Statistical procedures for network structures identification

Petr Koldanov

National Research University Higher School of Economics,
Laboratoty of Algorithms and Technologies for Network Analysis (LATNA)
Nizhny Novgorod, Russia
Statistical network analysis team
pkoldanov@hse.ru

Moscow, Russia, April 26, 2017

General problem

- $X = (X_1, X_2, \ldots, X_N)$-random vector.
- $f(x) \in \{f(x, \theta); \theta \in \Omega\}$
- $H_i : \theta \in \Omega_i, \Omega_i \subset \Omega, i = 1, \ldots, L$
- $x(1), x(2), \ldots, x(n)$ - sample of finite size from sample space $\mathcal{X} = R^{N \times n}$.
- Construct $\delta(x) : \mathcal{X} \to D, D = \{d_1, d_2, \ldots, d_L\}$

Random variable network

Random variable network is a pair (X, γ):

- $X = (X_1, \ldots, X_N)-$random vector,
- $\gamma-$measure of association.

Applications:

- Example 1 - market network (nodes correspond to the stocks, behaviour of stocks is described by returns (portfolio theory). Popular network in stock market:=Pearson network: $\gamma_{i,j}^{P} = \rho_{i,j} = \frac{\sigma_{i,j}}{\sqrt{\sigma_{i,i}\sigma_{j,j}}}$
- Example 2 - biological (gene experssion) network. Popular network in biology:=Partial correlation network: $\gamma_{i,j}^{part} = \rho^{i,j} = \frac{-\sigma^{i,j}}{\sqrt{\sigma^{i,i}\sigma^{j,j}}}$

Network model

Any random variable network generate network model.

- Complete weighted graph $G = (V, E, \gamma)$.
- Nodes of the network model - elements of the system.
- Weights of edges in the network model are given by some measure γ of connection between elements of the system.

Network structures - subgraphs of the network model.
$G' = (V', E') : V' \subseteq V, E' \subseteq E$

- Network structures contain useful information on the network model.
- Popular network structures for market network: maximum spanning tree (MST), market graph (MG), cliques and independent sets of MG.
- Popular network structures for gene expression network: Gaussian Graphical Model (concentration graph).

References

Market network

- Mantegna(1999) - MST for market network.
- Pardalos (2003) - MG for market network.
- Now there are around thousand papers.
- Main purpose - network structure calculation by numerical algorithms to real data and interpretation of obtained results. Examples of interpretation.
- No uncertainty analysis of obtained results.

Gene-expression network.

- Lauritzen(1996), Drton & Perlman(2007)
- Thousands of publications.
- Numerical algorothms. Only FWER under control.
- No results for finite sample size.

Our approach. Multiple decision approach

- (X, γ)-random variable network, $G = (V, E, \gamma)$-generated network model.
- $G' = (V', E') : V' \subseteq V, E' \subseteq E$ - network structure.
- X has a distribution from class $\mathcal{K} = \{(f(x, \theta), \theta \in \Omega)\}$.
- Let $S = (s_{i,j})$, $S \in \mathcal{G}$ - set of all adjacency matrices.
- $H_S : \theta \in \Omega_S$-hypothesis that network structure has adjacency matrix $S, S \in \mathcal{G}_1 \subseteq \mathcal{G}$.
- Observation $X(t) = (X_1(t), \ldots, X_N(t)), t = 1, \ldots, n$

Problem: construct statistical procedure $\delta(x)$ with appropriate properties to identify network structure from observations i.e. to select one from disjoint hypotheses H_S.

Quality of statistical procedures for network structure identification

- Statistical procedure $\delta(x) = \begin{cases} d_Q, & x \in D_Q \end{cases}$
  $\bigcup_{Q \in \mathcal{G}} D_Q = \mathcal{X}$ is the partition of sample space.
- $\delta(x) = d_Q$ - decision, that network structure has adjacency matrix $Q, Q \in \mathcal{G}$.
- $w(H_S; d_Q) = w(S, Q)$ - loss from the decision d_Q when the hypothesis H_S is true, $w(S, S) = 0, S \in \mathcal{G}$.
- Risk function (uncertainty) of statistical procedure $\delta(x)$ is defined by

$$Risk(S, \theta; \delta) = \sum_{Q \in \mathcal{G}} w(S, Q) P_\theta(\delta(x) = d_Q), \quad \theta \in \Omega_S, S \in \mathcal{G}$$

$P_\theta(\delta(x) = d_Q)$ - the probability that decision d_Q is taken while the true decision is d_S.

Main results

- optimal procedures for network structures identification.
- statistical procedures for network structures identification with invariant risk function.

Part 1. Procedures with invariant risk function.

Procedure with invariant risk function.

In practical applications it can be supposed that distribution of X from $\mathcal{K} = \{f(x;\theta), \theta \in \Omega\}$
Properties of statistical procedures for network structures identification may depend on distribution of $X \in \mathcal{K}$.
Problem: construct statistical procedure with invariant risk function for network structure identification.

Definition: statistical procedure δ has invariant risk function in class $\mathcal{K}$, if risk function $Risk(S, \theta, \delta)$ does not depend from distribution of vector X from class $\mathcal{K}$ for any S.

Class $\mathcal{K}$ of elliptical distributions

Most common models of stock market is the class of elliptical distributions.

$$f(x;\theta) = |\Lambda|^{-\frac{1}{2}} g\{(x-\mu)'\Lambda^{-1}(x-\mu)\} \tag{1}$$

where $\theta = (\mu, \Lambda, g)$, $\mu \in R^N$, Λ - symmetric positive definite matrix, $g(x) \geq 0$, and

$$\int_{-\infty}^{\infty} \dots \int_{-\infty}^{\infty} g(y'y)dy_1 \dots dy_N = 1$$

In the following we assume that μ is known.

Pearson correlation network, sign similarity network

Let $\mathcal{K}(\Lambda)$ be the subclass of class of elliptical distributions with fixed Λ.

- Pearson correlation network (X, γ^p) $\gamma^P_{i,j} = \frac{\lambda_{i,j}}{\sqrt{\lambda_{i,i}\lambda_{j,j}}}$. Then network models, generated by $(X, \gamma^p), X \in \mathcal{K}(\Lambda)$, coincide.
- Sign similarity network (X, γ^{Sg})
  $\gamma^{Sg}_{i,j} = p^{i,j} = P((X_i - E(X_i))(X_j - E(X_j) > 0)$.

Lemma 1.1: Probabilities $\gamma^{Sg}_{i,j} = P((X_i - E(X_i))(X_j - E(X_j) > 0)$ are defined by the matrix Λ and does not depend from g.
Then network models, generated by $(X, \gamma^{Sg}), X \in \mathcal{K}(\Lambda)$, coincide.

Threshold graph (TG) identification

We are looking for statistical procedures with invariant risk function in the class $\mathcal{K}(\Lambda)$ i.e. the risk function (for any S) does not depend from g.

Individual edge hypotheses:

$$h_{ij} : \gamma_{ij} \leq \gamma_0 \text{ vs } k_{ij} : \gamma_{ij} > \gamma_0.$$

Individual tests:

$$\varphi_{ij}(X) = \begin{cases} 1, & t_{ij}(X) > c_{ij} \\ 0, & t_{ij}(X) \leq c_{ij} \end{cases}$$

Multiple testing statistical procedure: statistical procedure, based on statistics of individual tests.

