Correlation screening in high dimension

Alfred Hero

University of Michigan, Ann Arbor

December 3, 2009

Acknowledgements

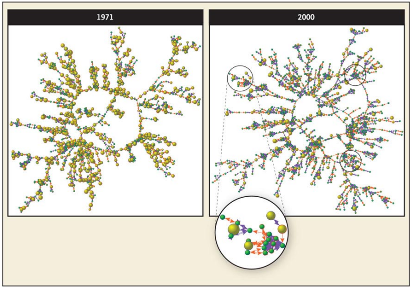
- Kumar Sricharan (UM Grad student)
- Bala Rajaratnam (Stanford)

- NSF: ITR CCR-032557
- AFOSR: FA9550-06-1-0324
- ONR: N00014-08-1-1065
- ARO: W911NF-05-1-0403
- DIGITEO, Paris France

Outline

Smoker network

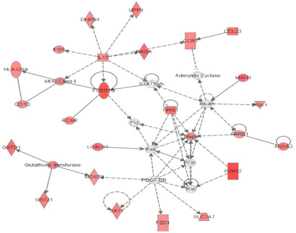
Social interaction network (Framingham study, NEJM 2008)



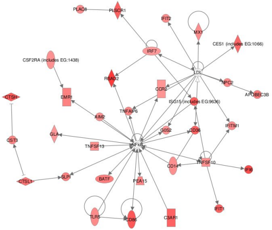
- By 2000 smokers more likely to be at periphery of their networks and in smaller subgroups than non-smokers (see dark circled areas)
- Size of circle: number of cigarettes per day
- Yellow circle: smoker
- Green circle: non-smoker

Curated gene expression networks

Canonical Pathway Involvement by Significant Genes: Immunological Diseases

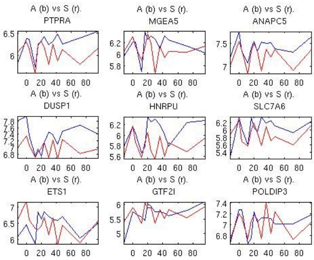


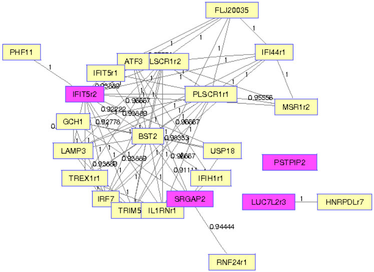
Canonical Pathway Involvement by Significant Genes: Cellular Growth and Proliferation / Organism Injury



Discovery of gene expression networks

For testing correlations between gene samples on a Affy gene microarray chip need to test $\binom{24{,}000}{2}$ sample correlations based on small sample size (here N=8).





Is discovered correlation network statistically significant?

Why sample correlation?

Sample correlation has been of great interest in signal processing

- Invariant to translation and scale transformations on variables
- Used to discover of dependency structure and graphical models (Willsky, Jordan)
- Used to estimate number of signals in a random mixture (Nadakaduti and Edelmann, Wax and Kailath)
- Used in spectral analysis and sensor array beamforming (Parzen, Schultheiss)

Outline

Correlation screening

- p-variate random sample: $\mathbf{X} = [X_1, \ldots, X_p]^T$
- $p \times p$ covariance matrix (unknown): $\Sigma = E[\mathbf{X}\mathbf{X}^T]$
- **Objective**: given n i.i.d. samples $\mathbb{X} = [\mathbf{X}_1, \ldots, \mathbf{X}_n]^T$ detect highest correlations
- Difficulty: $p \gg n$

Sample covariance matrix:

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

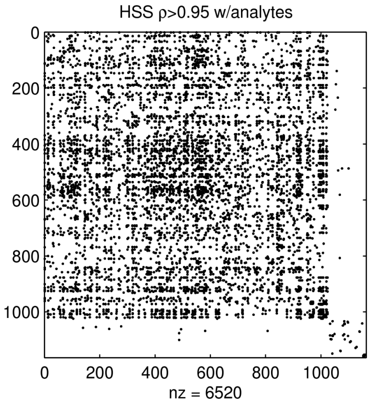
Sample correlation matrix:

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

where $\hat{\mathbf{D}} = \mathrm{diag}(\hat{\Sigma})$.

