Sparse correlation screening in high dimension

Alfred Hero

University of Michigan - Ann Arbor

Workshop on algorithms on Modern Massive Data Sets
June 17, 2010

Acknowledgements

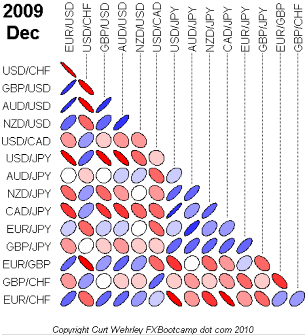
- Bala Rajaratnam (Stanford)
- Isaac Newton Institute
- DIGITEO, Paris France
- NSF: ITR CCR-032557

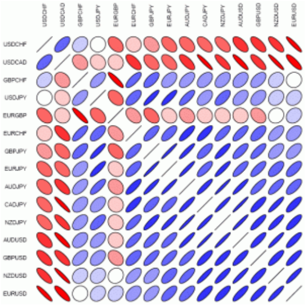
Outline

Correlation analysis of financial time series

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

p-variate correlation analysis of financial data





fxbootcamp.com

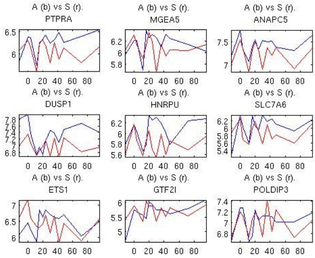
Sample covariance matrix:

$$\hat{\Sigma} = \frac{1}{n-1}\sum_{i=1}^{n}(\mathbf{X}_i - \hat{\mu})(\mathbf{X}_i - \hat{\mu})^T$$

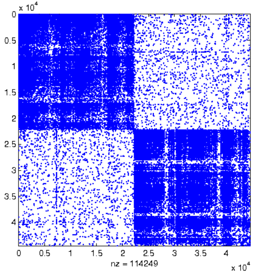
Sample correlation matrix:

$$\mathbf{R} = \mathbf{D}_{\hat{\Sigma}}^{-1/2}\hat{\Sigma}\mathbf{D}_{\hat{\Sigma}}^{-1/2}$$

Correlation analysis of gene expression arrays



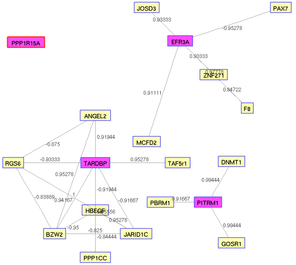
Gene expression profiles

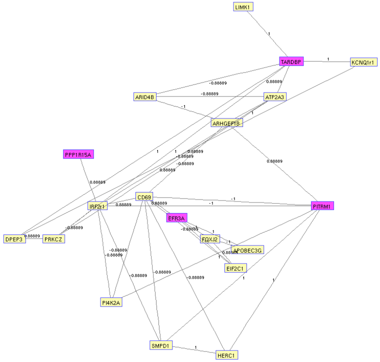


Correlation matrix **R**

Correlation screening and hub discovery

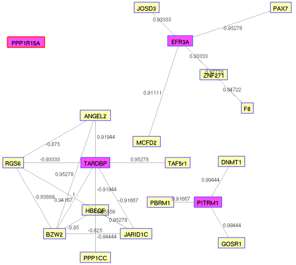
Blue condition

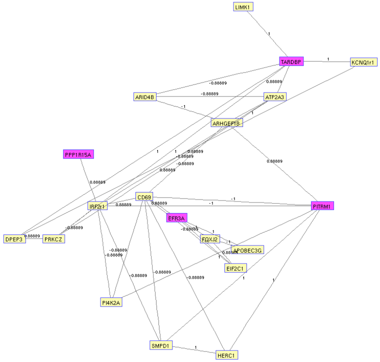




Correlation screening and hub discovery

Blue condition

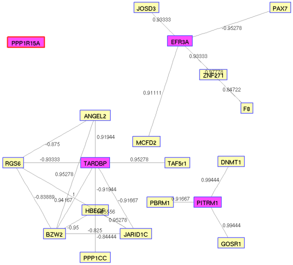


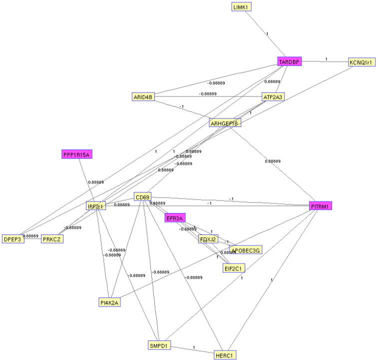


- Correlation screening finds hubs of high sample correlation

Correlation screening and hub discovery

Blue condition

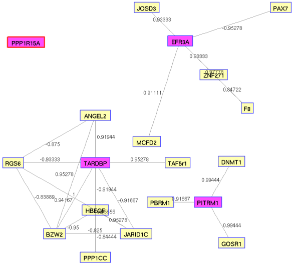


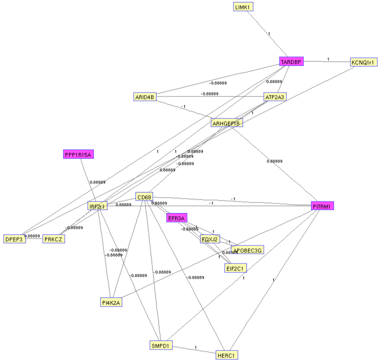


- Correlation screening finds hubs of high sample correlation
- Persistent correlation screening finds hubs surviving both treatments

