Measures of similarity and network structures on stock markets

Valery A. Kalyagin (a), Alexander P. Koldanov (a), Panos M. Pardalos (a, b) [1]

1 Introduction

Correlation networks represent an important class of network models and have various applications (Horvath, 2011). The use of correlations networks in financial analysis is associated with a network analysis of stock markets (Tumminello, 2010), (Boginski, 2005). Two research directions are presented in the network analysis of the stock markets. The first is associated with empirical analysis of specific stock markets. Publications in this area is quite numerous. The second direction is related with methodological and algorithmic aspects of the analysis (Boginski, 2006), (Tumminello, 2005), (Koldanov, 2013). An important issue in network analysis of the stock markets is the question of the choice of measure of similarity between stocks, which is associated with the weight of edges in the network. Correlation networks can be generated by different correlations: Pearson correlation, Fechner correlation, Kendall correlation, Spearman correlation, and others. In the literature on network analysis of stock markets there are numerous examples of applying different correlations for investigation of important network structures such as maximum spanning tree (MST), market graph (MG) and other. However, there are no works containing comparative analysis of different correlation networks. What is different in different correlation networks? An important characteristic in network study is to estimate the statistical uncertainty of identification of network structures (MST, MG, others) from observations. The main question addressed in this paper is: how to compare the correlation networks with respect to uncertainty of identification of network structures. To answer this question we develop a random variable network model.

[1] (a) - Laboratory of Algorithms and Technologies for Network Analysis, National Research University Higher School of Economics, Bolshaya Pecherskaya 25/12, Nizhny Novgorod, 603155 Russia, *vkalyagin@hse.ru* and (b) - University of Florida, ISE Department, Gainesville, FL 32611, USA.

2 Random variables network. Network structures.

Random variables network is a pair (X, γ), where X is a random vector, γ is a measure of association (dependence, similarity) of random variables. Random variables network generates a random variables network model. Nodes of random variables network model are associated with random variables X_i, $i = 1, 2, \ldots, N$. Links between nodes are represented by a weighted edges. The weight of the edge (i, j) is given by $\gamma(X_i, X_j)$. According to the choice of measure of association one get different correlation networks: Pearson correlation network, sign similarity network, Fechner correlation network, Kruskal correlation network, Kendall correlation network, Spearman correlation network and so on.

To understand a correlation network model one needs to filter the information in the complete weighted graph. A *network structure* is a subgraph of the complete weighted graph, which contains a valuable information about network. Different network structures are known in the literature: maximum spanning tree, market graph, cliques and independent sets in the market graph. *Maximum spanning three* (MST) is the spanning tree of maximal total weight. MST gives some information on hierarchical structure of stocks in the stock market (Tumminello 2005). An edge between two nodes is included in the *market graph* (MG), if the corresponding measure of similarity is larger than a given threshold. Maximum cliques, maximum independent sets in the market graph are useful sources of market data mining. In graph theory, a clique is a subset of nodes of a graph such that its induced subgraph is complete; that is, every two distinct nodes in the clique are connected. A clique in the market graph represents a set of closely interconnected stocks. The independent set is a set of nodes in a graph, which has no adjacent nodes (nodes connected by an edge). The independent set of the market graph represents a set of non-connected stocks. The concept of the market graph was introduced in (Boginski 2005). Since then, different aspects of the market graph approach have been developed in the literature.

3 Uncertainty of network structures identification.

An important problem in market network analysis is uncertainty problem. Uncertainty in the data implies uncertainty in the network structure. This uncertainty is related with the algorithm of network structure identification from observations. Let (X, γ) be a random variables network. *True network structure* is the network structure calculated for random variables network model (complete weighted graph) associated with the random variables network (X, γ). One needs to identify the true network structure from observations. We model the observations by random vectors

$$X(1), X(2), X(3), \dots, X(n)$$

where $X(t)$ are i.i.d. random vectors with the same distribution as X, n is the sample size. An identification procedure δ is a map from the sample space $R^{N\times n}$ into the decision space (set of possible structures). Two type of errors occur during identification

> Type I error: edge (i, j) is included in δ-structure when it is absent in the true structure.
>
> Type II error: edge (i, j) is not included in δ-structure when it is present in the true structure.

Each error generates a loss. For network structure identification it is natural to consider the case with additive losses. Denote by $a_{i,j}$ the loss associated with the type I error and $b_{i,j}$ the loss associated with the type II error for the edge (i, j). Additive loss function W_{add} is defined as the sum of individual losses. To measure uncertainty of identification procedure we use the risk function

$$Risk(W; \delta) = E(W; \delta),$$

i.e. the risk is the expected value of the loss. Larger is the value of risk, higher is uncertainty of identification procedure.

4 Connections between correlations networks

Consider the following measures of associations (Kruskal 1958):

Pearson correlation: $\gamma^P(X, Y) = \frac{Cov(X,Y)}{\sqrt{Cov(X,X)}\sqrt{Cov(Y,Y)}}$
Sign similarity: $\gamma^{Sg}(X, Y) = P\{(X - E(X))(Y - E(Y)) > 0\}$

Fechner correlation: $\gamma^{Fh}(X, Y) = 2\gamma^{Sg} - 1$

Kruskal correlation: $\gamma^{Kr}(X,Y) = 2P\{(X - \text{Med}X)(Y - \text{Med}Y) > 0\} - 1$

τ-Kendall correlation: $\gamma^{Kd}(X,Y) = 2P\{(X_1 - X_2)(Y_1 - Y_2) > 0\} - 1$,
where (X_1, Y_1), (X_2, Y_2) i.i.d. as (X, Y).

