Statistics of individual tests for market graph identification in market network

Petr Koldanov

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

Moscow, Russia, April 12, 2019

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), maximum cliques (MC) and maximum independent sets (MIS) 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. Maximum cliques and maximum independent sets.
- 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. Examples of interpretation.

Problem description

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

Random variable network

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

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

Example - market network (nodes correspond to the stocks, behaviour of stocks is described by returns)

- Popular network:=Pearson network: $\gamma_{i,j}^P = \rho_{i,j} = \frac{E(X_i - E(X_i))(X_j - E(X_j))}{\sigma_i \sigma_j}$
- Alternative network 1:=Sign similarity network: $\gamma_{i,j}^{Sg} = p^{i,j} = P((X_i - E(X_i))(X_j - E(X_j) > 0)$.
- Alternative network 2:=Kendall network: $\gamma_{i,j}^{Kd} = 2P(X_i(1) - X_i(2)(X_j(1) - X_j(2)) > 0) - 1$

Any random variable network generate network model. Network model is complete weighted graph $G = (V, E, \gamma)$

True network structures

Any network structure could be defined by adjacency matrix

$$S = \begin{pmatrix} 0 & s_{12} & \dots & s_{1N} \\ s_{12} & 0 & \dots & s_{2N} \\ \dots & \dots & \dots & \dots \\ s_{1N} & s_{2N} & \dots & 0 \end{pmatrix}.$$

$S-$true network structure.

$$s_{ij} = \begin{cases} 1, & \text{edge (i,j) is included to the true network structure} \\ 0, & \text{otherwise} \end{cases}$$

Statistical procedure

In real practice available data for analysis is sample of observations

$$\begin{pmatrix} X_1(1) \\ X_2(1) \\ \dots \\ X_N(1) \end{pmatrix}, \dots, \begin{pmatrix} X_1(n) \\ X_2(n) \\ \dots \\ X_N(n) \end{pmatrix}$$

The problem of the market graph identification can be considered as multiple hypotheses testing problem of the following individual hypotheses:

$$h_{ij} : \gamma_{i,j} \leq \gamma_0 \quad (s_{i,j} = 0) \text{ versus } k_{ij} : \gamma_{i,j} > \gamma_0 \quad (s_{i,j} = 1)$$

Statistical procedure

Any statistical procedure for the market graph identification is therefore based on individual tests $\varphi_{ij}(x)$ of testing the individual hypotheses $h_{ij} : \gamma_{i,j} \leq \gamma_0$ versus $k_{ij} : \gamma_{i,j} > \gamma_0$.

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

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

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

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

Statistical procedure. Pearson network with normal distribution

For Pearson correlation network with normal distribution individual hypotheses have the form: $h_{i,j} : \gamma^P_{i,j} \leq \gamma^P_0$. Individual test is:

$$\varphi^{PN}_{ij}(x) = \begin{cases} 1, & \sqrt{n-1}\left(\dfrac{r_{i,j} - \gamma^P_0}{\sqrt{1 - r^2_{i,j}}}\right) > c^{PN}_{i,j} \\ 0, & \sqrt{n-1}\left(\dfrac{r_{i,j} - \gamma^P_0}{\sqrt{1 - r^2_{i,j}}}\right) \leq c^{PN}_{i,j} \end{cases}$$

where $r_{i,j}$ is the sample correlation. $c^{PN}_{i,j}$ is chosen to make the significance level of the test equal to prescribed value $\alpha_{i,j}$. For $n \to \infty$

$$p^{PN}_{i,j} = 1 - \Phi\left(\sqrt{n-1}\left(\frac{r_{i,j} - \gamma^P_0}{\sqrt{1 - r^2_{i,j}}}\right)\right)$$

Pearson network with elliptical distribution [1]