Correlation screening and hub discovery

Blue condition





- Correlation screening finds hubs of high sample correlation
- Persistent correlation screening finds hubs surviving both treatments
- Edges shown are survivors after leave-one-out cross-validation

How much confidence can we have in such discoveries?

Confidence mitigated by

- Lack of principles for selecting correlation threshold

How much confidence can we have in such discoveries?

Confidence mitigated by

- Lack of principles for selecting correlation threshold
- Many (p) variables but few (n) observations

How much confidence can we have in such discoveries?

Confidence mitigated by

- Lack of principles for selecting correlation threshold
- Many (p) variables but few (n) observations
  - Affymetrix gene chip has $22,000$ probes (variables)
  - ....and has $\binom{22{,}000}{2} = 241,989,000$ sample correlations

How much confidence can we have in such discoveries?

Confidence mitigated by

- Lack of principles for selecting correlation threshold
- Many (p) variables but few (n) observations
  - Affymetrix gene chip has $22,000$ probes (variables)
  - ....and has $\binom{22,000}{2} = 241,989,000$ sample correlations
  - Often number of samples per treatment is less than 10

How much confidence can we have in such discoveries?

Confidence mitigated by

- Lack of principles for selecting correlation threshold
- Many (p) variables but few (n) observations
  - Affymetrix gene chip has $22,000$ probes (variables)
  - ....and has $\binom{22,000}{2} = 241,989,000$ sample correlations
  - Often number of samples per treatment is less than 10
- Cross validation cannot be relied upon in these situations

How much confidence can we have in such discoveries?

Confidence mitigated by

- Lack of principles for selecting correlation threshold
- Many (p) variables but few (n) observations
    - Affymetrix gene chip has $22,000$ probes (variables)
    - ....and has $\binom{22,000}{2} = 241,989,000$ sample correlations
    - Often number of samples per treatment is less than 10
- Cross validation cannot be relied upon in these situations

Objective: establish asymptotic (large p) theory.

Previous work

- Regularized l_2 or $l_{\mathcal{F}}$ covariance estimation
  - Shrinkage towards identity: Ledoit-Wolf (2005), Chen-Weisel-Eldar-H (2010)
  - Shrinkage towards banded: Bickel-Levina (2008)
  - Shrinkage towards sparse eigenvector: Johnstone-Lu (2007)

Previous work

- Regularized l_2 or $l_{\mathcal{F}}$ covariance estimation
  - Shrinkage towards identity: Ledoit-Wolf (2005), Chen-Weisel-Eldar-H (2010)
  - Shrinkage towards banded: Bickel-Levina (2008)
  - Shrinkage towards sparse eigenvector: Johnstone-Lu (2007)
- Gaussian graphical model selection
  - l_1 regularized GGM: Meinshausen-Bühlmann (2006), Wiesel-Eldar-H (2010).
  - Bayesian estimation: Rajaratnam-Massam-Carvalho (2008)

Previous work

- Regularized l_2 or $l_{\mathcal{F}}$ covariance estimation
  - Shrinkage towards identity: Ledoit-Wolf (2005), Chen-Weisel-Eldar-H (2010)
  - Shrinkage towards banded: Bickel-Levina (2008)
  - Shrinkage towards sparse eigenvector: Johnstone-Lu (2007)
- Gaussian graphical model selection
  - l_1 regularized GGM: Meinshausen-Bühlmann (2006), Wiesel-Eldar-H (2010).
  - Bayesian estimation: Rajaratnam-Massam-Carvalho (2008)
- Independence testing
  - Sphericity test for multivariate Gaussian: Wilks (1935)
  - Maximal correlation test: Moran (1980)
  - Ranked correlation test: Eagleson (1983)

Previous work

- Regularized l_2 or $l_{\mathcal{F}}$ covariance estimation
  - Shrinkage towards identity: Ledoit-Wolf (2005), Chen-Weisel-Eldar-H (2010)
  - Shrinkage towards banded: Bickel-Levina (2008)
  - Shrinkage towards sparse eigenvector: Johnstone-Lu (2007)
- Gaussian graphical model selection
  - l_1 regularized GGM: Meinshausen-Bühlmann (2006), Wiesel-Eldar-H (2010).
  - Bayesian estimation: Rajaratnam-Massam-Carvalho (2008)
- Independence testing
  - Sphericity test for multivariate Gaussian: Wilks (1935)
  - Maximal correlation test: Moran (1980)
  - Ranked correlation test: Eagleson (1983)

