Statistical procedures for network structures identification with invariant risk function

Petr Koldanov

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

Perm, Russia, September 28, 2018

Outline

Network model

One way to analyze a complex system is to consider associated 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.

Examples: social networks, market networks, biological network.

Network structures

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), planar maximally filtered graph (PMFG), market graph (MG), cliques and independent sets of MG.
- Market graph (TG) of network model $G = (V, E, \gamma)$ - subgraph $G'(\gamma_0) = (V', E') : V' = V; E' \subseteq E, E' = \{(i, j) : \gamma_{i,j} > \gamma_0\}$, where γ_0 - given threshold.
- MST of network model $G = (V, E, \gamma)$ - tree (graph without circle) $G' = (V', E') : V' = V; E' \subset E; |E'| = |V| - 1$; such that $\sum_{(i,j) \in E'} \gamma_{i,j}$ is maximal.

History of market network analysis

- Mantegna(1999) - MST for market network.
- Pardalos (2003) - MG for market network.
- Now there are around 3000 papers.
- Main purpose - network structure construction by numerical algorithms to real market data (stock returns) and interpretation of obtained results.
- But the quality of obtained results is unknown.

Problem description

- Stocks returns are random variables.
  1. to choose measure of association between random variables.
  2. to choose a joint distribution of random variables
- Key problem - identify these network structures by observations of complex system elements or to construct statistical procedure $\delta(x)$ with appropriate properties to identify network structure from observations.

Random variable network. Distribution

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

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

Example - market network (nodes X_i correspond to the stock returns).
Assume random vector $(X_1, \ldots, X_N)$ has elliptically contoured distribution $X - ECD(\mu, \Lambda, g)$
Definition: Class of elliptically contoured distribution is given by density functions:

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

where Λ is symmetric positive definite matrix, $g(x) \geq 0$, and $\int_{-\infty}^{\infty} \ldots \int_{-\infty}^{\infty} g(y'y) dy_1 \ldots dy_N = 1$

Random variable network. Measure

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

- $X = (X_1, \ldots, X_N)$−random vector,
- γ−measure of association.
- Popular network:=Pearson network: $\gamma^P_{i,j} = \rho_{i,j} = \frac{E(X_i - E(X_i))(X_j - E(X_j))}{\sigma_i \sigma_j}$
- Alternative network 1:=Sign similarity network: $\gamma^{Sg}_{i,j} = p^{i,j} = P((X_i - E(X_i))(X_j - E(X_j) > 0)$.
- Alternative network 2:=Kendall network $\gamma^{\tau}_{i,j} = P((X_i(1) - X_i(2))(X_j(1) - X_j(2)) > 0)$ where $\begin{pmatrix} X_i(1) \\ X_j(1) \end{pmatrix}, \begin{pmatrix} X_i(2) \\ X_j(2) \end{pmatrix}$ - iid vectors.

Network structures invariance.

Any RVN generate network model. Different RVN could generate the same network model.

- Class $\mathcal{K}$ of vector X distributions such that under fixed γ network models generated by $(X^{(1)}, \gamma)$, $(X^{(2)}, \gamma)$ are coincide

$$\gamma(X_i^{(1)}, X_j^{(1)}) = \gamma(X_i^{(2)}, X_j^{(2)}), \forall X^{(1)}, X^{(2)} \in \mathcal{K}, \forall i, j = 1, \ldots, N$$

- For all distributions from $\mathcal{K}$ network structures coincide also.
- Consider the subclass $\mathcal{K}(\Lambda)$ of $ECD(\mu, \Lambda, g)$ distributions with fixed Λ. Since $\gamma_{i,j}^P = \lambda_{i,j}/\sqrt{\lambda_{i,i}\lambda_{j,j}}$ if it exist, $\Lambda = (\lambda_{i,j})$ then network models, generated by RVNs $(X, \gamma^P), X \in \mathcal{K}(\Lambda)$ are coincide.