- Single step procedures (Bonferroni and others)
- Stepwise procedures (Holm, Hochberg and others)

Pearson correlation network

Individual hypotheses (Pearson measure): $h_{ij} : \rho_{i,j} \leq \rho_0$ vs $k_{ij} : \rho_{i,j} > \rho_0$

- $\varphi^P_{i,j}(x_i, x_j) = \begin{cases} 1, & \frac{r_{i,j}-\rho_0}{\sqrt{1-r^2_{i,j}}} > c^{St}_{i,j} \\ 0, & \frac{r_{i,j}-\rho_0}{\sqrt{1-r^2_{i,j}}} \leq c^{St}_{i,j} \end{cases}$ UMP in the class of invariant tests
- $c^{St}_{i,j}$ is $(1-\alpha_{ij})$-quantile of Student distribution t_{n-1},
- $\alpha_{i,j}$ is the given significance level for individual edge i, j test,
- $r_{i,j}$-sample correlation.

Pearson correlation network. Multiple testing procedures

- Single step rules

$$\partial^P(x) = \begin{pmatrix} 0, & \varphi^P_{1,2}(x), & \dots, & \varphi^P_{1,N}(x) \\ \varphi^P_{2,1}(x), & 0, & \dots, & \varphi^P_{2,N}(x) \\ \dots & \dots & \dots & \dots \\ \varphi^P_{N,1}(x), & \varphi^P_{N,2}(x), & \dots, & 0 \end{pmatrix}.$$

- Stepwise rules (Holm, Hochberg procedures with individual tests $\varphi^P_{i,j}$) with p-values of individual tests.

Sign similarity network

$$\gamma_{i,j}^{Sg} = p^{i,j} = P((X_i - E(X_i))(X_j - E(X_j) > 0)$$

- Individual hypotheses: $h_{ij} : p^{i,j} \leq p_0$ vs $k_{ij} : p^{i,j} > p_0$
- $I_{i,j}(t) = \begin{cases} 1, & sign(x_i(t) - \mu_i) = sign(x_j(t) - \mu_j) \\ 0, & \text{else} \end{cases}$
- Define $T_{i,j}^{Sg} = \sum_{t=1}^{n} I_{i,j}(t)$,
- $\varphi_{i,j}^{Sg} = \begin{cases} 0, & T_{i,j}^{Sg} \leq c_{i,j} \\ 1, & T_{i,j}^{Sg} > c_{i,j} \end{cases}$,
  where $c_{i,j}$ is defined from equation: $\sum_{k=c_{i,j}}^{n} \frac{n!}{k!(n-k)!}(p_0)^k(1-p_0)^{n-k} \leq \alpha$

Sign similarity network

- Multiple decision single step (Bonferroni type) procedure

$$\delta^{Sg}(x) = \begin{pmatrix} 0, & \varphi^{Sg}_{1,2}(x), & \dots, & \varphi^{Sg}_{1,N}(x) \\ \varphi^{Sg}_{2,1}(x), & 0, & \dots, & \varphi^{Sg}_{2,N}(x) \\ \dots & \dots & \dots & \dots \\ \varphi^{Sg}_{N,1}(x), & \varphi^{Sg}_{N,2}(x), & \dots, & 0 \end{pmatrix}.$$

- Holm, Hochberg procedures with the use of statistics $T^{Sg}_{i,j}$

Statistical procedure with invariant risk function

Theorem 1.1 Let random vector $X - \mathcal{K}(\Lambda)$.Then

1. single step statistical procedure δ^{Sg} for TG identification
2. Holm statistical procedure δ_H^{Sg} for TG identification
3. Hochberg statistical procedure δ_{Hg}^{Sg} for TG identification
4. Kruskal procedure δ^{Sg} for MST identification

have invariant risk function i.e. risk of these procedures does not depend from g.

Proof.

a

□

[a]Kalyagin V. A., Koldanov A. P., Petr A. Koldanov. Robust identification in random variables networks // Journal of Statistical Planning and Inference. 2017. Vol. 181, P. 30-40.

Statistical procedure with invariant risk function

Lemma 1.2 Let random vector $(X_1, \ldots, X_N) - EC(\mu, \Lambda, g)$. Then the probabilities

$$p(i_1, \ldots, i_N) := P_\Lambda(i_1(X_1 - \mu_1) > 0, \ldots, i_N(X_N - \mu_N) > 0)$$

are defined by the matrix Λ and does not depend on the function g for any $i_k \in \{-1, 1\},\ k = 1, 2, \ldots, N$.

Lemma 1.3 Let random vector $(X_1, \ldots, X_N) - EC(\mu, \Lambda, g)$.Then joint distribution of the statistics $T_{i,j}^{Sg}\ (i, j = 1, 2, \ldots, N; i \neq j)$ are defined by the matrix Λ and does not depend on the function g.

Comparison

1. Statistical procedures, based on sample Pearson correlations, have not invariant risk function in the class $EC(\mu, \Lambda, g)$ with fixed Λ.
2. Procedures, based on sample signs correlations, have invariant risk function in the class $EC(\mu, \Lambda, g)$ with fixed Λ.

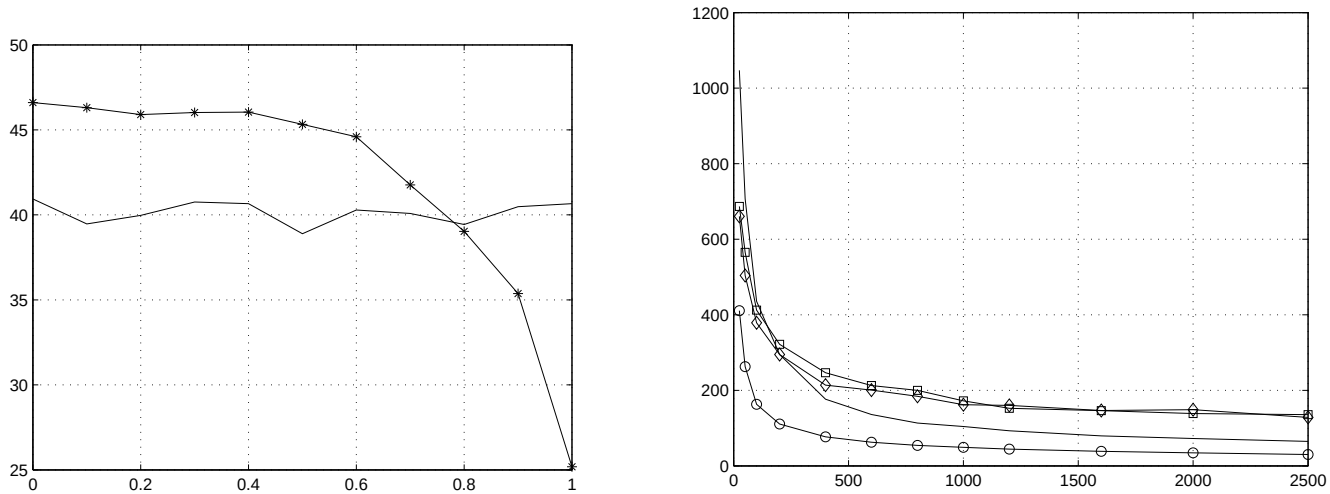

Figure: 2: Risk function for MG, $\rho_0 = 0.64$. Left - $n = 400$, star line - δ^P, line - δ^S, right: circle - $\gamma = 1$, δ^P; diamond - $\gamma = 0,5$, δ^P; square - $\gamma = 0$, δ^P, line - δ^S.The model is the mixture distribution consisting of multivariate normal distribution and multivariate Student distribution with 3 degree of freedom.

Advantages of sign similarity network

Sign similarity network:

- easy to interpretation;
- statistical procedures in sign similarity network have invariant risk function;
- statistical procedures in sign similarity network can be applied for network structure identification in other network models.

Comparison of Pearson correlations and sign similarity networks

If $X - EC(\mu, \Lambda, g)$ then network structures in Pearson correlation network are connected to network structures in sign similarity network.