New framework: screening for highly correlated variables

Previous work

- Regularized l_2 or $l_{\mathcal{F}}$ covariance estimation
  - Shrinkage towards identity: Ledoit-Wolf (2005), Chen-Weisel-Eldar-H (2010)
  - Shrinkage towards banded: Bickel-Levina (2008)
  - Shrinkage towards sparse eigenvector: Johnstone-Lu (2007)
- Gaussian graphical model selection
  - l_1 regularized GGM: Meinshausen-Bühlmann (2006), Wiesel-Eldar-H (2010).
  - Bayesian estimation: Rajaratnam-Massam-Carvalho (2008)
- Independence testing
  - Sphericity test for multivariate Gaussian: Wilks (1935)
  - Maximal correlation test: Moran (1980)
  - Ranked correlation test: Eagleson (1983)

New framework: screening for highly correlated variables
No particular distribution or sparsity patterns imposed

Previous work

- Regularized l_2 or $l_{\mathcal{F}}$ covariance estimation
  - Shrinkage towards identity: Ledoit-Wolf (2005), Chen-Weisel-Eldar-H (2010)
  - Shrinkage towards banded: Bickel-Levina (2008)
  - Shrinkage towards sparse eigenvector: Johnstone-Lu (2007)
- Gaussian graphical model selection
  - l_1 regularized GGM: Meinshausen-Bühlmann (2006), Wiesel-Eldar-H (2010).
  - Bayesian estimation: Rajaratnam-Massam-Carvalho (2008)
- Independence testing
  - Sphericity test for multivariate Gaussian: Wilks (1935)
  - Maximal correlation test: Moran (1980)
  - Ranked correlation test: Eagleson (1983)

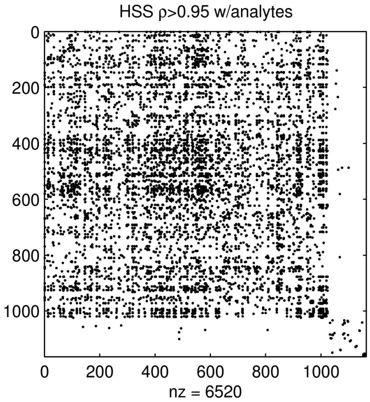
New framework: screening for highly correlated variables
No particular distribution or sparsity patterns imposed
Scalable: computational complexity can be as low as $O(logp)$

Correlation screening = screening rows (variables) of $\mathbf{R}$

- For $\mathbf{r}_{ij} = (\mathbf{R})_{ij}$ let ρ be a user-defined threshold in $[0, 1]$
- Variable i passes correlation screen if: $\max_{j \neq i} |\mathbf{r}_{ij}| \geq \rho$

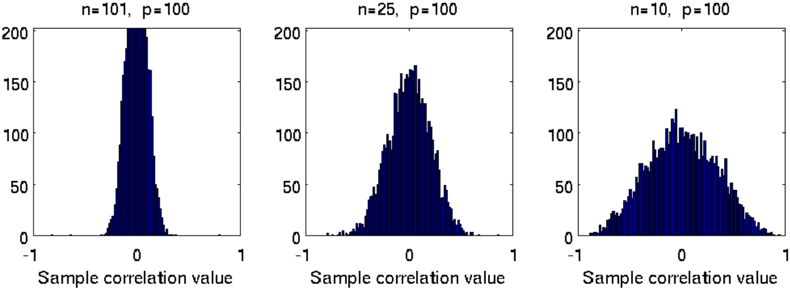
Correlation screening = screening rows (variables) of $\mathbf{R}$

- For $\mathbf{r}_{ij} = (\mathbf{R})_{ij}$ let ρ be a user-defined threshold in $[0,1]$
- Variable i passes correlation screen if: $\max_{j \neq i} |\mathbf{r}_{ij}| \geq \rho$
- *Discovered* variables have high correlation with some other variable



Phase transitions in correlation screening

- Number of discoveries exhibit phase transition phenomenon
- This phenomenon gets worse as p/n increases.



Overview of mathematical results

Two types of results for auto-correlation and persistent correlation screening

- Characterize large p phase transition and its threshold.
- Poisson asymptotics for predicting false positive rates.

Overview of mathematical results

Two types of results for auto-correlation and persistent correlation screening

- Characterize large p phase transition and its threshold.
- Poisson asymptotics for predicting false positive rates.