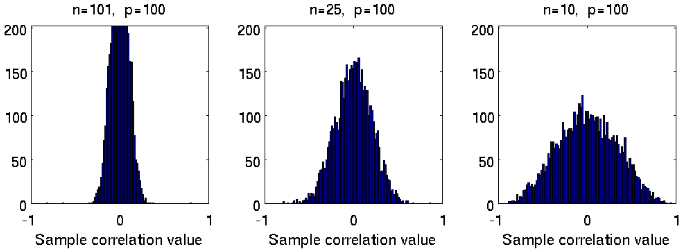
Thresholded sample correlation matrix

- Define $\rho_{ij} = (\mathbf{R})_{ij}$ and ρ a user-defined threshold in $[0,1]$
- Fisher's correlation screening test: $|\rho_{ij}| > \rho$
- Screening gives set of "discovered" (i,j) correlation pairs



Phase transitions in correlation screening

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



Mathematical results

Two types of results obtained

- Characterize large p phase transition and its threshold.
- Predict mean discovery rate and p-values for correlation screening and persistent correlation screening.

How we approach the analysis

- Start with assuming Gaussian diagonal covariance null model
- Extend results to dependent or non-Gaussian null model

Basis for analysis

- Projected Z-scores embedding of sample correlation
- Geometric probability on $(n-1)$-sphere $S_{n-1} \subset \mathbb{R}^{n-1}$
- Exchangeable process theory for handling dependent variables

Z-score representation of sample correlation

- Z-score representation of correlation matrix

$$\mathbf{R} = \mathbb{Z}^T \mathbb{Z}$$

$$\mathbb{Z} = [\mathbf{Z}_1, \ldots, \mathbf{Z}_p] = (n-1)^{-1/2}(\mathbf{I} - n^{-1}\mathbf{1}\mathbf{1}^T)\mathbb{X}\mathbf{D}^{-1/2}.$$

- $\mathbf{Z}_i$ standardizes $\mathbf{X}_i$ by scale/translation transformation

$$\mathbf{Z}_i = \frac{\mathbf{X}_i - \hat{\mu}_i \mathbf{1}}{\hat{s}_i \sqrt{n-1}}, \quad i = 1, \ldots, p$$

$$\hat{\mu}_i = \frac{1}{n}\sum_{i=1}^{n} X_{ij}, \quad \hat{s}_i^2 = \frac{1}{n-1}\sum_{j=1}^{n}(X_{ij} - \hat{\mu}_j)^2$$

- n-dimensional $\mathbf{Z}_i$ lies in $n-2$ dimensional subpsace

$$\mathbf{1}^T \mathbf{Z}_i = 0 \text{ and } \|\mathbf{Z}_i\| = 1$$

Sample correlation and Z-score distances

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

S_{n-1} embedding via projected Z-scores

Easier to work with projected Z-scores $\mathbb{U} = [\mathbf{U}_1, \ldots, \mathbf{U}_p]$

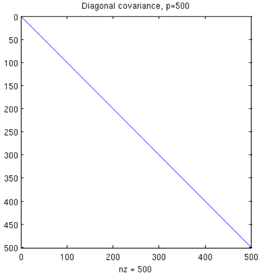
- $\mathbf{U}_i$ are $(n-1)$-element summaries of n-element $\mathbf{Z}_i$
- $\mathbf{U}_i$ satisfy $\|\mathbf{U}_i\| = 1$ and lie on sphere $S_{n-1} \subset \mathbb{R}^{n-1}$
- $\mathbb{U}$ gives more parsimonious representation than $\mathbb{Z}$