Theorem 1.2 Let vector $X = (X_1, \ldots, X_N) - EC(\mu, \Lambda, g)$. Then:

$$\gamma_{i,j}^{Sg} = \frac{1}{2} + \frac{1}{\pi} \arcsin \frac{\lambda_{i,j}}{\sqrt{\lambda_{i,i}\lambda_{j,j}}} = \frac{1}{2} + \frac{1}{\pi} \arcsin \gamma_{i,j}^{P} \tag{2}$$

Corollary 1: MST in network model generated by Pearson correlation network coincide with MST in network model generated by sign similarity network.

Corollary 2: TG with threshold ρ_0 in network model generated by Pearson correlation network coincide with TG with threshold $p_0 = \frac{1}{2} + \frac{1}{\pi} \arcsin(\rho_0)$ in network model generated by sign similarity network.

$\gamma_{i,j}^{P} = 0.1 \Leftrightarrow \gamma_{i,j}^{Sg} = 0.53; \gamma_{i,j}^{P} = 0.6 \Leftrightarrow \gamma_{i,j}^{Sg} = 0.705$

Part 2. Optimal statistical procedure.

Optimal statistical procedure.

- **Definition:** Statistical procedure δ is optimal if $R(S, \theta, \delta) \leq R(S, \theta, \delta'), \forall S, \forall \theta \in \Omega_S, \forall \delta' \in \mathcal{D}$.
- Restrict attention to W-unbiased statistical procedures

$$E_\theta w(\theta, \delta) \leq E_\theta w(\theta', \delta), \forall \theta, \theta' \in \Omega$$

$$R(S, \theta, \delta) \leq R(S', \theta, \delta), \forall S, S', \theta \in \Omega_S$$

Gaussian Graphical Model selection

- Gaussian Graphical Model identification problem. Identify the concentration graph. The method can be applied to other problem.
- Concentration graph - edge (i, j) is included in the concentration graph if random variables X_i and X_j are conditionally dependent.
- Model selection: identify concentration graph by observations.

Random variables network

Random variables network settings.

- (X, γ).
- $X = (X_1, \ldots, X_N) - N(\mu, \Sigma)$
- Measure of association - partial correlation $\gamma_{i,j} = \rho^{i,j}$

Multiple decision approach

Let $x_i(t), i = 1, \ldots, N, t = 1, \ldots, n$ be a sample from multivariate normal distribution.
Let $G \in \mathcal{G}$-adjacency matrix, $\mathcal{G}$-set of all adjacency matrices.

$$H_G : \rho^{ij} = 0 \text{ if } g_{i,j} = 0, \rho^{ij} \neq 0 \text{ if } g_{i,j} = 1$$

Problem: construct optimal in the class of unbiased multiple decision statistical procedure to select one from disjoint hypothesis H_G

Individual hypotheses

Individual hypotheses

$$h_{i,j} : \rho^{i,j} = 0 \text{ vs } k_{i,j} : \rho^{i,j} \neq 0$$

According to Lauritzen S.L.[1]

$$\rho^{i,j} = \frac{-\sigma^{i,j}}{\sqrt{\sigma^{i,i}\sigma^{j,j}}}$$

Then

$$h_{i,j} : \sigma^{i,j} = 0 \text{ vs } k_{i,j} : \sigma^{i,j} \neq 0$$

[1] Lauritzen S.L.(1996) Graphical model. Oxford university press.

Statistical procedure

Let $\varphi_{i,j}(x)$ tests of individual hypotheses.
Define

$$\Phi(x) = \begin{pmatrix} 0, & \varphi_{1,2} & \dots & \varphi_{1,N} \\ \varphi_{1,2} & 0, & \dots & \varphi_{2,N} \\ \dots & \dots & \dots & \dots \\ \varphi_{1,N} & \varphi_{2,N} & \dots & 0 \end{pmatrix}$$

Define $\delta(x) = d_G$ if $\Phi(x) = G$

Existing statistical procedures. Single step procedure.[3]

Test of individual edge inclusion is

$$\varphi_{ij}(x) = \begin{cases} 1, & |z^{ij}| > c_{ij} \\ 0, & |z^{ij}| \leq c_{ij} \end{cases}$$

where $z^{ij} = \frac{1}{2} \ln\left(\frac{1+r^{ij}}{1-r^{ij}}\right)$, $r^{ij} = \frac{-s^{ij}}{\sqrt{s^{ii}s^{jj}}}$-sample partial correlation,
s^{ij}-elements of matrix S^{-1}.
c_{ij} from [2] $P_{\rho^{ij}=0}(|z^{ij}| > c_{ij}) = \alpha$
Properties of the associated multiple decision statistical procedure were not investigated.

[2] Anderson T.W.(2003) An introduction to multivariate statistical analysis.3-d edition. Wiley-Interscience, New York

[3] Edwards, D.M.(2000) Introduction to Graphical Modeling. New York, Springer.

Stepdown procedure.[4]

Let p_k are p-values of tests $\varphi_{ij}(x)$ $k = 1, \ldots, \frac{N(N-1)}{2}$. Order $p_{(1)} \leq p_{(2)} \leq \ldots \leq p_{(M-1)} \leq p_{(M)}$ and let $h_{(1)}, h_{(2)}, \ldots, h_{(M)}$ be the corresponding hypotheses.

- Step 1: If $p_{(1)} \geq \frac{\alpha}{M}$ then the decision is : accept all hypotheses $h_{(i)}$, $i = 1, 2, \ldots, M$ and stop, else reject hypothesis $h_{(1)}$ and go to the step 2.
- Step 2: If $p_{(2)} \geq \frac{\alpha}{M-1}$ then the decision is : accept all hypotheses $h_{(i)}$, $i = 2, \ldots, M$ and stop, else reject hypothesis $h_{(2)}$ and go to the step 3.
- ...
- Step M: If $|p_M| \geq \alpha$ then the decision is: accept hypothesis $h_{(M)}$ else reject all hypotheses.

Properties - control of FWER.

Type II error are not under control.

[4] M. Drton, M.D. Perlman.(2007) Multiple testing and error control in Gaussian graphical model selection. Statistical Science, 22,3, 430-449.

Our approach. Additive loss function

$l'_{i,j}$-loss from false inclusion of edge (i,j)
$l''_{i,j}$-loss from false non inclusion of the edge (i,j) $i,j = 1,2,\ldots,N;\ i \neq j$.
Loss function $w(S,Q)$ is *additive*[5] if:

$$w(S,Q) = \sum_{(i,j):s_{i,j}=0,q_{i,j}=1} l'_{i,j} + \sum_{(i,j):s_{i,j}=1,q_{i,j}=0} l''_{i,j} \qquad (3)$$

Theorem 2.1[6] Let the loss function w be defined by (3), and $l'_{i,j} = l'$, $l''_{i,j} = l''$, $i \neq j$, $i,j = 1,2,\ldots,N$. Then

$$Risk(S,\theta;\delta) = \sum_{i,j} r(s_{i,j},\varphi_{i,j}) = l' E_\theta[Y_I(S,\delta)] + l'' E_\theta[Y_{II}(S,\delta)], \quad \theta \in \Omega_S$$

where $Y_I(S,\delta)$, $Y_{II}(S,\delta)$ are the numbers of Type I and Type II errors by δ when the true decision is d_S.

[5]E.L.Lehmann (1957) A theory of some multiple decision problems.I Ann.Math.Stat.,28,1-25, 547-572.