Spearman correlation: $\gamma^{Sp}(X,Y) = 6P\{(X_1 - X_2)(Y_1 - Y_3) > 0\} - 3$,
where (X_1, Y_1), (X_2, Y_2), (X_3, Y_3) i.i.d. as (X, Y).

Each correlation generates a correlation network. To find a connections between network structures we use a large class of elliptically contoured distributions well known in stock market analysis (Gupta 2013). Elliptically contoured distribution $X \sim Ell(\mu, \Lambda; g)$ is defined by the density function

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

where Λ is positive definite matrix, $g(y) \geq 0$. Particular cases of elliptically contoured distributions are Gaussian and Student distributions. For Gaussian multivariate distribution one has

$$f_N(x; \mu, \Lambda) = \frac{1}{(2\pi)^{N/2}|\Lambda|^{1/2}} \exp\{-(1/2)(x - \mu)'\Lambda^{-1}(x - \mu)\}$$

For Student multivariate distribution one has

$$f_S(x; \mu, \Lambda, m) = \frac{\Gamma(\frac{m+N}{2})}{\Gamma(\frac{m}{2}) m^{N/2} \pi^{N/2} |\Lambda|^{1/2}} (1 + \frac{(x - \mu)'\Lambda^{-1}(x - \mu)}{m})^{-\frac{m+N}{2}}$$

There are a connections between different correlations in the class of elliptically contoured distributions:

Theorem: Under some general conditions for any $X \sim Ell(\mu, \Lambda; g)$ one has

$$\gamma^{Fh}(X_i, X_j) = \gamma^{Kr}(X_i, X_j) = \gamma^{Kd}(X_i, X_j) = \frac{2}{\pi} \text{arcsin}(\gamma^P(X_i, X_j)),$$

$$\gamma^{Sp}(X_i, X_j) = \frac{6}{\pi} \text{arcsin} \frac{\gamma^P(X_i, X_j)}{2}$$

$$\gamma^{Sg}(X_i, X_j) = \frac{1}{2} + \frac{1}{\pi} \text{arcsin}(\gamma^P(X_i, X_j))$$

5 Discussions

Interconnections between correlations implies some interesting consequences. First, one can prove that true MST (maximum spanning tree) is the same in all correlation networks

(X, γ), for a fixed X. Indeed, to calculate the true MST one can use the Kruskal algorithm. The main steps of the algorithm are: order the edges according to decreasing weights, add the edge in the MST if this does not produce a cycle. Any transformation above from one correlation to another one is an increasing function. It means that the edge ordering will be the same in any correlation network. Second, the true market graphs in different networks are interconnected too. One can obtain a true market graph in one network as a true market graph in another network but for different value of threshold. Indeed, because the transformation functions from one correlation to another one are increasing the relation $\gamma^{(1)} \geq \gamma_0^{(1)}$ is equivalent to the relation $\gamma^{(2)} \geq \gamma_0^{(2)}$ for some appropriate choice of $\gamma_0^{(1)}$, $\gamma_0^{(2)}$.

However, uncertainty of network structure (MST, MG) identification procedures in different correlation networks are different. This surprising phenomenon produce a new problems. First, one is interested to construct a robust (distribution free) identification procedures in correlation networks. Second, it is important to construct optimal identification procedures, with minimal uncertainty. One can show that it is possible to construct a robust (distribution free in the class of elliptically contoured distributions) identification procedures for MST and MG for each correlation network. Moreover, for MG network structure it is possible to construct identification procedures with minimal risk.

Aknowledgement: This work is conducted at the National Research University Higher School of Economics, Laboratory of algorithms and technologies for network analysis. The author Valery A. Kalyagin is supported by RFHR grant 15-32-01052.

References

[1] Boginski V., Butenko S., Pardalos P.M. Statistical analysis of financial networks, Computational Statistics and Data Analysis. 48 (2), 431–443 (2005).

[2] Boginski V., Butenko S., Pardalos P.M. Mining market data: a network approach J. Computers and Operations Research. 33 (11) 3171–3184 (2006).

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

[4] Horvath S. Weighted Network Analysis: Application in Genomics and Systems Biology, Springer book, 2011.

[5] Kalyagin V.A., Koldanov A.P., Koldanov P.A., Pardalos P.M., Zamaraev V.A. Measures of uncertainty in market network analysis, Physica A: Statistical Mechanics and its Applications, v. 413, No. 1, pp. 59-70 (2014).

[6] Koldanov A.P., Koldanov P.A., Kalyagin V.A., Pardalos P.M. Statistical procedures for the market graph construction. Computational Statistics and Data Analysis 68 17–29 (2013).

[7] Kruskal W.H. Ordinal Measures of Association, Journal of American Statistical Association, v. 53, pp. 814-861 (1958).

[8] Tumminello M., Aste T., Matteo T.D., Mantegna R.N. A tool for filtering information in complex systems. Proceedings of the National Academy of Sciences. 102 (30), 10421–10426 (2005).

[9] Tumminello M., Lillo F., Mantegna R.N. (2010). Correlation, Hierarchies and Networks in Financial Markets, J. of Econ. Behavior Organization. Vol. 75. P. 40-58.