Main ingredients in our analysis

- Z-score representation: $\mathbf{R} = \mathbb{U}^T\mathbb{U}$

$$\mathbb{U} = [\mathbf{U}_1, \ldots, \mathbf{U}_p], \quad \mathbf{U}_i \in S_{n-2} \subset \mathbb{R}^{n-1}$$

- Geometric probability on unit-sphere S_{n-2}
- Exchangeable process theory for dependent variables

Sample correlation and Z-score distances

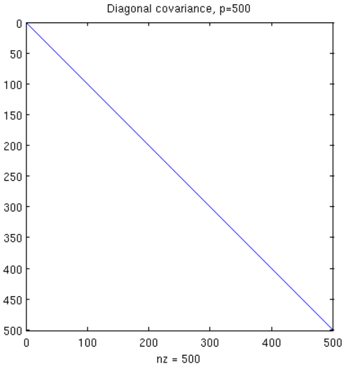
- Sample correlation between $\mathbf{X}_i$ and $\mathbf{X}_j$ is equal to Z-score inner product

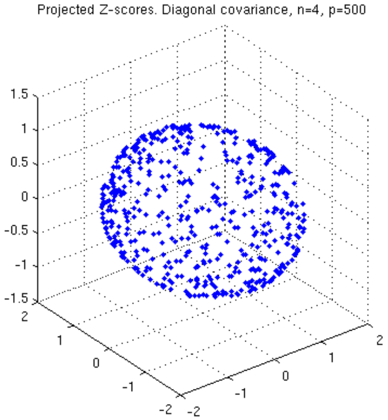
$$\mathbf{r}_{ij} = \mathbf{U}_i^T \mathbf{U}_j$$

Sample correlation and Z-score distances

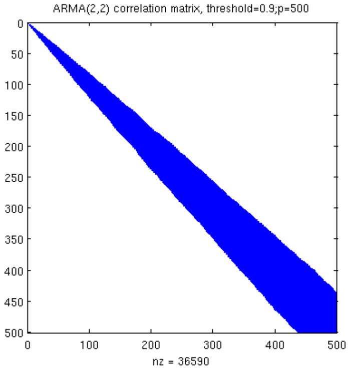
- Sample correlation between $\mathbf{X}_i$ and $\mathbf{X}_j$ is equal to Z-score inner product
$$\mathbf{r}_{ij} = \mathbf{U}_i^T \mathbf{U}_j$$
- Relate to Euclidean distance between $\mathbf{U}_i$ and $\mathbf{U}_j$
$$\|\mathbf{U}_i - \mathbf{U}_j\| = \sqrt{2(1 - \mathbf{r}_{ij})}$$

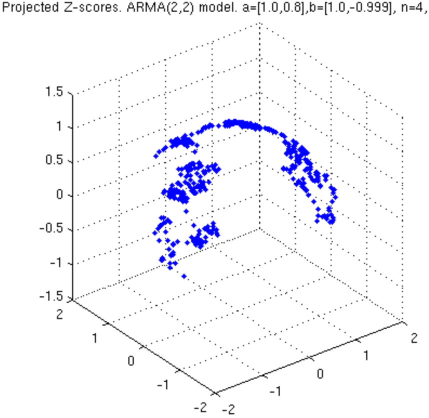
Example: Z-scores for diagonal Gaussian





Example : Z-scores for ARMA(2,2) Gaussian





Outline

Mathematical analysis

Define N the number of discoveries:

$$N = \sum_{i=1}^{p} \phi_i$$

Where $\phi = [\phi_1, \ldots, \phi_p]$ is "discovery" indicator sequence:

$$\phi_i = \begin{cases} 1, & \max_{j \neq i} |\mathbf{r}_{ij}| \geq \rho \\ 0, & o.w. \end{cases}$$

Mathematical analysis

Define N the number of discoveries:

$$N = \sum_{i=1}^{p} \phi_i$$

Where $\phi = [\phi_1, \ldots, \phi_p]$ is "discovery" indicator sequence:

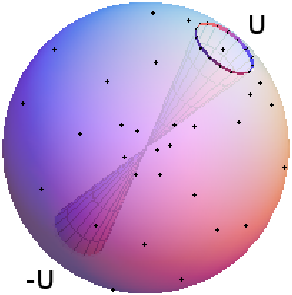
$$\phi_i = \begin{cases} 1, & \max_{j \neq i} |\mathbf{r}_{ij}| \geq \rho \\ 0, & o.w. \end{cases}$$

Objective: Find mathematical expressions for $E[N]$ as a function of p, n, ρ.

Mathematical analysis

Conditional expectation of ϕ_i has representation

$$E[\phi_i|\mathbf{U}_i] = P(\cup_{j\neq i}\mathbf{U}_j \in C_{\rho,\mathbf{U}_i} \cup C_{\rho,-\mathbf{U}_i}|\mathbf{U}_i)$$



Mathematical analysis

Given $\mathbf{U}_i$ define the binary sequence $\mathbf{b} = [b_1, \ldots, b_{p-1}]$

$$b_i = \begin{cases} 1, & \mathbf{U}_j \in C_{\rho,\mathbf{U}_i} \cup C_{\rho,-\mathbf{U}_i} \\ 0, & o.w. \end{cases}$$

Then, have equivalent representation

$$E[\phi_i|\mathbf{U}_i] = P(\sum_{i=1}^{p-1} b_i > 0|\mathbf{U}_i)$$

Mathematical analysis

Given $\mathbf{U}_i$ define the binary sequence $\mathbf{b} = [b_1, \ldots, b_{p-1}]$

$$b_i = \begin{cases} 1, & \mathbf{U}_j \in C_{\rho,\mathbf{U}_i} \cup C_{\rho,-\mathbf{U}_i} \\ 0, & o.w. \end{cases}$$