[6]V.A. Kalyagin, A.P. Koldanov,P.A. Koldanov, P.M. Pardalos. Optimal statistical decision for Gaussian graphical model selection.arXiv:1701.02071v1

UMPU test for individual hypotheses

Theorem 2.2 Optimal in the class of unbiased statistical level α test for individual edge inclusion $h_{ij} : \rho^{i,j} = 0$ against $k_{ij} : \rho^{i,j} \neq 0$ is:

$$\varphi_{ij}^{opt} = \begin{cases} 0, & \frac{|as_{ij} - \frac{b}{2}|}{\sqrt{\frac{b^2}{4} + ac}} < 1 - 2c_{\alpha}^{beta} \\ 1, & \frac{|as_{ij} - \frac{b}{2}|}{\sqrt{\frac{b^2}{4} + ac}} > 1 - 2c_{\alpha}^{beta} \end{cases} \tag{4}$$

where $\det(s_{kl}) = -as_{ij}^2 + bs_{ij} + c$, c_{α}^{beta} is the α-quantile of Beta distribution. $(a = a(\{s_{kl}\}), b = b(\{s_{kl}\}), c = c(\{s_{kl})\})$.
7

[7] Koldanov P., Koldanov A. P., Kalyagin V. A., Pardalos P. M. Uniformly most powerful unbiased test for conditional independence in Gaussian graphical model // Statistics & Probability Letters, 2017, Vol. 122, P. 90-95.

Wishart distribution

$$S = \begin{pmatrix} s_{11} & s_{12} & \dots & s_{1N} \\ s_{21} & s_{22} & \dots & s_{2N} \\ \dots & \dots & \dots & \dots \\ s_{N1} & s_{N2} & \dots & s_{NN} \end{pmatrix} \tag{5}$$

$$f(\{s_{k,l}\}) = \frac{[\det(\sigma^{kl})]^{n/2} \times [\det(s_{kl})]^{(n-N-2)/2} \times \exp[-(1/2)\sum_k \sum_l s_{k,l}\sigma^{kl}]}{2^{(Nn/2)} \times \pi^{N(N-1)/4} \times \Gamma(n/2)\Gamma((n-1)/2)\cdots\Gamma((n-N+1)/2)}$$

if the matrix (s_{kl}) is positive definite, and $f(\{s_{kl}\}) = 0$ otherwise. Let I be the interval of positive definiteness of the matrix. One has for a fixed $i < j$:

$$f(\{s_{kl}\}) = C(\{\sigma^{kl}\}) \times \exp[-\sigma^{ij}s_{ij} - \frac{1}{2} \sum_{(k,l)\neq(i,j);(k,l)\neq(j,i)} s_{kl}\sigma^{kl}] \times h(\{s_{kl}\})$$

UMPU test

UMPU test for testing hypothesis

$$h_{ij} : \rho^{i,j} = 0 \text{ vs } k_{ij} : \rho^{i,j} \neq 0$$

has the Neyman structure and can be written as

$$\delta_{i,j}(\{s_{kl}\}) = \begin{cases} \partial_{i,j}, & \text{if } c_1(\{s_{kl}\}) \leq s_{ij} \leq c_2(\{s_{kl}\}),\ (k,l) \neq (i,j) \\ \partial_{i,j}^{-1}, & \text{if } s_{ij} < c_1(\{s_{kl}\})\ \ s_{ij} > c_2(\{s_{kl}\}),\ (k,l) \neq (i,j) \end{cases} \tag{6}$$

where constants are defined from

$$\frac{\int_{I \cap [c_1;c_2]} \exp[-\sigma_0^{ij} s_{ij}][\det(s_{kl})]^{(n-N-2)/2} ds_{ij}}{\int_I \exp[-\sigma_0^{ij} s_{ij}][\det(s_{kl})]^{(n-N-2)/2} ds_{ij}} = 1 - \alpha_{i,j}, \tag{7}$$

$$\begin{aligned} &\int_{I \cap [-\infty;c_1]} s_{ij} \exp[-\sigma_0^{ij} s_{ij}][\det(s_{kl})]^{(n-N-2)/2} ds_{ij} + \\ &+ \int_{I \cap [c_2;+\infty]} s_{ij} \exp[-\sigma_0^{ij} s_{ij}][\det(s_{kl})]^{(n-N-2)/2} ds_{ij} = \\ &= \alpha_{i,j} \int_I s_{ij} \exp[-\sigma_0^{ij} s_{ij}][\det(s_{kl})]^{(n-N-2)/2} ds_{ij}, \end{aligned} \tag{8}$$

UMPU test.

Under $\sigma_0^{i,j} = 0$ equation (7) is

$$\frac{\int_{I\cap[c_1;c_2]}[\det(s_{kl})]^{(n-N-2)/2}ds_{ij}}{\int_I[\det(s_{kl})]^{(n-N-2)/2}ds_{ij}} = 1-\alpha_{i,j} \tag{9}$$

Let $K = \frac{n-N-2}{2}, x = s_{ij}$. Then

$$\int_f^d (ax^2-bx-c)^K dx = (-1)^K a^K (x_2-x_1)^{2K+1} \int_{\frac{f-x_1}{x_2-x_1}}^{\frac{d-x_1}{x_2-x_1}} u^K(1-u)^K du$$

Equation (9) can be written as

$$\int_{\frac{c_1-x_1}{x_2-x_1}}^{\frac{c_2-x_1}{x_2-x_1}} u^K(1-u)^K du = (1-\alpha)\int_0^1 u^K(1-u)^K du = (1-\alpha)\frac{\Gamma(K+1)\Gamma(K+1)}{\Gamma(2K+2)} \tag{10}$$

Acceptance region is: $c_\alpha^{beta} \le \frac{s_{i,j}-x_1}{x_2-x_1} \le 1-c_\alpha^{beta}$ or

$2c_\alpha^{beta}-1 \le \frac{as_{i,j}-b/2}{\sqrt{b^2/4+ac}} \le 1-2c_\alpha^{beta}$

UMPU test is equivalent to partial correlation test

Sample partial correlation test for testing hypothesis $\rho^{i,j} = 0$:

$$\varphi_{i,j} = \begin{cases} 0, & |r^{i,j}| \leq c_{i,j} \\ 1, & |r^{i,j}| > c_{i,j} \end{cases} \tag{11}$$

where $c_{i,j}$ is $(1 - \alpha/2)$-quantile of the distribution with density function

$$f(x) = \frac{1}{\sqrt{\pi}} \frac{\Gamma(n - N + 1)/2)}{\Gamma((n - N)/2)} (1 - x^2)^{(n-N-2)/2}, \quad -1 \leq x \leq 1$$

Theorem 2.3 Sample partial correlation test (11) is equivalent to UMPU test (4) for testing hypothesis $\rho^{i,j} = 0$ vs $\rho^{i,j} \neq 0$.

Equivalence of partial correlation and UMPU tests.

It is sufficient to prove that

$$\frac{S^{i,j}}{\sqrt{S^{i,i}S^{j,j}}} = \frac{as_{i,j} - \frac{b}{2}}{\sqrt{\frac{b^2}{4} + ac}} \tag{12}$$

Let $A = (a_{k,l})$ be an $(N \times N)$ symmetric matrix. Fix $i < j$, $i, j = 1, 2, \ldots, N$. Denote by $A(x)$ the matrix obtained from A by replacing the elements $a_{i,j}$ and $a_{j,i}$ by x. Denote by $A^{i,j}(x)$ the cofactor of the element (i, j) in the matrix $A(x)$. Then the following statement is true **Lemma 2.1** One has $[\det A(x)]' = -2A^{i,j}(x)$.

Equivalence of partial correlation and UMPU tests.

$$\det(S(x)) = -ax^2 + bx + c \to [\det S(x)]' = -2ax + b = -2S^{i,j}(x)$$

i.e. $S^{i,j}(x) = ax - b/2$.

$$x = s_{i,j} \to as_{i,j} - \frac{b}{2} = S^{i,j}$$

It is sufficient to prove that $\sqrt{S^{i,i}S^{j,j}} = \sqrt{\frac{b^2}{4} + ac}$.