Sign measure invariance

- Theorem 1: *If* $X = (X_i, X_j)$ *has elliptical distribution* $ECD(\mu, \Lambda, g)$, *then probability of sign coincidence* $\gamma_{i,j}^{Sg} = P((X_i - \mu_i)(X_j - \mu_j) > 0)$ *does not depend from* g .
- Theorem 2: *If* $X = (X_1, X_2, \ldots, X_N)$ *has distribution from* $ECD(\mu, \Lambda, g)$, *then market graph in Pearson correlation network with threshold* ρ_0 *coincide with market graph in sign network with threshold* $p_0 = \frac{1}{2} + \frac{1}{\pi} \arcsin(\rho_0)$.
- Theorem 3: *If* $X = (X_1, X_2, \ldots, X_N)$ *has distribution from* $ECD(\mu, \Lambda, g)$, *then MST in Pearson correlation network coincide with MST in sign network.*

Identification problem statement

- (X, γ)-RVN, $G = (V, E, \gamma)$-network model.
- $G' = (V', E') : V' \subseteq V, E' \subseteq E$ - network structure.
- X has distribution from $\mathcal{K}(\Lambda)$
- Let $S = (s_{i,j})$, $S \in \mathcal{G}$ - set of adjacency matrices.
- $H_S : \Lambda \in \Omega_S$-hypothesis that network structure has adjacency matrix $S, S \in \mathcal{G}_1 \subseteq \mathcal{G}$.
- observations $X(t) = (X_1(t), \ldots, X_N(t)), t = 1, \ldots, n$

Problem: to construct statistical procedure $\delta(x)$ of selection one from the set of hypotheses H_S, with invariant risk function (does not depend from g).

Statistical procedure

- $\delta(x) = d_Q$ - decision, that network structure has adjacency matrix $Q, Q \in \mathcal{G}$ iff $\Phi(x) = Q$

$$\Phi(x) = \begin{pmatrix} 0 & \varphi_{12}(x) & \dots & \varphi_{1N}(x) \\ \varphi_{12}(x) & 0 & \dots & \varphi_{2N}(x) \\ \dots & \dots & \dots & \dots \\ \varphi_{1N}(x) & \varphi_{2N}(x) & \dots & 0 \end{pmatrix}.$$

$\Phi(x)-$sample network structure.

$$\varphi_{ij}(x) = \begin{cases} 1, & \text{edge (i,j) is added to network structure} \\ 0, & \text{otherwise} \end{cases}$$

Statistical procedure. Pearson network

Let $X(1), \ldots, X(n)$–i.i.d. vectors with existing second moments.
Note $\rho(X_i, X_j) = \rho((X_i(t) - \overline{X_i}), (X_j(t) - \overline{X_j})), \forall t = 1, \ldots, n$
For Pearson correlation network individual hypotheses have the form:
$h_{i,j} : \gamma_{i,j}^P \leq \gamma_0^P$. Individual test is:

$$\varphi_{ij}^{Corr}(x) = \begin{cases} 1, & \dfrac{r_{i,j} - \gamma_0^P}{\sqrt{1 - r_{i,j}^2}} > c_{i,j} \\ 0, & \dfrac{r_{i,j} - \gamma_0^P}{\sqrt{1 - r_{i,j}^2}} \leq c_{i,j} \end{cases}$$

where $r_{i,j}$ is the sample correlation. $c_{i,j}$ is chosen to make the significance level of the test equal to prescribed value $\alpha_{i,j}$.

Statistical procedure. Sign network. Known μ

For sign similarity network (X, γ^{Sg}) individual hypotheses have the form: $h_{i,j} : \gamma_{i,j}^{Sg} \leq \gamma_0^{Sg}$. Define

$$I_{i,j}(t) = \begin{cases} 1, & (x_i(t) - \mu_i)(x_j(t) - \mu_j) \geq 0 \\ 0, & (x_i(t) - \mu_i)(x_j(t) - \mu_j) < 0 \end{cases}$$

$$T_{i,j}^{sg} = \sum_{t=1}^{n} I_{i,j}(t)$$

Individual test is: $\varphi_{ij}^{Sg} = \begin{cases} 1, & T_{i,j}^{sg} > c_{i,j}^{Sg} \\ 0, & T_{i,j}^{sg} \leq c_{i,j}^{Sg} \end{cases}$