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

$\overline{\kappa} = \frac{\sum_{t=1}^n (x(t)-\overline{x})' S^{-1}(x(t)-\overline{x})}{(n-1)N(N+2)}$, $S = \sum_{t=1}^n (x(t)-\overline{x})'(x(t)-\overline{x})$. For $n \to \infty$

$$p_{i,j}^P = 1 - \Phi\left(\sqrt{\frac{n-1}{1+\overline{\kappa}}}\left(\frac{r_{i,j}-\gamma_0^P}{\sqrt{1-r_{i,j}^2}}\right)\right)$$

---

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

Statistical procedure. Sign network

For sign similarity network (X, γ^{Sg}) individual hypotheses have the form: $h_{i,j} : \gamma^{Sg}_{i,j} \le \gamma^{Sg}_0$. Define

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

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

Individual test is: $\varphi^{Sg}_{ij} = \begin{cases} 1, & T^{sg}_{i,j} > c^{Sg}_{i,j} \\ 0, & T^{sg}_{i,j} \le c^{Sg}_{i,j} \end{cases}$

$$p^{Sg}_{i,j} = 1 - F_{\gamma^{Sg}_0}\left(T^{sg}_{i,j}\right)$$

where $F_{\gamma^{Sg}_0}(x)$ is the distribution function of the binomial distribution $b(n, \gamma^{Sg}_0)$. $c^{Sg}_{i,j}$ is chosen to make the significance level of the test equal to $\alpha_{i,j}$. In the case of unknown μ replace μ_i by $\overline{x_i}$.

Statistical procedure. Kendall network

For Kendall network (X, γ^{Kd}) individual hypotheses have the form: $h_{i,j} : \gamma_{i,j}^{Kd} \leq \gamma_0^{Kd}$. Individual test is:

$$\varphi_{ij}^{Kd} = \begin{cases} 1, & T_{ij}^{Kd} > c_{i,j}^{Kd} \\ 0, & T_{ij}^{Kd} \leq c_{i,j}^{Kd} \end{cases}$$

where

$$T_{ij}^{Kd} = \frac{1}{n(n-1)} \sum_{t \neq s} sign\left((x_i(t) - x_i(s))(x_j(t) - x_j(s))\right)$$

$c_{i,j}^{Kd}$ is chosen to make the significance level of the test equal to $\alpha_{i,j}$. For $n \to \infty$ and $\gamma_{i,j}^{Kd} = 0$

$$p_{i,j}^{Kd} = 1 - \Phi\left(\sqrt{\frac{9n(n-1)}{2(2n+5)}}\left(T_{ij}^{Kd} - \gamma_0^{Kd}\right)\right)$$

Experimental results.[2]

Class of elliptically contoured distribution with fixed matrix Λ.
The hypothesis $\lambda_{i,j} = 0$ is tested.

- Robustness of significance level with respect to g.
- Robustness of power function with respect to g.

The mixture distribution - $X = (X_1, \dots, X_N)$ takes value from $N(0, \Lambda)$ with probability ϵ and from $t_3(0, \Lambda)$ with probability $1 - \epsilon$.

- $\epsilon = 1$ - normal case.
- $\epsilon = 0$ - Student case.

---

[2] 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
$\int_{-\infty}^{\infty} \dots \int_{-\infty}^{\infty} g(y'y) dy_1 \dots dy_N = 1$

Experimental results. Robustness of significance level

- For $\alpha = 0.1$ and $\lambda_{ij} = 0$ test $\varphi_{ij}^{PN}(x)$ does not robust to deviation from normality. Namely under $n = 50, \epsilon = 1$ one has 104 rejection from 1000 experiments. But for decreasing of ϵ the number of rejection is increased. For $\epsilon = 0$ one has 255 rejections.
- For $\alpha = 0.1$ and $\lambda_{ij} = 0$ test $\varphi_{ij}^{P}(x)$ does not robust to deviation from normality. Namely under $n = 50, \epsilon = 1$ one has 108 rejection from 1000 experiments. But for decreasing of ϵ the number of rejection is increased. For $\epsilon = 0$ one has 177 rejections.