Let $x_2 = \frac{b+\sqrt{b^2+4ac}}{2a}$ be the maximum root of equation $ax^2 - bx - c = 0$.

Then $ax_2 - \frac{b}{2} = \sqrt{\frac{b^2}{4} + ac}$.

Equivalence of partial correlation and UMPU tests.

Consider

$$r^{i,j}(x) = \frac{-S^{i,j}(x)}{\sqrt{S^{i,i}S^{j,j}}}$$

According to Silvester determinant identity:

$$S^{\{i,j\},\{i,j\}} \det S(x) = S^{i,i}S^{j,j} - [S^{i,j}(x)]^2$$

Therefore for $x = x_1$ and $x = x_2$ one has

$$S^{i,i}S^{j,j} - [S^{i,j}(x)]^2 = 0$$

For $x = x_1$ and $x = x_2$ one has $r^{i,j}(x) = \pm 1$. The equation $S^{i,j}(x) = ax - \frac{b}{2}$ implies that when x is increasing from x_1 to x_2 then $r^{i,j}(x)$ is decreasing from 1 to -1. That is $r^{i,j}(x_2) = -1$, i.e. $ax_2 - \frac{b}{2} = \sqrt{S^{i,i}S^{j,j}}$. Therefore

$$\sqrt{S^{i,i}S^{j,j}} = \sqrt{\frac{b^2}{4} + ac}$$

Multiple decision statistical procedure

$$\Phi^{opt}(x) = \begin{pmatrix} 0, & \varphi_{12}^{opt}(x), & \dots, & \varphi_{1N}^{opt}(x) \\ \varphi_{21}^{opt}(x), & 0, & \dots, & \varphi_{2N}^{opt}(x) \\ \dots & \dots & \dots & \dots \\ \varphi_{N1}^{opt}(x), & \varphi_{N2}^{opt}(x), & \dots, & 0 \end{pmatrix}. \quad (13)$$

Define multiple statistical procedure for concentration graph identification

$$\delta^{opt}(x) = d_G, \text{ iff } \Phi^{opt}(x) = G \quad (14)$$

Theorem 2.4 Let the loss function w be defined by (3) and

$$\alpha_{i,j} = \frac{I''_{i,j}}{I'_{i,j} + I''_{i,j}}, \ i \neq j, \ i,j = 1,2,\dots,p. \quad (15)$$

Then the procedure δ^{opt} is optimal multiple decision statistical procedure for Gaussian graphical model selection in the class of w-unbiased procedures. [8]

[8] V.A. Kalyagin, A.P. Koldanov, P.A. Koldanov, P.M.Pardalos Optimal statistical decision for Gaussian graphical model selection.

Conclusions of part 2

- UMPU test for testing hypothesis $h^{i,j} : \rho^{i,j} = 0$ vs $k_{i,j} : \rho^{i,j} \neq 0$ is constructed.
- Statistical procedure $\Phi^{opt}(x)$ is optimal unbiased under additive loss function.
- Existing multiple single step procedure is asymptotically optimal for additive loss function.
- Multiple stepdown procedure is not unbiased for additive loss function but control probability of at least one incorrectly included edge and designed to another loss function.

Resume

- 1. The general approach to network structures identification in the framework of random variable network is proposed:
  - concept of random variable network;
  - problems of network structures identification as multiple decision problems;
  - statistical uncertainty - risk function;
  - linear combination of expectations of type I and type II errors .
- 2. Statistical procedures for network structures identification with invariant risk function in the class of elliptically contoured distributions are constructed:
  - the probability of sign coincidence;
  - concept of sign similarity network;
  - sign identification statistical procedures;
  - network structures (TG, MST) identification statistical procedures have invariant risk function;
  - pearson correlation network;
  - conditions of optimality.

Resume

- 3. Optimal statistical procedure for GGM identification is constructed.
  - UMPU Neyman structure test.
  - UMPU test and ML test.
  - Additive loss function.
  - Unbiased multiple decision procedure.
  - Optimal unbiased multiple decision procedure for GGM.
- 4. Properties of standard procedures are investigated:
  - Threshold graph identification problem
  - Pearson sample correlation
  - Invariant procedures
  - Additive loss
  - Restricted class of procedures
  - Optimality

Resume

- 5. Properties of statistical procedure for nodes selection are investigated.
  - UMPU Neyman structure test for expectations of components of random vector with multivariate normal distribution.
  - Additive loss function.
  - Unbiased multiple procedure.
  - Optimal unbiased multiple decision procedure.
  - Risk function does not depend from correlation matrix.
- 6. Application to market network analysis
  - Network structures comparison.
  - Three-steps procedure for portfolio selection.
  - Testing of the symmetry conditions.

Our publications on market network analysis.

- Koldanov P., Kalyagin V. A., Bautin G. A. On some statistical procedures for stock selection problem // Annals of Mathematics and Artificial Intelligence. 2016. Vol. 76. No. 1. P. 47-57.
- Kalyagin V. A., Koldanov A. P., Petr A. Koldanov. Robust identification in random variables networks // Journal of Statistical Planning and Inference. 2017. Vol. 181, P. 30-40.
- Koldanov P., Koldanov A. P., Kalyagin V. A., Pardalos P. M. Uniformly most powerful unbiased test for conditional independence in Gaussian graphical model // Statistics & Probability Letters, 2017, Vol. 122, P. 90-95.
- Koldanov A.P., Koldanov P.A., Kalyagin V.A., Pardalos P.M. Statistical Procedures for the Market Graph Construction, Computational Statistics & Data Analysis, v.68, pp.17-29 (2013).
- Kalyagin V.A., Koldanov A.P., Koldanov P.A., Pardalos P.M., Zamaraev V.A.: Measures of uncertainty in market network analysis, Physica A. Vol 413 (2014), pp.59-70.

Our publications on market network analysis.

- V.A. Kalyagin, A.P. Koldanov,P.A. Koldanov, P.M. Pardalos. Optimal statistical decision for Gaussian graphical model selection. arXiv: 1701.02071v1.
- Bautin G.A., Kalyagin V.A., Koldanov A.P., Koldanov P.A., Pardalos P.M. Simple measure of similarity for the market graph construction, Computational Management Science, Volume 10, Issue 2 (2013), Page 105-124.
- V. A. Kalyagin, A. P. Koldanov, P. A. Koldanov, P. M. Pardalos Optimal decision for the market graph identification problem in a sign similarity network. Ann. Oper. Res, DOI 10.1007/s10479-017-2491-6
- Koldanov P., Lozgacheva N. Multiple testing of sign symmetry for stock return distributions // International Journal of Theoretical and Applied Finance. 2016. Vol. 19. No. 8. P. 1650049-1-1650049-14.
- Koldanov P.A. Bautin G.A. Multiple decision problem for stock selection in market network, Vol. 8426: Lectute Notes in Computer Science. Switzerland : Springer, 2014. P. 98-110.

Our publications on market network analysis.

- Vizgunov A.N., Goldengorin B.I., Kalyagin V.A., Koldanov A.P., Koldanov P.A., Pardalos P. M., Network approach for the Russian stock market, Computational Management Science, Springer-Verlag, Volume 11, Issue 1 (2014), Page 45-55.
- Koldanov A. P., Koldanov P.A: Optimal Multiple Decision Statistical Procedure for Inverse Covariance Matrix. Springer Optimization and Its Applications. Vol. 87 (2014), pp 205-216.
- Kalyagin V.A., Koldanov P.A., Zamaraev V.A. Network structures uncertainty for different markets, Springer Optimization and Its Applications. 2014. Vol. 100. P. 181-197.
- Latyshev A., Koldanov P. Testing of connections between Pearson and sign correlations in market network. in: Models, Algorithms, and Technologies for Network Analysis. Springer Proceedings in Mathematics and Statistics, 2016. P. 175-182.