Then, have equivalent representation

$$E[\phi_i|\mathbf{U}_i] = P(\sum_{i=1}^{p-1} b_i > 0|\mathbf{U}_i)$$

Classical result [Thm. 4.5.4]{TW Anderson, 2003}:

Lemma

Let $\mathbf{X}$ be a p-variate elliptical vector with diagonal dispersion matrix $\mathbf{\Sigma}$. The Z-scores $\{\mathbf{U}_i\}_{i=1}^{p}$ are i.i.d. random vectors uniformly distributed on S_{n-2}.

Mathematical analysis

Given $\mathbf{U}_i$ define the binary sequence $\mathbf{b} = [b_1, \ldots, b_{p-1}]$

$$b_i = \begin{cases} 1, & \mathbf{U}_j \in C_{\rho,\mathbf{U}_i} \cup C_{\rho,-\mathbf{U}_i} \\ 0, & o.w. \end{cases}$$

Then, have equivalent representation

$$E[\phi_i|\mathbf{U}_i] = P(\sum_{i=1}^{p-1} b_i > 0|\mathbf{U}_i)$$

Classical result [Thm. 4.5.4]{TW Anderson, 2003}:

Lemma

Let $\mathbf{X}$ be a p-variate elliptical vector with diagonal dispersion matrix $\mathbf{\Sigma}$. The Z-scores $\{\mathbf{U}_i\}_{i=1}^{p}$ are i.i.d. random vectors uniformly distributed on S_{n-2}.

Implication: b_i's are Bernoulli and $E[\phi_i|\mathbf{U}_i] = 1 - (1 - P_0)^{p-1}$.

Main result: correlation screening

Proposition

Let the $n \times p$ data matrix $\mathbb{X}$ have i.i.d. rows but possibly dependent columns. Let the sequence $\{\rho_p\}_p$ of correlation thresholds be such that $\rho_p \to 1$ and $p(p-1)\left(1-\rho_p^2\right)^{(n-2)/2} \to d_n$ for some finite constant d_n. Then

$$\lim_{p\to\infty} E[N] = \kappa_n J(\overline{f_{\mathbf{U}_\bullet,\mathbf{U}_{*-\bullet}}}), \tag{1}$$

where $\kappa_n = a_n d_n/(n-2)$ and $\overline{f_{\mathbf{U}_\bullet,\mathbf{U}_{*-\bullet}}}$ is limit of average density

$$\overline{f_{\mathbf{U}_\bullet,\mathbf{U}_{*-\bullet}}}^{(p)}(\mathbf{u},\mathbf{v}) = \frac{1}{p}\sum_{i=1}^{p}\frac{1}{p-1}\sum_{j\neq i}^{p}\left(\tfrac{1}{2}f_{\mathbf{U}_i,\mathbf{U}_j}(\mathbf{u},\mathbf{v}) + \tfrac{1}{2}f_{\mathbf{U}_i,\mathbf{U}_j}(\mathbf{u},-\mathbf{v})\right). \tag{2}$$

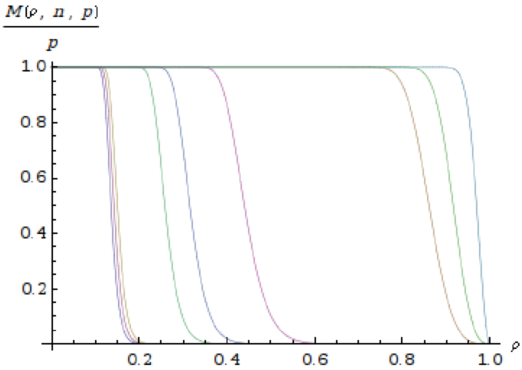
Implication: uniform Z-score density is minimax

- $J(\overline{f_{\mathbf{U}_\bullet, \mathbf{U}_{*-\bullet}}})$: related to Hellinger divergence and Rényi entropy