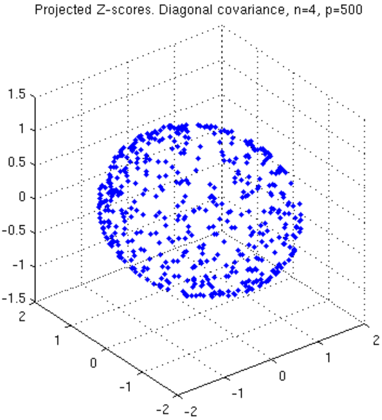
$$\mathbf{R} = \mathbb{U}^T \mathbb{U}$$

- $\rho_{ij} = \mathbf{U}_i^T \mathbf{U}_j$ and geodesic distance between $\mathbf{U}_i$ and $\mathbf{U}_j$ satisfies

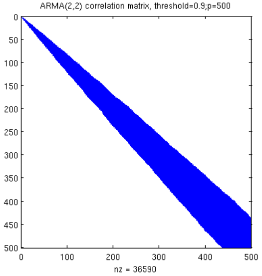
$$d(\mathbf{U}_i, \mathbf{U}_j) = \arccos(\rho_{ij})$$

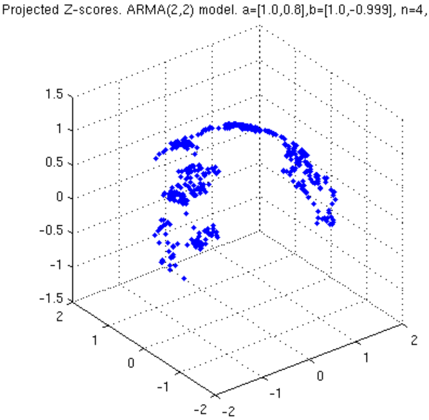
S_{n-1} embedding example: diagonal Gaussian





S_{n-1} embedding example : ARMA(2,2) Gaussian





Phase transition analysis

Define $\phi = [\phi_1, \ldots, \phi_p]$ the "discovery" indicator sequence:

$$\phi_i = \begin{cases} 1, & \max_{j \neq i} |\rho_{ij}| > \rho \\ 0, & o.w. \end{cases}$$

Define N the number of discoveries:

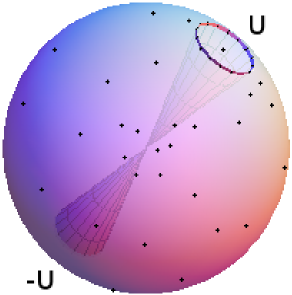
$$N = \sum_{i=1}^{p} \phi_i$$

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

Phase transition analysis

Conditional expectation of ϕ_i has representation

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



Phase transition analysis: diagonal Gaussian case

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 of multivariate statistics [Thm. 4.5.4]{TW Anderson, 2003}:

Lemma

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

Phase transition analysis: diagonal Gaussian case

Implication: $[b_1, \ldots, b_{p-1}]$ is i.i.d. Bernoulli sequence and

$$E[\phi_i|\mathbf{U}_i] = 1 - B(0, P_0, p-1)$$

where

$$B(k, \theta, m) = \binom{m}{k} \theta^k (1-\theta)^{m-k}$$

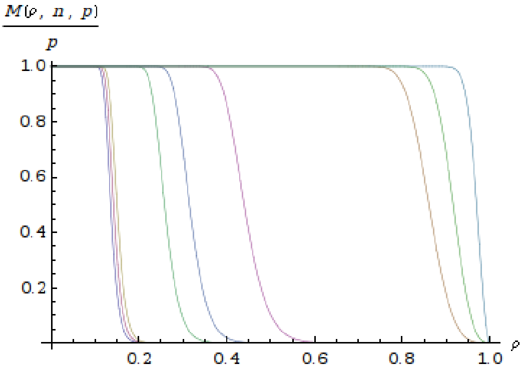
and

$$P_o = P_o(\rho, n) = \frac{2\Gamma((n-1)/2)}{\sqrt{\pi}\Gamma((n-2)/2)} \int_0^{\arccos(\rho)} \sin^{(n-3)}(\theta) d\theta. \qquad (1)$$