Our publications on market network analysis.

- Kalyagin V. Koldanov A. Koldanov P. Zamaraev V. Market Graph and Markovitz Model, Optimization in Science and Engineering, 2014, pp 293-306
- Koldanov P., Komissarova A. Statistical uncertainty of minimum spanning tree in market network, in: Models, Algorithms, and Technologies for Network Analysis. Springer Proceedings in Mathematics and Statistics, 2016. P. 157-164.
- Koldanov P.A. Efficiency Analysis of Branch Network, in: Models, Algorithms, and Technologies for Network Analysis. Springer Proceedings in Mathematics and Statistics, 2013. 71-84.
- Koldanov P., Grechikhin I. How independent are stocks in an independent set of a market graph, in: Models, Algorithms and Technologies for Network Analysis. Springer Proceedings in Mathematics and Statistics, 2014, V.104., P. 45-53.
- Koldanov P., Pardalos P. M., Zamaraev V. A. Statistical Uncertainty of Market Network Structures, in: DATA ANALYTICS 2014, The Third International Conference on Data Analytics, 2014. P. 91-94.

THANK YOU FOR YOUR ATTENTION!

Optimality of ∂^P

Theorem

Let loss function w be additive, individual test statistics $t_{i,j}$ depends only on observations $X_i(t), X_j(t)$ and vector $X = (X_1, \ldots, X_N)$ has a multivariate normal distribution. Then for single step statistical procedure ∂^P for threshold graph identification ($\rho_0 = 0$) in Pearson correlation network one has $Risk(S, \partial^P) \leq Risk(S, \partial)$ for any adjacency matrix S and any $w-$unbiased δ.

Theorem

Let loss function w be additive, individual test statistics $t_{i,j}$ depends only on observations $X_i(t), X_j(t)$ and vector $X = (X_1, \ldots, X_N)$ has a multivariate normal distribution. Then for single step statistical procedure ∂^P for threshold graph identification in Pearson correlation network one has $Risk(S, \partial^P) \leq Risk(S, \delta)$ for any adjacency matrix S and any invariant δ.

Assumption of normality can not be removed.

Optimality of δ^{Sg}

Theorem

Let loss function w be additive, individual test statistics $t_{i,j}$ depends only on $I_{i,j}(t)$, $E(X_i)$ are known $\forall i = 1, \ldots, N$ and distribution of vector $X = (X_1, \ldots, X_N)$ satisfy the symmetry condition below.
Then for single step statistical procedure δ^{Sg} for threshold graph identification in sign similarity network one has $Risk(S, \delta^{Sg}) \leq Risk(S, \delta)$ for any adjacency matrix S and any $w-$unbiased δ.

Symmetry condition:

$$p_{1,1}^{i,j} = p_{-1,-1}^{i,j}; p_{1,-1}^{i,j} = p_{-1,1}^{i,j}$$

$$p_{1,1}^{i,j} = P(X_i - E(X_i) > 0, X_j - E(X_j) > 0)$$

$$p_{-1,-1}^{i,j} = P(X_i - E(X_i) \leq 0, X_j - E(X_j) \leq 0)$$

$$p_{-1,1}^{i,j} = P(X_i - E(X_i) \leq 0, X_j - E(X_j) > 0)$$

$$p_{1,-1}^{i,j} = P(X_i - E(X_i) > 0, X_j - E(X_j) \leq 0)$$

Proof of lemma 2

Lemma 2. Probabilities
$p(i_1, i_2, \ldots, i_N) := P_\Lambda(i_1 X_1 > 0, i_2 X_2, \ldots, i_N X_N > 0)$ are defined by the matrix Λ and does not depend on the function g.

Proof

$$P(i_1 X_1 > 0, i_2 X_2, \ldots, i_N X_N > 0) = \int_{i_k x_k > 0, k=1,2,\ldots,N} |\Lambda|^{-\frac{1}{2}} g(x'\Lambda x) dx_1 \ldots dx_N \tag{16}$$

Matrix Λ is positive definite, therefore there exists a matrix C such that $C'\Lambda C = I$. Put $y = C^{-1}x$. Then $x = Cy$ and

$$\int_{i_k x_k > 0, k=1,2,\ldots,N} |\Lambda|^{-\frac{1}{2}} g(x'\Lambda x) dx_1 \ldots dx_N = \int_D g(y'y) dy_1 \ldots dy_N \tag{17}$$

where D is given by

$$0 < i_k(c_{k,1} y_1 + c_{k,2} y_2 + \ldots + c_{k,N} Y_N) < \infty, \quad k = 1, 2, \ldots, N \tag{18}$$

Proof of lemma 2

Vector y can be written in polar coordinates as:

$$\begin{aligned} y_1 &= r\sin(\theta_1) \\ y_2 &= r\cos(\theta_1)\sin(\theta_2) \\ y_3 &= r\cos(\theta_1)\cos(\theta_2)\sin(\theta_3) \\ &\dots \\ y_{N-1} &= r\cos(\theta_1)\cos(\theta_2)\dots\cos(\theta_{N-2})\sin(\theta_{N-1}) \\ y_N &= r\cos(\theta_1)\cos(\theta_2)\dots\cos(\theta_{N-2})\cos(\theta_{N-1}) \end{aligned} \tag{19}$$

where $-\frac{\pi}{2} \le \theta_i \le \frac{\pi}{2}, i = 1\dots, N-2; -\pi \le \theta_{N-1} \le \pi, 0 \le r \le \infty$ The Jacobian of the transformation (19) is

$$r^{N-1}\cos^{N-2}(\theta_1)\cos^{N-3}(\theta_2)\dots\cos(\theta_{N-2})$$

In polar coordinates region (18) is transformed to the region $D' \times R_+^1$ where D' given by $(k = 1, 2, \dots, N)$:

$$0 < i_k(c_{11}\sin(\theta_1) + \dots + c_{1N}\cos(\theta_1)\cos(\theta_2)\dots\cos(\theta_{N-2})\cos(\theta_{N-1})) < \infty \tag{20}$$

Proof of lemma 2

Then $p(i_1, i_2, \ldots, i_N)$ can be written as

$$\int_{D'} \int_0^{\infty} r^{N-1} \cos^{N-2}(\theta_1) \cos^{N-3}(\theta_2) \ldots \cos(\theta_{N-2}) g(r^2) dr d\theta_1 \ldots d\theta_{N-1} =$$

$$= \int_{D'} \cos^{N-2}(\theta_1) \cos^{N-3}(\theta_2) \ldots \cos(\theta_{N-2}) d\theta_1 \ldots d\theta_{N-1} \int_0^{\infty} r^{N-1} g(r^2) dr$$

It is known that

$$\int_0^{\infty} r^{N-1} g(r^2) dr = \frac{1}{C(N)}$$

where

$$C(N) = \int_{-\frac{\pi}{2}}^{\frac{\pi}{2}} \ldots \int_{-\frac{\pi}{2}}^{\frac{\pi}{2}} \int_{-\pi}^{\pi} \cos^{N-2}(\theta_1) \cos^{N-3}(\theta_2) \ldots \cos(\theta_{N-2}) d\theta_1 \ldots d\theta_{N-1}$$

Region D' is defined by the matrix Λ and does not depend on the function g. Then $p(i_1, i_2, \ldots, i_N)$ are defined by the matrix Λ and does not depend on the function g.

Proof of lemma 3

Lemma 3. Joint distribution of the statistics $T_{i,j}^{Sg}$ are defined by the matrix Λ and does not depend on the function g.

Proof.

