Optimal decision for the market graph identification problem in a sign similarity network

V. A. Kalyagin1 · A. P. Koldanov1 · P. A. Koldanov1 · P. M. Pardalos1,2

Abstract Research into the market graph is attracting increasing attention in stock market analysis. One of the important problems connected with the market graph is its identification from observations. The standard way of identifying the market graph is to use a simple procedure based on statistical estimations of Pearson correlations between pairs of stocks. Recently a new class of statistical procedures for market graph identification was introduced and the optimality of these procedures in the Pearson correlation Gaussian network was proved. However, the procedures obtained have a high reliability only for Gaussian multivariate distributions of stock attributes. One of the ways to correct this problem is to consider different networks generated by different measures of pairwise similarity of stocks. A new and promising model in this context is the sign similarity network. In this paper the market graph identification problem in the sign similarity network is reviewed. A new class of statistical procedures for the market graph identification is introduced and the optimality of these procedures is proved. Numerical experiments reveal an essential difference in the quality between optimal procedures in sign similarity and Pearson correlation networks. In particular, it is observed that the quality of the optimal identification procedure in the sign similarity network is not sensitive to the assumptions on the distribution of stock attributes.

Keywords Pearson correlation network · Sign similarity network · Market graph · Multiple decision statistical procedures · Loss function · Risk function · Optimal multiple decision procedures
1 Introduction

There is a variety of data mining techniques applied to stock markets. Different optimization techniques are used for portfolio selection (Boginski et al. 2014; Cesarone et al. 2015). Other techniques are based on the analysis of the market network and its structures (Mantegna 1999; Boginsky et al. 2005). The market network is a complete weighted graph where the nodes are associated with stocks and weights of edges are given by some measure of similarity in the stock’s behavior. The market graph is an important structure in the market network. An edge between two nodes is included in the market graph, if the corresponding measure of similarity is larger than a given threshold. Maximum cliques, maximum independent sets, and degree distribution 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 degree of a node of a graph is the number of edges incident to this node. The degree of a node in the market graph can be interpreted as the influence of the associated stock on the market (see Boginsky et al. 2005).

The concept of the market graph was introduced in Boginsky et al. (2003, 2004). Since then, different aspects of the market graph approach have been developed in the literature. Most publications are related to the experimental study of real markets. The power law phenomenon first observed for the US stock market in Boginsky et al. (2005) was then developed in Huang et al. (2009) and Tse et al. (2010). Clustering in the Pearson correlation-based financial network was studied in Onella et al. (2004). The market graph structure was applied to study the dynamics of the US market in Boginsky et al. (2006). Market crashes of the US stock market were analyzed in Emmert-Streib and Dehmer (2010) with the help of the market network model. The specifics of different financial markets were investigated in Bautin et al. (2013) and Vizgunov et al. (2014) for the Swedish and the Russian market, Garas and Argyrakis (2007) for the Greek market, Huang et al. (2009) for the Chinese market, and in Namaki et al. (2011) for the Iranian market. The sign measure of similarity for the market network model was introduced in Bautin et al. (2013), and it was compared with the Pearson correlation in Bautin et al. (2013). It was shown that the sign measure is easy to interpret and it leads to robust identification of the market graph. Hub discovery in a large-scale partial correlation network was studied in Hero and Rajaratnam (2012), where a theory and algorithms for discovering hubs were developed. Dominant stocks which drive the correlations present among stocks were investigated in Kenett et al. (2010) using the partial correlation network. The evolution of the market graph for the US market in Spearman correlation network was studied in Shirokikh et al. (2013), where some cohesive clusters of stocks were revealed. A dynamic time warping measure of similarity was used in Wang et al. (2012) to study the topology evolution of foreign exchange markets. It was shown that USD and Euro are the predominant world currencies, but USD losses the most central position while Euro acts as a stable center during the crisis. Some efficient algorithms related to the calculation of isolated cliques in a market graph were presented in Gunawardena et al. (2012) and Huffner et al. (2008).

However, an economic interpretation of market network data mining is not complete without an estimation of the reliability of the obtained results. The reliability of the minimum spanning tree in the Pearson correlation network was studied by the bootstrap method in Tumminello et al. (2007). In this paper we study the reliability of the market graph using a
Our approach is based on the multiple decision theory and the random variables network model. The nodes of the network are random variables and the edges’ weights are given by a measure of pairwise similarity between the random variables. The observed values of stock attributes are modeled by a sample from the distribution of random variables. The reliability of the network structure can now be measured by the risk function of a statistical procedure for its identification. Statistical procedures with minimal risk (maximum reliability) are of great practical interest. A class of optimal statistical procedures for the identification of the market graph in the Pearson correlation based network was introduced and studied in Koldanov et al. (2013). One can see from numerical experiments in Koldanov et al. (2013) that the value of the risk function of the procedures of this class essentially depends on the assumptions on multivariate distributions of stock attributes. Taking this into account, it is of interest to study distribution-free identifying statistical procedures. A new and promising approach in this context is to consider a sign correlation based (sign similarity) network.

In this paper we study optimal statistical procedures for market graph identification in a sign similarity (sign correlation based) network. Our construction of optimal procedures is based on the simultaneous inference of optimal two-decision procedures. We prove that the constructed procedure is optimal under the following assumptions: additivity of the loss function, unbiasedness of the procedure, and sign symmetry of the distributions. We give a direct proof which simplifies the general approach by Lehmann (1957). In addition, we compare the risk function of the optimal procedure in the sign similarity network with the risk function of the optimal procedure in the Pearson correlation network. Numerical experiments show an essential difference in risk behavior for two optimal statistical procedures. For the multivariate Gaussian distribution, both procedures assure a control of risk, when the significance level of individual tests is changing. In contrast, for the multivariate Student distribution, the optimal procedure in the Pearson correlation network does not control risk while the optimal procedure in the sign similarity network does. Therefore, the quality of the optimal identification procedure in the sign similarity network is not sensitive to the assumptions on the distribution of stock attributes.