Phase transition analysis: diagonal Gaussian case

Result: mean number of false discoveries

$$E[N] = M(\rho, n, p) \stackrel{\text{def}}{=} p(1 - B(0, P_0, p-1)) = p(1-(1-P_0)^{p-1})$$



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

Phase transition analysis: diagonal Gaussian case

Proposition

The slope of $E[N]$ *is*

$$dE[N]/d\rho = -p(p-1)(1-P_o)^{p-2}(1-\rho^2)^{\frac{n-4}{2}} c_n,$$

where

$$c_n = (2\Gamma((n-1)/2)/(\sqrt{\pi}\Gamma(n/2-1)))^{-2/(n-4)}.$$

Critical threshold $\rho_c = \max\{\rho : dE[N]/d\rho = -1\}$ *is*

$$\rho_c = \sqrt{1 - c_n(p-1)^{-2/(n-4)}}, \quad (pP_o \ll 1) \tag{2}$$

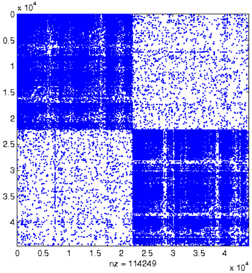
Outline

Persistent correlation screening

- Pair of p-variate random vectors: $\mathbf{X}^a = [X_1^a, \ldots, X_p^a]^T$, $\mathbf{X}^b = [X_1^b, \ldots, X_p^b]^T$
- $p \times p$ covariance matrices: Σ^a, Σ^b
- **Objective**: Discover variables with correlations that persist in *a* and *b* given samples
  - $\mathbb{X}^a = [\mathbf{X}_1^a, \ldots, \mathbf{X}_{n_a}^a]^T$
  - $\mathbb{X}^b = [\mathbf{X}_1^b, \ldots, \mathbf{X}_{n_b}^b]^T$
- **Method**: jointly screen sample correlation matrices: $\mathbf{R}^a$ and $\mathbf{R}^b$.

Thresholded sample correlation matrices

- Given sample correlations ρ_{ij}^a, ρ_{ij}^b and thresholds ρ^a, ρ^b
- Variable i declared PC if both $\max_{j\neq i} |\rho_{ij}^a| > \rho^a$ and $\max_{j\neq i} |\rho_{ij}^b| > \rho^b$
- L is number of persistent correlation discoveries



PC phase transition analysis

Define $\phi^a = [\phi_1^a, \ldots, \phi_p^a]$ and $\phi^b = [\phi_1^b, \ldots, \phi_p^b]$ the a and b discovery indicator vectors.
Define M and N the number of discoveries in a and b

$$M = \sum_{i=1}^{p} \phi_i^a, \qquad N = \sum_{i=1}^{p} \phi_i^b$$

Then L, the number of common discoveries, is

$$L = \sum_{i=1}^{p} \phi_i^a \phi_i^b$$

Objective: Find expressions for $E[L]$ as a function of $p, n_a, n_b, \rho^a, \rho^b$.

PC phase transition analysis: diagonal Gaussian case

Proposition

Assume that the two sets of observations $\{\mathbf{X}_n^a\}_{n=1}^{n_a}$ and $\{\mathbf{X}_n^b\}_{n=1}^{n_b}$ are mutually independent and each is composed of i.i.d. p-variate Gaussian random vectors with diagonal covariances. Then

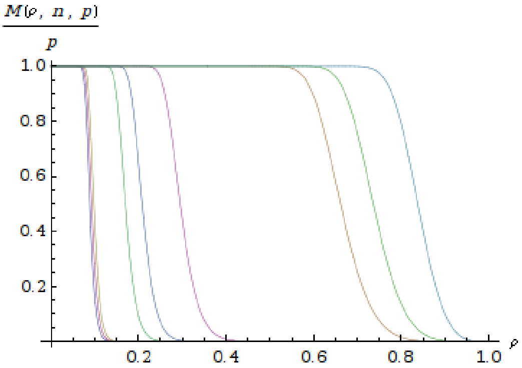
$$P(L=k) = \frac{1}{k!}\left(\frac{E[N]E[M]}{p}\right)^k (1+O(1/p)),\ 0 < k \le p \qquad (3)$$

and

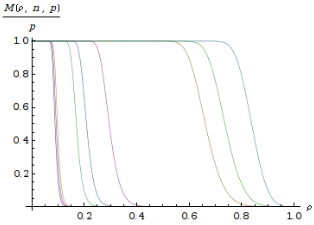
$$P(L=0) = \exp\left(-\frac{E[N]E[M]}{p}\right)(1+O(1/p)).$$

