) and complex valued screening in (Firouzi-W-H 2014)
Critical threshold $\rho_c$ as function of $n$ (H-Rajaratnam 2012)
Critical phase transition threshold in $n$ and $p$ ($\delta = 1$)

Outline

1. Correlation mining
2. High dimensional analysis
3. Sample complexity
4. Two-stage Sampling, Prediction and Adaptive Regression via Correlation Screening (SPARCS)
5. Application to predicting health and disease
6. Conclusions
Sample complexity regimes for different tasks

<table>
<thead>
<tr>
<th>Task</th>
<th>Screening</th>
<th>Detection</th>
<th>Support detection</th>
<th>Param. estimation</th>
<th>Perform. estimation</th>
</tr>
</thead>
<tbody>
<tr>
<td>Risk</td>
<td>$P(N_e &gt; 0)$</td>
<td>$P(N_e &gt; 0)$</td>
<td>$P(\text{card}{S \Delta \hat{S}} = \phi)$</td>
<td>$E[|\Omega - \hat{\Omega}|_F^2]$</td>
<td>$\int E[(f_\Omega(x) - \hat{f}(x))^2]dx$</td>
</tr>
<tr>
<td>Bound</td>
<td>$1 - e^{-\kappa n}$</td>
<td>$pe^{-n\alpha}$</td>
<td>$2pe^{-n\alpha}$</td>
<td>$\frac{p \log p}{n \alpha}$</td>
<td>$n^{-2/(1+p)\alpha}$</td>
</tr>
<tr>
<td>Regimes</td>
<td>$\frac{\log p}{n} \to \infty$</td>
<td>$\frac{\log p}{n} \to \alpha$</td>
<td>$\frac{p}{n} \to \alpha$</td>
<td>$\frac{p \log p}{n} \to \alpha$</td>
<td>$\frac{p}{\log n} \to \alpha$</td>
</tr>
<tr>
<td>Threshold</td>
<td>$\rho_c \to 1$</td>
<td>$\rho_c \to \rho^*$</td>
<td>$\rho_c \to 0$</td>
<td>$\rho_c \to 0$</td>
<td>$\rho_c \to 0$</td>
</tr>
</tbody>
</table>


- 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

- 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
  ⇒ match the mining task to the available sample size.
- Multistage correlation mining for budget limited applications
  ⇒ Screen small exploratory sample prior to big collection
Outline

1. Correlation mining
2. High dimensional analysis
3. Sample complexity
4. Two-stage Sampling, Prediction and Adaptive Regression via Correlation Screening (SPARCS)
5. Application to predicting health and disease
6. Conclusions
Sampling, Prediction and Adaptive Regression via Correlation Screening

Experiment: Stage 1
- \( p \) probes
- \( q \) responses
- \( n \) replicates

\[ \gamma^q \]

Experiment: Stage 2
- \( k \) probes
- \( q \) responses
- \( t-n \) replicates

\[ \chi^k \]

Predictive Correlation Screening (\( \delta=1 \))

Pooled OLS predictor:

\[ \text{argmin}_A \sum_{\exp1,\exp2} |Y^q - AX^k|^2 \]

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 \( \pi_0 \).

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

- Total budget for two-stage experiment: $\mu$.
- 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*

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

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

1. Correlation mining
2. High dimensional analysis
3. Sample complexity
4. Two-stage Sampling, Prediction and Adaptive Regression via Correlation Screening (SPARCS)
5. Application to predicting health and disease
6. Conclusions
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
Huang et al, PLoS Genetics, 2011
Bazot et al, BMC Bioinformatics, 2013
Zaas et al, Science Translation Medicine, 2014
Critical threshold $\rho_c$ 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
Support recovery (simu)  

- Firouzi, H and Rajaratnam, "Predictive correlation screening: Application to two-stage predictor design in high dimension," AISTATS 2013

Outline

1. Correlation mining
2. High dimensional analysis
3. Sample complexity
4. Two-stage Sampling, Prediction and Adaptive Regression via Correlation Screening (SPARCS)
5. Application to predicting health and disease
6. Conclusions
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:

  \[ n \ \text{fixed} \quad \text{while} \quad p \to \infty \]

- 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)


available as Arxiv preprint arXiv:1109.6846.