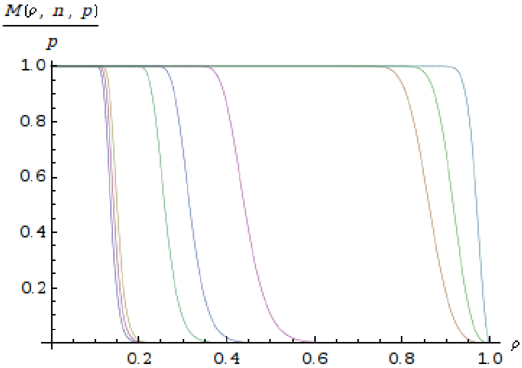
$$
\begin{aligned}
J(f_{\mathbf{U},\mathbf{V}}) &= |S_{n-2}| \int f_{\mathbf{U},\mathbf{V}}(\mathbf{w},\mathbf{w}) d\mathbf{w} \\
&= |S_{n-2}| \int \left(f_{\mathbf{U}|\mathbf{V}}(\mathbf{w}|\mathbf{w}) f_{\mathbf{V}|\mathbf{U}}(\mathbf{w}|\mathbf{w})\right)^{1/2} \left(f_{\mathbf{U}}(\mathbf{w}) f_{\mathbf{V}}(\mathbf{w})\right)^{1/2} d\mathbf{w} \\
&\le |S_{n-2}| \left(\int f_{\mathbf{U}|\mathbf{V}}(\mathbf{w}|\mathbf{w}) f_{\mathbf{V}|\mathbf{U}}(\mathbf{w}|\mathbf{w})\right)^{1/2} \left(\int f_{\mathbf{U}}(\mathbf{w}) f_{\mathbf{V}}(\mathbf{w})\right)^{1/2} \\
&\le H_2^{1/4}(f_{\mathbf{U}|\mathbf{V}}) H_2^{1/4}(f_{\mathbf{V}|\mathbf{U}}) H_2^{1/4}(f_{\mathbf{U}}) H_2^{1/4}(f_{\mathbf{V}}),
\end{aligned}
$$

- Equalities iff $f_{\mathbf{U},\mathbf{V}}(\mathbf{u},\mathbf{u}) = f_{\mathbf{U}}(\mathbf{u}) f_{\mathbf{V}}(\mathbf{u})$ and $f_{\mathbf{U}}(\mathbf{u}) = f_{\mathbf{V}}(\mathbf{u})$
- Right side of (3) minimized when $f_{\mathbf{U}}$ is uniform over S_{n-2}.

Implication: phase transition for correlation screening



Implication: phase transition for correlation screening



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 approximation: $\rho_c = \max\{\rho : dE[N]/d\rho = -1\}$

$$\rho_c = \sqrt{1 - c_n(p-1)^{-2/(n-4)}} \quad (3)$$

Main result: persistent correlation screening

Proposition

Let the $n_a \times p$ and $n_b \times p$ data matrices $\mathbb{X}^b$ and $\mathbb{X}^a$ be independent. Let $\rho_p^a \to 1$ and $\rho_p^b \to 1$ while for $\gamma = a, b$

$$p^{1/2}(p-1)\left(1-(\rho_p^\gamma)^2\right)^{(n_\gamma-2)/2} \quad \to \quad d_{n_\gamma}$$

Then

$$\lim_{p\to\infty} E[N^{a\wedge b}] = \kappa_n^{a\wedge b} \lim_{p\to\infty} \frac{1}{p}\sum_{i=1}^{p} J(\overline{f_{\mathbf{U}_i^a,\mathbf{U}_{*-i}^a}})J(\overline{f_{\mathbf{U}_i^b,\mathbf{U}_{*-i}^b}}), \tag{4}$$

where, for $\mathbf{U} \in \{\mathbf{U}^a, \mathbf{U}^b\}$,

$$\overline{f_{\mathbf{U}_i,\mathbf{U}_{*-i}}}(\mathbf{u},\mathbf{v}) = \frac{1}{p-1}\sum_{j\neq i}^{p}\left(\tfrac{1}{2}f_{\mathbf{U}_i,\mathbf{U}_j}(\mathbf{u},\mathbf{v}) + \tfrac{1}{2}f_{\mathbf{U}_i,\mathbf{U}_j}(\mathbf{u},-\mathbf{v})\right). \tag{5}$$

Persistent correlation screening: observations

- $\rho_a \to 1$, $\rho_b \to 1$ at slower rates than before.
- When $J(\overline{f_{\mathbf{U}_i^a,\mathbf{U}_{*-i}^a}})$, $J(\overline{f_{\mathbf{U}_i^b,\mathbf{U}_{*-i}^b}})$ are asymptotically *incoherent*

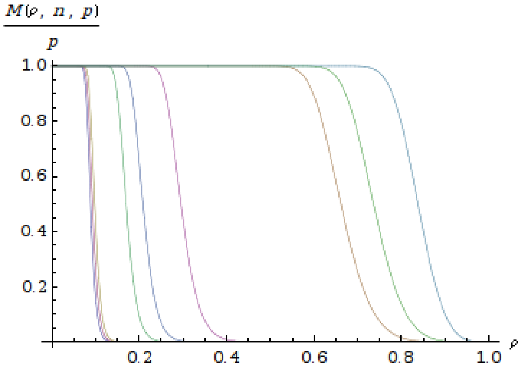
$$\lim_{p\to\infty} \frac{1}{p} \sum_{i=1}^{p} J(\overline{f_{\mathbf{U}_i^a,\mathbf{U}_{*-i}^a}}) J(\overline{f_{\mathbf{U}_i^b,\mathbf{U}_{*-i}^b}}) = J(\overline{f_{\mathbf{U}_\bullet^a,\mathbf{U}_{*-\bullet}^a}}) J(\overline{f_{\mathbf{U}_\bullet^b,\mathbf{U}_{*-\bullet}^b}})$$