Constants $c_{i,j}^{Sg}$ are defined for given α_{ij} from binomial equations $b(n, \gamma_0^{Sg})$

Quality of statistical procedures for network structure identification

- $w(H_S; d_Q) = w(S, Q)$ - loss from the decision d_Q when the hypothesis H_S is true, $w(S, S) = 0, S \in \mathcal{G}$.
- Risk function of statistical procedure $\delta(x)$ is defined by

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

Statistical procedures with invariant risk function

- Definition 1. *Statistical procedure δ for network structure S identification in network model $G = (V, E, \gamma)$, generated by RVN $(X, \gamma) : X \in EC(\mu, \Lambda, g)$, has invariant risk function in the class $\mathcal{K}(\Lambda)$, if risk function $R(S, \theta, \delta), \theta = (\mu, \Lambda, g)$ does not depend from g.*
- Theorem 4: *Let $(X_1, \ldots, X_N)$ has distribution from $ECD(\mu, \Lambda, g)$ Then joint distribution of statistics $T^{sg}_{i,j}$ $(i, j = 1, 2, \ldots, N; i \neq j)$ does not depend from function g.*
- Corollary: If X has distribution from $ECD(\mu, \Lambda, g)$ then algorithms for network structures identification based on statistics $T^{sg}_{i,j}$ has invariant risk function.
- Experimental results shows that procedures based on sample pearson correlation have not invariant risk function.

The case of unknown μ

It was noted

$$\rho(X_i, X_j) = \rho((X_i(1) - \overline{X_i}), (X_j(1) - \overline{X_j}))$$

for all distributions with existing second moments.
But what about of

$$P((X_i - \mu_i)(X_j - \mu_j) > 0) = P((X_i - \overline{X_i})(X_j - \overline{X_j}) > 0) - ?$$

lid random variables

Let $\overline{X_i} = \frac{1}{n}\sum_{t=1}^{n} X_i(t)$

$$Y_i(t) = X_i(t) - \overline{X_i}, i = 1, \ldots, p$$

Theorem 5. If $X - ECD(\mu, \Lambda, g)$ and

$$\int_0^{\infty} r^{p+1} g(r^2) dr < +\infty$$

then

$$P(Y_i(t)Y_j(t) > 0) = P((X_i(t) - \mu_i)(X_j(t) - \mu_j) > 0)$$

lid random variables

Corollary 1: Let random vector $(X_1, \ldots, X_N)$ has elliptically contoured distribution. Then for any function g one has $((i,j) : i,j = 1, \ldots, N, i \neq j)$

$$\gamma_{i,j}^{Sg} = P((X_i(t) - \overline{X_i})(X_j(t) - \overline{X_j}) > 0) = \frac{1}{2} + \frac{1}{\pi}\arcsin(\gamma_{i,j}^{P})$$

Corollary 2: For $n = 2$ one has $P((X_i(t) - \overline{X_i})(X_j(t) - \overline{X_j}) > 0) = P((X_i(1) - X_i(2))(X_j(1) - X_j(2)) > 0) = \gamma_{i,j}^{\tau}$

Corollary 3: If $X - \mathcal{K}(\Lambda)$ then network structures in Pearson correlation network, network structures in sign similarity network and network structures in $\tau-$Kendall network are equivalent and defined by the matrix Λ only.

Statistical procedure. Sign network. Unknown μ

For sign similarity network (X, γ^{Sg}) individual hypotheses have the form: $h_{i,j} : \gamma_{i,j}^{Sg} \leq \gamma_0^{Sg}$. Define

$$I_{i,j}^1(t) = \begin{cases} 1, & (x_i(t) - \overline{x_i})(x_j(t) - \overline{x_j}) \geq 0 \\ 0, & (x_i(t) - \overline{x_i})(x_j(t) - \overline{x_j}) < 0 \end{cases}$$

$$T_{i,j}^s = \sum_{t=1}^{n} I_{i,j}^1(t)$$

Individual test is: $\varphi_{ij}^S = \begin{cases} 1, & T_{i,j}^s > c_{i,j}^S \\ 0, & T_{i,j}^s \leq c_{i,j}^S \end{cases}$

Constants $c_{i,j}^S$ are defined for given α_{ij} from asymptotic normal distribution.