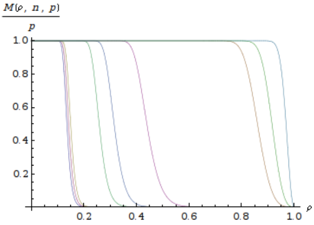
PC phase transition analysis: diagonal Gaussian case

Mean number of persistent discoveries: $E[L] = \frac{E[M]E[N]}{p}$



PC phase transition vs previous phase transition





Proof of PC Proposition

Note: L is number of matching "1"s in binary sequences ϕ^a, ϕ^b.
As these sequences are Bernoulli, conditioned on N, M we have

$$P(L = k|N, M) = \frac{\left(\frac{p!}{k!(N-k)!(M-k)!(p-N-M+k)!}\right)}{\binom{p}{M}\binom{p}{N}}, \quad 0 \le k \le \min\{N, M\}.$$

or, applying Stirling approximation to terms involving p,

$$P(L = k|N, M) = \frac{N!M!}{k!(N-k)!(M-k)!} p^{-k} \left(1 + O(NM/p)\right).$$

As M,N are binomial, elementary combinatorial identities yield

$$P(L = k) = \frac{1}{k!} \left(\frac{E[N]E[M]}{p}\right)^k \left(1 + O(1/p)\right), \; 0 < k \le p$$

Outline

Extension to arbitrary distributions

Central concept: invariance of M, N and L to index reordering.

- For π an arbitrary permutation:

$$N = \sum_{i=1}^{p} \phi_i = \sum_{i=1}^{p} \phi_{\pi(i)}$$

- An **exchangeable sequence** of binary random variables $\mathbf{b}_1, \ldots, \mathbf{b}_p$ has probability mass function f that satisfies

$$f_{\mathbf{b}_{\pi(1)},\ldots,\mathbf{b}_{\pi(p)}}(b_1, \ldots, b_p) = f_{\mathbf{b}_1,\ldots,\mathbf{b}_p}(b_1, \ldots, b_p)$$

Extension to arbitrary distributions: de Finetti theorem

Proposition

(Diaconis and Freedman, 1980) A length p subsequence of a length P exchangeable binary sequence $\mathbf{b}_1, \ldots, \mathbf{b}_P$, is almost i.i.d. in the sense that there exists a distribution μ on $[0,1]$ such that

$$\|f_{\mathbf{b}_1,\ldots,\mathbf{b}_p}(b_1, \ldots, b_p) - \int \theta^N (1-\theta)^{p-N} \mu(d\theta)\| \leq \frac{4p}{P}$$

where $N = \sum_{i=1}^{p} b_i$. Furthermore,

$$E[\theta] = \frac{1}{P} \sum_{i=1}^{P} E[b_i]$$

Correlation screening with dependencies

Single treatment correlation screening with dependencies.

Proposition

Let the $n \times p$ random matrix $\mathbb{X}$ have independent rows but possibly dependent columns. Then

$$E[N] = p\left((p-1)P_0 H_2(\overline{f_{\mathbf{U}}}) + \epsilon\right), \tag{4}$$

where

$$H_2(\overline{f_{\mathbf{U}}}) = |S_{n-1}| \int_{S_{n-1}} \overline{f_{\mathbf{U}}}^2(\mathbf{u})d\mathbf{u}$$

with $\overline{f_{\mathbf{U}}} = p^{-1}\sum_{i=1}^p f_{\mathbf{U}_i}$ the avg population density and