Then, as $p \to \infty$,

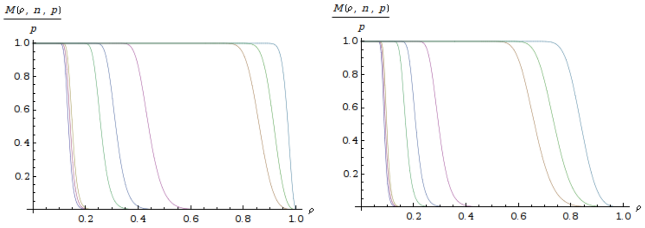
$$E[N^{a\wedge b}] \to \frac{E[N^a]E[N^b]}{p}$$

- $p^{-1/2}E[N_a]$, $p^{-1/2}E[N_b]$ converge but $E[N_a]$, $E[N_b]$ do not.

Implication: phase transition for persistent correlation screening



Phase transitions: correlation vs persistent correlation screening



Outline

Application: correlation screening with spike-in

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

Table: Achievable limits in FPR (α) for TPR =0.8 (β), as function of n, minimum detectable threshold, and correlation threshold ($\rho_1 \backslash \rho$). To obtain entries $\rho_1 \backslash \rho$ a Poisson approximation determined $\rho = \rho(\alpha)$ and a Fisher-Z Gaussian approximation determined $\rho_1 = \rho_1(\beta)$. Here $p = 1000$ on Gaussian sample having diagonal covariance with a spike-in correlated pair.

Application: correlation screening with spike-in

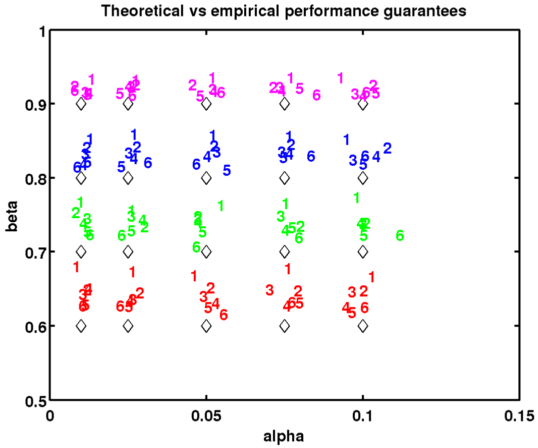


Figure: Comparison between predicted (diamonds) and actual (numbers) operating points (α, β) using Poisson approximation to false positive rate (α) and Fisher approximation to false negative rate (β). Each number is located at an operating point determined by the sample size n ranging over $n = 10, 15, 20, 25, 30, 35$. These numbers are color coded according to the target value of β.

Application: persistent correlation discovery

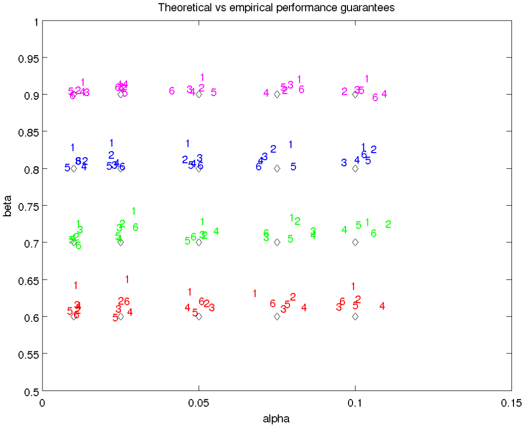


Figure: Comparison between predicted (diamonds) and actual (numbers) operating points (α, β) for perisistent correlation screening.

Application: gene expression data

Beverage Data from Gene Expression Omnibus (GEO) NCBI

- Reference: Florent Baty *etal* (2006) BMC Bioinformatics
- Subjects: 6 individuals at 5 time points (0, 1, 2, 4, 12 hours)
- Treatments: post-baseline intake of
  - A: alcohol ($n_1 = 20$)
  - G: grape juice ($n_2 = 22$)
  - H: water ($n_3 = 23$)
  - W: red wine ($n_4 = 22$)
- 87 Affymetrix HU133 Genechip peripheral blood samples
- Each sample contains $p = 22,283$ gene probes

Application: observed Z-scores

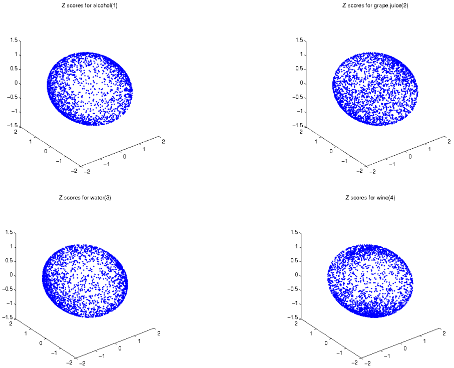


Figure: 3 dimensional projections of the Z-scores for the experimental beverage data under each of the treatments A,G,H,W. For visualization the 22,238 variables (gene probes) were downsampled by a factor of 8 and a randomly selected set of four samples in each treatment were used to produce these figures.