- $T_{i,j}^{Sg} = \frac{n}{2} + \frac{1}{2}\sum_{t=1}^{n} \text{sign}(X_i(t))\text{sign}(X_j(t))$
- from the lemma 2 - joint distribution of the $\text{sign}(X) = (\text{sign}(X_1), \text{sign}(X_2), \ldots, \text{sign}(X_N))$ is defined by the matrix Λ and does not depend on the function g.
- random vectors $\text{sign}(X(t))$, $t = 1, 2, \ldots, n$ are independent and identically distributed.
- then the joint distribution $\text{sign}(X_i(t))$, $i = 1, 2, \ldots, N$, $t = 1, 2, \ldots, n$ is defined by the matrix Λ and does not depend on the function g.
- then joint distribution of statistics $T_{i,j}^{Sg}$, $i, j = 1, 2, \ldots, N; i < j$ does not depend on the function g.

Data for stability

1. We consider the real-world data from USA stock market. We take $N = 83$ largest by capitalization companies and consider the daily returns of these companies for the period from 03.01.2011 up to 31.12.2013, total 751 observations.
2. We calculate correlation matrix Σ by this data and consider the matrix Σ as true matrix. Structures of the matrix are considered as true structures.
3. We simulate a certain number of observation (n) using the mixture distribution. The mixture distribution is constructed as follow - random vector $X = (X_1, \ldots, X_N)$ takes value from $N(0, \Sigma)$ with probability γ and from $t_3(0, \Sigma)$ with probability $1 - \gamma$.
4. We estimate the matrix Σ using the chosen association measure (Pearson $\rho_{i,j}$ or probability $p^{i,j}$).
5. We construct the sample threshold graph basing on the estimation and compare it to the true threshold graph.

Appendix: Proof of the Theorem 2

Theorem 2: Let vector $X = (X_1, \ldots, X_N)$ has elliptical distribution (1). Then:

$$\gamma_{i,j}^{Sg} = \frac{1}{2} + \frac{1}{\pi} \arcsin \frac{\lambda_{i,j}}{\sqrt{\lambda_{i,i}\lambda_{j,j}}} = \frac{1}{2} + \frac{1}{\pi} \arcsin \gamma_{i,j}^{P} \tag{21}$$

Prove: It is known $E(X) = \mu$. Without loss of generality let $\mu = 0$. Define matrix $A = (a_{i,j}) = \Lambda^{-1}$. Density of random vector (X_i, X_j) has the form:

$$f(x_i, x_j) = |A^{-1}|^{-\frac{1}{2}} g(a_{i,i}x_i^2 + 2a_{i,j}x_i(x_j + a_{j,j}x_j^2)$$

The prove is based on the following lemma:

Proof of the Theorem 2

Lemma 1: Probability $\gamma_{i,j}^{Sg} = P(X_i X_j > 0)$ defined by the matrix Λ and does not depend from g.

Prove: Matrix $A_{i,j} = \begin{pmatrix} a_{i,i} & a_{i,j} \\ a_{i,j} & a_{j,j} \end{pmatrix}$ is positive definite, then exists $C = \begin{pmatrix} c_{i,i} & c_{i,j} \\ c_{j,i} & c_{j,j} \end{pmatrix}$ such that $C'AC = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}$.

Define $U = c_{i,i}X_i + c_{i,j}X_j$, $V = c_{j,i}X_i + c_{j,j}X_j$. Then random vector (U, V) has distribution with density $f(u, v) = g(u^2 + v^2)$. Then

$$P(X_i > 0, X_j > 0) = P(\frac{c_{ii}}{c_{jj}}V < U < \frac{c_{ij}}{c_{jj}}V) + P(\frac{c_{ij}}{c_{jj}}V < U < \frac{c_{ii}}{c_{jj}}V) =$$

$$= \begin{cases} \frac{arctg(c_{ij}/c_{jj}) - arctg(c_{ii}/c_{jj})}{2\pi}, & \frac{c_{ij}}{c_{jj}} > \frac{c_{ii}}{c_{jj}} \\ \frac{arctg(c_{ii}/c_{jj}) - arctg(c_{ij}/c_{jj})}{2\pi}, & \frac{c_{ij}}{c_{jj}} < \frac{c_{ii}}{c_{jj}} \end{cases}$$

Then $P(X_i > 0, X_j > 0)$ does not depend from g. Similarly $P(X_i < 0, X_j < 0)$ does not depend from g. Then $P(X_i X_j > 0) = P(X_i > 0, X_j > 0) + P(X_i < 0, X_j < 0)$ does not depend from g.

Holm procedure

- Step 1: If $\max_{i,j=1,\ldots,N} T_{i,j} \leq c_1^H$ then accept all hypotheses $h_{i,j}, i,j = 1,2,\ldots,N$, else if $\max_{i,j=1,\ldots,N} T_{i,j} = T_{i_1,j_1}$ then reject hypothesis h_{i_1,j_1} and go to step 2.
- ...
- Step K: Let $I = \{(i_1,j_1),(i_2,j_2),\ldots,(i_{K-1},j_{K-1})\}$ be the set of indexes of previously rejected hypotheses. If $\max_{(i,j)\notin I} T_{i,j} \leq c_K^H$ then accept all hypotheses $h_{i,j}$, $(i,j) \notin I$, else if $\max_{(i,j)\notin I} T_{i,j} = T_{i_K,j_K}$ then reject hypothesis h_{i_K,j_K} and go to step (K+1).
- ...
- Step M: Let $I = \{(i_1,j_1),\ldots,(i_{M-1},j_{M-1})\}$ be the set of indexes of previously rejected hypotheses. Let $(i_M,j_M) \notin I$. If $T_{i_M,j_M} \leq c_M^H$ then accept the hypothesis h_{i_M,j_M}, else reject hypothesis h_{i_M,j_M} (reject all hypotheses).

For a given significance level α the critical values c_K^H for Holm procedure are given by $F_{\gamma_0}(c_K^H) = 1 - \dfrac{\alpha}{M-K+1}, \quad K = 1,2,\ldots,M$

Hochberg procedure

- Step 1: If $T_{i_1,j_1} = \min_{i,j=1,\dots,N} T_{i,j} > c_1^{Hg}$ then reject all individual hypotheses $h_{i,j}$, else accept hypothesis h_{i_1,j_1} and go to step 2.
- ...
- Step K: Let $I = \{(i_1,j_1),\dots,(i_{K-1},j_{K-1})\}$ be the set of indexes of previously accepted hypotheses. If $T_{i_K,j_K} = \min_{i,j=1,\dots,N;(i,j)\notin I} T_i(x) > c_K^{Hg}$ then reject all hypotheses $h_{i,j}$, $(i,j) \notin I$, else accept hypothesis h_{i_K,j_K} and go to step (K+1).
- ...
- Step M: Let $I = \{(i_1,j_1),\dots,(i_{M-1},j_{M-1})\}$ be the set of indexes of previously accepted hypotheses. Let $(i_M,j_M) \notin I$. If $T_{i_M,j_M} > c_M^{Hg}$ then reject the hypothesis h_{i_M,j_M} else accept the hypothesis h_{i_M,j_M} (accept all hypothesis).

For a given significance level α the critical values c_K^{Hg} for Hochberg procedure are given by $F_{\gamma_0}(c_K^{Hg}) = 1 - \frac{\alpha}{K}, \quad K = 1,2,\dots,M$

Role of measure of association

Therefore the following statistical procedure for threshold graph identification in Pearson correlation network will be distribution free:

- fix a threshold ρ_0.
- Take δ a threshold graph identification statistical procedure in sign similarity network distribution free in the class of elliptically contoured distributions.
- Apply statistical procedures δ for threshold graph identification with the threshold

$$p_0 = \frac{1}{2} + \frac{1}{\pi} \arcsin \rho_0$$

.

- Consider obtained graph as the threshold graph in Pearson correlation network.