$$\epsilon \leq (pP_0 \sup \overline{f_{\mathbf{U}}})^2.$$

Implications of Proposition

- Effect of multivariate dependency on $E[N]$ is inflation by factor
$$H_2(\overline{f_{\mathbf{U}}}) = |S_{n-1}| \int_{S_{n-1}} \overline{f_{\mathbf{U}}}^2(\mathbf{u})d\mathbf{u}.$$
- $1 \leq H_2(\overline{f_{\mathbf{U}}}) < \infty$, with "$=1$" iff $\overline{f_{\mathbf{U}}}$ is uniform over S_{n-1} and $= \infty$ iff $\overline{f_{\mathbf{U}}}$ is dirac.
- $H_2(\overline{f_{\mathbf{U}}})$ is decreasing in Rényi α-entropy of order $\alpha = 2$.
- Phase transition threshold is
$$\rho_c = \sqrt{1 - d_n(p-1)^{-2/(n-4)}},$$
where $d_n = c_n H_2(\overline{f_{\mathbf{U}}})$.

Proof of Proposition

Recall definitions: $\phi_i = I(\sum_{i=1}^{p-1} b_i > 0)$, $N = \sum_{i=1}^{p} \phi_i$,

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

Wrt N, b_i is subsequence of infinite exchangable sequence. Therefore, to order $O(p^2 E^2[\theta|\mathbf{U}_i])$:

$$E[\phi_i|\mathbf{U}_i] = 1 - \int B(0,\theta,p-1)\mu(d\theta) = (p-1)E[\theta|\mathbf{U}_i]$$

By the de Finetti representation, to order $O(\sup \overline{f_{\mathbf{U}}}/p)$

$$E[\theta|\mathbf{U}_i] = \frac{1}{p-1}\sum_{j\neq i} E[b_i|\mathbf{U}_i] = \int_{C_{\rho,\mathbf{U}_i}\cup C_{\rho,-\mathbf{U}_i}} \overline{f_{\mathbf{U}}}(u)du.$$

Therefore, applying MVT and summing over i,

$$E[N] = \sum_{i=1}^{p} E[\phi_i] = p(p-1)|S_{n-1}P_0 \int_{S_{n-1}} \overline{f_{\mathbf{U}}}^2(u)du$$

Persistent correlation screening with dependencies

Proposition

Assume that two sets of observations $\{\mathbf{X}_n^a\}_{n=1}^{n_a}$ *and* $\{\mathbf{X}_n^b\}_{n=1}^{n_b}$ *are mutually independent, each composed of i.i.d. p-variate random vectors. Then the mean number of discovered PC's is*

$$E[L] = E_0[L]\ H_2(\overline{f_{\mathbf{U}^a} f_{\mathbf{U}^b}}) H_2(\overline{f_{\mathbf{U}^a}}\ \overline{f_{\mathbf{U}^b}})\ A(\overline{f_{\mathbf{U}^a} f_{\mathbf{U}^b}}, \overline{f_{\mathbf{U}^a}}\ \overline{f_{\mathbf{U}^b}})$$

where $E_0[L]$ *is mean for diagonal Gaussian case,* $A(g,h)$ *is*

$$A(g,h) = \frac{\int gh}{\sqrt{\int g^2}\sqrt{\int h^2}}$$

and $\overline{f_{\mathbf{U}_i^a} f_{\mathbf{U}^b}} = \frac{1}{p}\sum_{i=1}^{p} f_{\mathbf{U}_i^a} f_{\mathbf{U}_i^b}$

Implications of dependent PC Proposition

- Affinity $A(g,h)$ is normalized l_2 inner product between distributions h and g on $S_{n_a-1} \times S_{n_b-1}$

$$0 \leq A(g,h) \leq 1$$

- $A(\overline{f_{\mathbf{U}^a} f_{\mathbf{U}^b}}, \overline{f_{\mathbf{U}^a}}\ \overline{f_{\mathbf{U}^b}}) = 1$ iff $f_{\mathbf{U}^a_i}$ and $f_{\mathbf{U}^b_i}$ do not depend on i.
- $E[L] = E_0[L]$ if $f_{\mathbf{U}^a}$ and $f_{\mathbf{U}^b}$ uniform on S_{n_a-1} and S_{n_b-1}.