Then corrected Pearson test does not valid $\alpha-$level test under deviation from normality.

Experimental results. Robustness of significance level

- For $\alpha = 0.05$ and $\lambda_{ij} = 0$ test $\varphi_{ij}^{Kd}(x)$ does not robust to deviation from normality. Namely under $n = 50, \epsilon = 1$ one has 52 rejection from 1000 experiments. But for decreasing of ϵ the number of rejection is increased. For $\epsilon = 0$ one has 94 rejections.
- For all α and $\lambda_{ij} = 0$ test $\varphi_{ij}^{Sg}(x)$ is robust to deviation from normality. [3]

[3] 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.

Experimental results. Robustness of power function

- For

$$\alpha = 0.05, n = 100, \epsilon = 1, \lambda_{ij} = 0.3$$

power function of test $\varphi_{ij}^{PN}(x)$ is 0.927 ($\hat{\alpha} = 0.046$). But for

$$\alpha = 0.05, n = 100, \epsilon = 0, \lambda_{ij} = 0.3$$

power function of test $\varphi_{ij}^{PN}(x)$ is 0.771 ($\hat{\alpha} = 0.21$).

- For

$$\alpha = 0.05, n = 100, \epsilon = 1, \lambda_{ij} = 0.3$$

power function of test $\varphi_{ij}^{P}(x)$ is 0.933 ($\hat{\alpha} = 0.046$). But for $\alpha = 0.05$, $n = 100$, $\epsilon = 0$ and $\lambda_{ij} = 0.3$ power function of test $\varphi_{ij}^{P}(x)$ is 0.611 ($\hat{\alpha} = 0.111$).

Experimental results. Robustness of power function

- For

$$\alpha = 0.1, n = 25, \epsilon = 1, \lambda_{ij} = 0.45$$

power function of test $\varphi_{ij}^{Kd}(x)$ is 0.828 ($\hat{\alpha} = 0.103$). But for

$$\alpha = 0.1, n = 25, \epsilon = 0, \lambda_{ij} = 0.45$$

power function of test $\varphi_{ij}^{Kd}(x)$ is 0.780 ($\hat{\alpha} = 0.125$).

- Power function of the test $\varphi_{ij}^{Sg}(x)$ is robust to deviation from normality.[4]

[4] 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.

Experimental results. Robustness of significance level and power function for $\alpha = 0.5$

- For $\alpha = 0.5$ the probability of first kind error is equal to 0.5 for any ϵ.
- For power function the result does not valid. Namely for $\varphi_{ij}^{PN}(x), \varphi_{ij}^{P}(x)$ power function is 0.94 for $\epsilon = 1, \lambda = 0.15, n = 100$. But power function is 0.95 for $\epsilon = 0, \lambda = 0.35, n = 100$.
- For $\alpha = 0.5$ significance levels and power functions of the tests $\varphi_{ij}^{Sg}(x)$ and $\varphi_{ij}^{Kd}(x)$ are robust to deviation from normality.
- For $\alpha = 0.5$ power function of $\varphi_{ij}^{Kd}(x)$ is uniformly better than $\varphi_{ij}^{Sg}(x)$. Namely for $\epsilon = 0, \lambda = 0.3, n = 50$ power function of $\varphi_{ij}^{Kd}(x)$ is 0.97. For $\varphi_{ij}^{Sg}(x)$ power function is 0.97 for $\lambda = 0.4$ or power function of $\varphi_{ij}^{Sg}(x)$ is 0.97 for $n = 100$.

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.
- Kalyagin V. A., Koldanov A. P., Koldanov P., Pardalos P. M. Optimal decision for the market graph identification problem in a sign similarity network // Annals of Operations Research. 2018. P. 1-15
- Koldanov P. Probability of sign coincidence centered with respect to sample mean random variables// Vestnik TvGU. Series: Applied mathematics. 2018. N 4. p. 23-30.

THANK YOU FOR YOUR ATTENTION!