Matrix elliptically contoured distribution

Let

$$X = \begin{pmatrix} X_1(1) & X_1(2) & \dots & X_1(n) \\ X_2(1) & X_2(2) & \dots & X_2(n) \\ \dots & \dots & \dots & \dots \\ X_p(1) & X_p(2) & \dots & X_p(n) \end{pmatrix}$$

be the random matrix $p \times n$.

Definition[1] Matrix X has matrix elliptically contoured distribution $X - E_{p,n}(M, \Sigma \otimes \Phi, \psi)$, if its characteristic function has the form:

$$\phi_X(t) = \exp\left(tr(iT'M)\right)\psi\left(tr(T'\Sigma T\Phi)\right)$$

where

$$T : p \times n; M : p \times n; \Sigma : p \times p; \Phi : n \times n; \Sigma \geq 0; \Phi \geq 0$$

and $\psi : [0, +\infty) \to R$

[1] Gupta F.K. Varga T. Bodnar T. Elliptically Contoured Models in Statistics and Portfolio Theory, Springer, 2013, ISBN: 978-1-4614-8153-9.

Sign measure equivalence. Unknown μ.

Theorem 6 If matrix

$$X = \begin{pmatrix} X_1(1) & X_1(2) & \dots & X_1(n) \\ X_2(1) & X_2(2) & \dots & X_2(n) \\ \dots & \dots & \dots & \dots \\ X_p(1) & X_p(2) & \dots & X_p(n) \end{pmatrix} - E_{p,n}(M, \Sigma \otimes \Phi, \psi)$$

then for any matrix Φ matrix

$$\begin{pmatrix} X_i(t) - \overline{X_i} \\ X_j(t) - \overline{X_j} \end{pmatrix} - E_{2,1}\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} \varphi_{t,t}(\Phi)\sigma_{11} & \varphi_{t,t}(\Phi)\sigma_{12} \\ \varphi_{t,t}(\Phi)\sigma_{12} & \varphi_{t,t}(\Phi)\sigma_{22} \end{pmatrix}, \psi\right)$$

$\forall i, j = 1, \dots, p; i \neq j$ for $\forall t = 1, \dots, n$

Corollary 4: For $\forall t = 1, \dots, n$

$$P((X_i(t) - \overline{X_i})(X_j(t) - \overline{X_j}) > 0) = P((X_i(t) - \mu_i)(X_j(t) - \mu_j) > 0)$$

Our publications.

- 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. Probability of sign coincidence centered with respect to sample mean random variables// Vestnik TvGU. Series: Applied mathematics. Accepted to publication.

THANK YOU FOR YOUR ATTENTION!

Experimental results

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 reference matrix. Structures of the matrix are considered as reference 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 (γ^P or γ^{Sg}).
5. We construct the sample MG basing on the estimations and compare it to the reference network structure.

Experimental results

The model is the mixture distribution consisting of multivariate normal distribution and multivariate Student distribution with 3 degree of freedom.