Application: persistent correlation discoveries

$\{A\}, \{G\}, \{H\}, \{W\}$		42	50	82	424	
$\{A,G\}, \{A,H\}, \{A,W\}, \{G,H\}, \{G,W\}, \{H,W\}$	493	748	1069	677	864	1445
$\{G,H,W\}, \{A,H,W\}, \{A,G,W\}, \{A,G,H\}$		2242	2530	1893	1690	
$\{A,G,H,W\}$			3313			

Table: Number of genes discovered by auto-screening (top row) and persistency screening (lower three rows) for various combinations of treatments in the experimental data. Auto-screening threshold determined using Poisson approximation to Type I error of level 10^{-5}.

Application: set-inclusion diagram

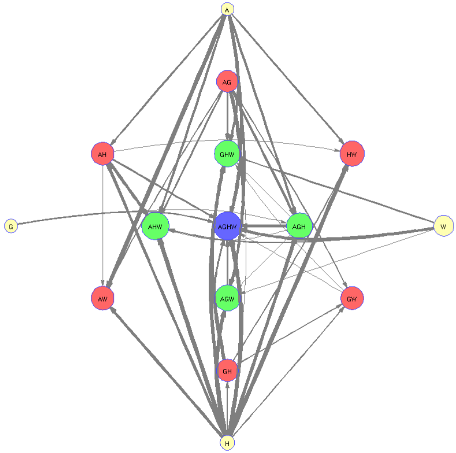


Figure: Set-inclusion graph between genes discovered by correlation screening in various combinations of treatments. Size of node is proportional to the log of number of associated correlation screening discoveries given in Table 2. A directed edge from node i to node j exists if at least 90% of the genes discovered in node i are also discovered in node j and the thickest edges indicate 100% set inclusion. The asymmetry of diagram indicates that treatments have different effects on gene expression.

Application: persistent covariance network

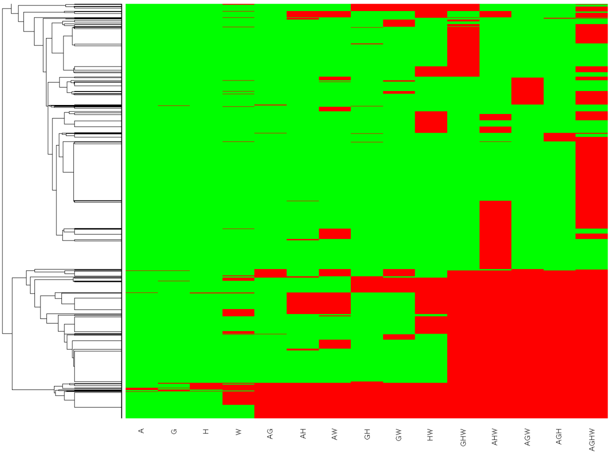


Figure: Heatmap of 4444 genes discovered in at least one of the set inclusion tests shown in Table 2.

Application: persistent covariance network

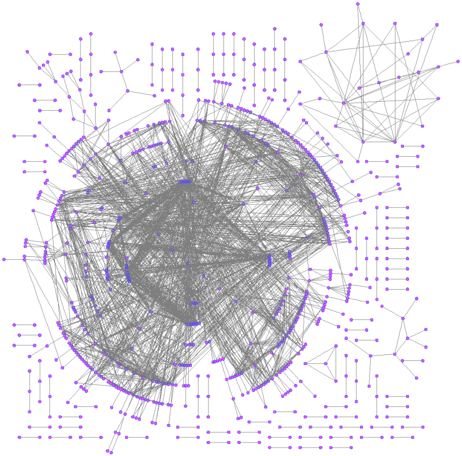


Figure: 774 gene subnetwork of the 3313 gene persistent-correlation network across all four treatments corresponding to the last row of Table 2. Two nodes in this network are linked by an edge if for all 4 treatments their sample correlation is above the 10^{-5} FWER correlation-screening threshold.

Outline

Conclusions

- Correlation screening can be performed with confidence
  - Screening affected by phase transition as threshold decreases
  - Large p expressions for critical PT threshold ρ_c are available
  - Effect of pairwise dependence manifested through Hellinger divergence

Conclusions

- Correlation screening can be performed with confidence
    - Screening affected by phase transition as threshold decreases
    - Large p expressions for critical PT threshold ρ_c are available
    - Effect of pairwise dependence manifested through Hellinger divergence
- Key concepts:
    - Z-score representation of sample correlation
    - Geometry of unit sphere

Conclusions

- Correlation screening can be performed with confidence
  - Screening affected by phase transition as threshold decreases
  - Large p expressions for critical PT threshold ρ_c are available
  - Effect of pairwise dependence manifested through Hellinger divergence
- Key concepts:
  - Z-score representation of sample correlation
  - Geometry of unit sphere
- Persistence: Strongest specialists are not strongest generalists