In particular one can construct single step, Holm and Hochberg distribution free statistical procedures for threshold graph identification in Pearson correlation network.

Symmetry conditions. Tests for individual hypotheses

Individual hypotheses:
$h_1^{i,j} : p_{1,1}^{i,j} = p_{-1,-1}^{i,j}$ vs $k_1^{i,j} : p_{1,1}^{i,j} \neq p_{-1,-1}^{i,j}$; $i, j = 1, \ldots, N$; $i \neq j$
Statistics $T_{1,1}^{i,j} = \sum_{t=0}^{n} T_{1,1}^{i,j}(t)$, $T_{-1,-1}^{i,j} = \sum_{t=0}^{n} T_{-1,-1}^{i,j}(t)$,

$$T_{1,1}^{i,j}(t) = \begin{cases} 1, & X_i(t) \geq 0, X_i(t) \geq 0 \\ 0, & else \end{cases}$$

$$T_{-1,-1}^{i,j}(t) = \begin{cases} 1, & X_i(t) < 0, X_i(t) < 0 \\ 0, & else \end{cases}$$

Symmetry conditions. Tests for individual hypotheses

Individual hypotheses:
$h_2^{i,j} : p_{1,-1}^{i,j} = p_{-1,1}^{i,j}$ vs $k_2^{i,j} : p_{1,-1}^{i,j} \neq p_{-1,1}^{i,j};\ i,j = 1,\dots,N;\ i \neq j$
Statistics $T_{1,-1}^{i,j} = \sum_{t=0}^{n} T_{1,-1}^{i,j}(t), \quad T_{-1,1}^{i,j} = \sum_{t=0}^{n} T_{-1,1}^{i,j}(t)$

$$T_{1,-1}^{i,j}(t) = \begin{cases} 1, & X_i(t) \geq 0, X_i(t) < 0 \\ 0, & else \end{cases}$$

$$T_{-1,1}^{ij}(t) = \begin{cases} 1, & X_i(t) < 0, X_i(t) \geq 0 \\ 0, & else \end{cases}$$

Optimal tests for individual hypotheses testing $h^1_{i,j}$

Exponential form for the joint distribution of statistics $T_{k,l}$:

$$P(T_{1,1} = k_1, T_{-1,-1} = k_2, T_{1,-1} = k_3, T_{-1,1} = k_4) =$$

$$= C \exp\{k_1 \ln \frac{p_{1,1}}{p_{-1,-1}} + (k_1 + k_2) \ln \frac{p_{-1,-1}}{p_{-1,1}} + k_3 \ln \frac{p_{1,-1}}{p_{-1,1}}\}$$

where

$$C = \frac{n!}{k_1!k_2!k_3!k_4!}(1 - p_{1,1} - p_{-1,-1} - p_{1,-1})^n$$

Then UMPU test for testing hypothesis $h^1_{i,j}$ has Neymann structure and can be written as:

$$\varphi^1_{i,j} = \begin{cases} 0, & C_1(k, k_3) < k_1 < C_2(k, k_3) \\ 1, & else \end{cases} \tag{22}$$

where k_1, k_2, k_3, k_4 are the observed values of statistics $T^{i,j}_{1,1}$, $T^{i,j}_{-1,-1}$, $T^{i,j}_{1,-1}$, $T^{i,j}_{-1,1}$, $k = k_1 + k_2$.

Optimal tests for individual hypotheses testing $h^1_{i,j}$

The constants C_1, C_2 are defined from conditional distribution of statistic $T_{1,1}$ under conditions $T_{1,1} + T_{-1,-1} = k$, $T_{1,-1} = k_3$ and assumption that the hypothesis $h^1_{i,j}$ is true. One has

$$P(T_{1,1} = k_1 | T_{1,1} + T_{-1,-1} = k, T_{1,-1} = k_3) =$$

$$= \frac{P(T_{1,1} = k_1, T_{-1,-1} = k - k_1, T_{1,-1} = k_3)}{P(T_{1,1} + T_{-1,-1} = k, T_{1,-1} = k_3)}$$

$$P(T_{1,1}+T_{-1,-1} = k, T_{1,-1} = k_3) = \sum_{i=0}^{k} P(T_{1,1} = i, T_{-1,-1} = k-i, T_{1,-1} = k$$

$$= \frac{n!}{k_3!(n-k_3-k)!k!} p_{1,-1}^{k_3} (p_{1,1}+p_{-1,-1})^k (1-p_{1,1}-p_{-1,-1}-p_{1,-1})^{n-k_3-k}$$

$$P(T_{1,1} = k_1, T_{-1,-1} = k - k_1, T_{1,-1} = k_3) =$$

$$= \frac{n!}{k_1!k_3!(k-k_1)!(n-k_3-k)!} p_{1,1}^{k_1} p_{1,-1}^{k_3} p_{-1,-1}^{k-k_1} (1-p_{1,1}-p_{-1,-1}-p_{1,-1})^{n-k_3-k}$$

Optimal tests for individual hypotheses testing $h^1_{i,j}$

$$P(T_{1,1} = k_1 | T_{1,1} + T_{-1,-1} = k, T_{1,-1} = k_3) =$$

$$= C_k^{k_1} \left(\frac{p_{1,1}}{p_{1,1} + p_{-1,-1}} \right)^{k_1} \left(\frac{p_{-1,-1}}{p_{1,1} + p_{-1,-1}} \right)^{k-k_1}$$

Under $h^1_{i,j}$ one has $p_{1,1} = p_{-1,-1}$. Optimal test is

$$\varphi^1_{i,j} = \begin{cases} 0, & C_1(k) < k_1 < C_2(k) \\ 1, & else \end{cases} \tag{23}$$

where $C_1(k)$ and $C_2(k)$ are defined by

$$C_1(k) = \max\{C : (\frac{1}{2})^k \sum_{i=0}^{C} C_k^i \leq \frac{\alpha}{2}\}$$

$$C_2(k) = \min\{C : (\frac{1}{2})^k \sum_{i=C}^{k} C_k^i \leq \frac{\alpha}{2}\}$$

p-values

The p-value of the test can be calculated by

$$p^1_{i,j} = 2\min\{(\frac{1}{2})^k \sum_{i=k_1}^{k} C^i_k,\ (\frac{1}{2})^k \sum_{i=0}^{k_1} C^i_k\} \tag{24}$$

On the same way one can construct the uniformly most powerful test for the hypothesis $h^2_{i,j}$. The test can be written as

$$\varphi^2_{i,j} = \begin{cases} 0, & C_1(m) < k_3 < C_2(m) \\ 1, & else \end{cases} \tag{25}$$

where $m = k_3 + k_4$ The p-value of the test (25) can be calculated by

$$p^2_{i,j} = 2\min\{(\frac{1}{2})^m \sum_{i=k_3}^{m} C^i_m,\ (\frac{1}{2})^m \sum_{i=0}^{k_3} C^i_m\} \tag{26}$$

Note that by construction all individual tests are distribution free uniformly most powerful tests of Neymann structure.

Rejection graph

We select 100 stocks from US market with a highest trading volume during the period of 8 years, from 01.01.2006 to 31.12.2013. We compare results for different periods of observations: 8 periods of 1 year each, 4 periods of 2 years each, 2 periods of 4 years each and 1 period of 8 years. Significance level of multiple tests are set to $\alpha = 0,1$ and $\alpha = 0,5$. To describe the results of multiple testing we introduce a *rejection graph*. Edge (i, j) is included in the rejection graph for hypotheses h^1 iff the hypothesis $h^1_{i,j}$ is rejected by multiple testing procedure. Nodes of the rejection graph are vertices adjacent to these edges.

Rejection graph

The Figure illustrates the structure of the rejection graph for the year 2006, $\alpha = 0.5$, US market.