Proof of dependent PC Proposition

Wrt $L = \sum_{i=1}^{p} \phi_i^a \phi_i^b$, $\{\phi_i^a \phi_i^b\}_{i=1}^{p}$ is a segment of an infinite exchangable sequence. Therefore, by de Finetti

$$P(L = k | \mathbf{U}_i^a, \mathbf{U}_i^b) = \binom{p}{k} \int \theta^k (1-\theta)^{p-k} \mu(d\theta)$$

with

$$E[\theta | \mathbf{U}_i^a, \mathbf{U}_i^b] = p^{-1} \sum_{i=1}^{p} E[\phi_i^a | \mathbf{U}_i^a] E[\phi_i^b | \mathbf{U}_i^b].$$

From previous proposition

$$\begin{aligned} E[\phi_i^a | \mathbf{U}_i^a = \mathbf{u}^a] &= (p-1) P_0(\rho^a, n^a) |S_{n_a-1}| f_{\mathbf{U}_i^a}(\mathbf{u}^a) \\ E[\phi_i^b | \mathbf{U}_i^b = \mathbf{u}^b] &= (p-1) P_0(\rho^b, n^b) |S_{n_b-1}| f_{\mathbf{U}_i^b}(\mathbf{u}^b) \end{aligned}$$

Mean is therefore

$$E[L] = E_0[L] \int d\mathbf{u}^a \int d\mathbf{u}^b \left(\frac{1}{p} \sum_{i=1}^{p} f_{\mathbf{U}_i^a}(\mathbf{u}^a) f_{\mathbf{U}_i^b}(\mathbf{u}^b) \right) \overline{f_{\mathbf{U}_i^b}}(\mathbf{u}^b) \overline{f_{\mathbf{U}_i^b}}(\mathbf{u}^b).$$

Outline

Application: screening for high correlations

Consider testing simple null hypotheses on γ_{ij}

$$H_i \quad : \quad g(\gamma_{ij}) = 0, \;\; \forall j \neq i, \; j = 1, \ldots, p$$

Objective

For given p, n and average false positive rate $\alpha = P(N > 0|H)$ what is the minimum detectable level ρ_1 of correlation?

- $N = \sum_{i=1}^{p} \phi_i$ is number of false positives
- For large p, fpr $P(N > 0|H)$ is approximately Poisson$(E[N])$
- Using Gaussian distribution of Fisher Z transform, tpr $P(\phi_{true} = 1|H^c)$ can be computed

Application: screening for high correlations

$n \setminus \alpha$	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: Minimum detectable correlation and level-α threshold (given as entry ρ_1/ρ in table) for $p = 1000$ and $\beta = 0.8$.

Application: screening for high correlations

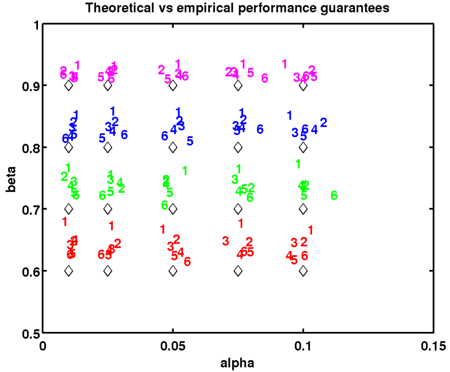


Figure: Comparison between predicted (diamonds) and actual (numbers) operating points (α, β) using the star-shaped decomposition and Poisson approximation to false positive rate (α) and Fisher approximation to false negative rate (β). Each number is located at an operating point determined by the sample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

Outline

Conclusions

- Correlation and persistent correlation screening are important in applications
- Screening negatively affected by false positive phase transition as function of threshold
- Asymptotic expression for critical PT threshold ρ_c is available for single treatment
- Effect of dependency on phase transitions is mediated by Rényi 2-entropy of average marginal density on sphere
- Key concepts:
  - Stochastic representation of sample correlation on sphere
  - Exchangeable processes