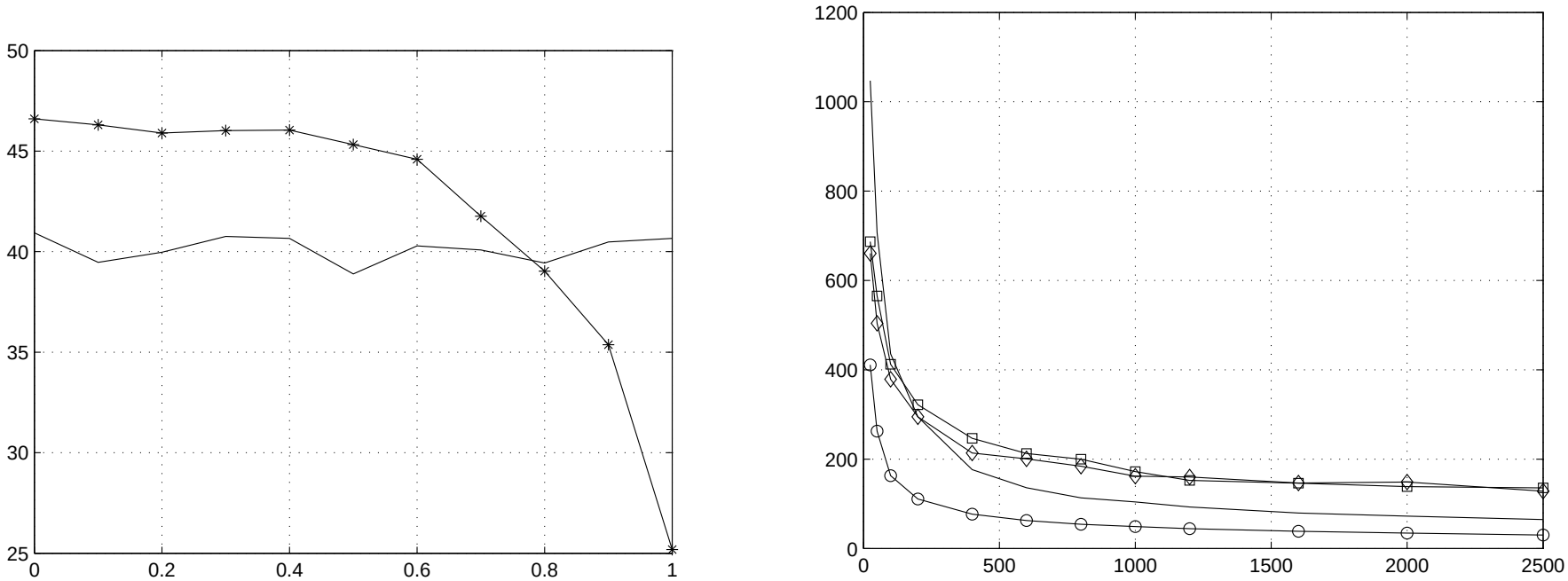

Figure: Risk function for MG. Left - $n = 400$, star line - δ_P, line - δ_{Sg}, right: circle - $\gamma = 1$, δ_P; diamond - $\gamma = 0,5$, δ_P; square - $\gamma = 0$, δ_P, line - δ_S.

Matrix elliptically contoured distribution. Theorems[2]

Theorem 2.1 Matrix $X : p \times n$ has matrix elliptically contoured distribution $X - E_{p,n}(M, \Sigma \otimes \Phi, \psi)$ iff $x = vec(X')$ has elliptically contoured distribution $ECD_{pn}(vec(M'), \Sigma \otimes \Phi, \psi)$.

Theorem 2.2 Let $X - E_{p,n}(M, \Sigma \otimes \Phi, \psi)$, $A : q \times p; B : n \times m; C : q \times m$ be the matrix with elements from R^1. Then
$AXB + C - E_{q,m}(AMB + C, (A\Sigma A)' \otimes (B'\Phi B), \psi)$

Theorem 2.8 Let $X - E_{p,n}(M, \Sigma \otimes \Phi, \psi)$. Divide X, M, Σ by submatrices

$$X = \begin{pmatrix} X_1 \\ X_2 \end{pmatrix}, M = \begin{pmatrix} M_1 \\ M_2 \end{pmatrix}, \Sigma = \begin{pmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{pmatrix}$$

where X_1- matrix of $q \times n$, M_1-matrix of $q \times n$, $\Sigma_{11}-$matrix of $q \times q$. Then $X_1 - E_{q,n}(M_1, \Sigma_{11} \otimes \Phi, \psi)$

[2]Gupta F.K. Varga T. Bodnar T. Elliptically Contoured Models in Statistics and Portfolio Theory, Springer, 2013, ISBN: 978-1-4614-8153-9.