The paper is organized as follows. In Sect. 2 we give the basic definitions and notations. In Sect. 3 we describe a multiple decision framework for the threshold graph identification problem. In Sect. 4 we discuss the concept of optimality of identification procedures. In Sect. 5 we construct a multiple decision identification procedure in the sign similarity network. In Sect. 6 we give a proof of the optimality of this procedure. In Sect. 7 we present numerical experiments to compare the optimal procedures in sign similarity and Pearson correlation networks. In Sect. 8 we give some concluding remarks.

2 Market graph identification problem

Consider a network generated by a random vector \( X = (X_1, X_2, \ldots, X_N) \). The network nodes are random variables \( X_i, i = 1, \ldots, N \) and the weight of edge \((i, j)\) is given by a pairwise measure of association \( \gamma \):

\[
\gamma_{i,j} = \gamma(X_i, X_j), \quad \text{for } i, j = 1, 2, \ldots, N.
\]

The obtained network is a complete weighted graph which we will call the random variables network. The random variables network is defined by the multivariate distribution of the vector \( X \) and by the choice of the measure of association \( \gamma \). The network based on the
Pearson correlation of random variables will be called the Pearson correlation network. The network based on the probability of pairwise sign coincidence will be called the sign similarity network.

For any network, the market graph is constructed as follows: the edge between two vertices $i$ and $j$ is included in the market graph, if $\gamma_{i,j} > \gamma_0$, where $\gamma_0$ is a given threshold. Henceforth we will call this network structure the reference (true) market graph.

In the sign similarity network the weight of the edge $(i, j)$ is defined by

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

For a given threshold $p_0$ the reference market graph in the sign similarity network is constructed as follows: the edge between two nodes $i$ and $j$ is included in the reference market graph if $p^{i,j} > p_0$, where $p^{i,j}$ is the probability of the sign coincidence of the random variables associated with nodes $i$ and $j$.

In practice we are given a sample of observations $x(1), x(2), \ldots, x(t)$ from $X$. The identification of the reference market graph from observations is called in this paper the market graph identification problem. Two types of losses are possible for the identification procedure. The loss from Type I error occurs if the identification procedure includes an edge in the market graph when it is absent in the reference market graph. The loss from Type II error occurs if the identification procedure does not include an edge in the market graph when it is present in the reference market graph. For the market graph identification problem it is important to minimize the expectation of the total loss from errors.

### 3 Multiple decision framework

We model observations as a family of random vectors

$$X(t) = (X_1(t), X_2(t), \ldots, X_N(t)), \quad t = 1, 2, \ldots, n$$

where $n$ is the number of observations (sample size) and vectors $X(t)$ are independent and identically distributed as $X = (X_1, X_2, \ldots, X_N)$. Henceforth we assume that expectations $E(X_i), i = 1, 2, \ldots, N$ are known. We put (for simplicity) $E(X_i) = 0, i = 1, 2, \ldots, N$. In this case

$$p^{i,j} = P \left( X_i X_j > 0 \right), \quad i, j = 1, 2, \ldots, N$$

The random vector $X$ with measures of association $(p^{i,j})$ defines the sign similarity network. For a given threshold $p_0$, the reference market graph is defined by its adjacency matrix $TG = (tg_{i,j})$, where $tg_{i,j} = 0$ if $p^{i,j} \leq p_0$ and $tg_{i,j} = 1$ if $p^{i,j} > p_0$, $tg_{i,i} = 0$, $i, j = 1, 2, \ldots, N$.

Let $x_i(t)$ be observations of the random variables $X_i(t), t = 1, 2, \ldots, n, i = 1, 2, \ldots, N$. Define the $N \times n$ matrix $x = (x_i(t))$. Consider the set $G$ of all $N \times N$ symmetric matrices $G = (g_{i,j})$ with $g_{i,j} \in \{0, 1\}, i, j = 1, 2, \ldots, N$. $g_{i,i} = 0, i = 1, 2, \ldots, N$. Matrices $G \in G$ represent adjacency matrices of all simple undirected graphs with $N$ vertices. The total number of matrices in $G$ is equal to $L = 2^M$ with $M = N(N-1)/2$. The market graph identification problem in the sign similarity network can be formulated as a multiple decision problem of selection of one from the set of $L$ hypotheses:

$$H_G : p^{i,j} \leq p_0, \text{ if } g_{i,j} = 0, \quad p^{i,j} > p_0, \text{ if } g_{i,j} = 1; \quad i \neq j$$
Consider some examples of hypotheses (3). For the matrix
\[
G_1 = \begin{pmatrix}
0 & 0 & \ldots & 0 \\
0 & 0 & \ldots & 0 \\
\vdots & \ddots & \ddots & \vdots \\
0 & 0 & \ldots & 0 \\
\end{pmatrix}
\]
the corresponding graph has \(N\) isolated vertices and the associated hypothesis \(H_{G_1}\) is
\[
H_{G_1} : p^{i,j} \leq p_0, \forall i, j = 1, \ldots, N, \ i \neq j
\]
For the matrix:
\[
G_2 = \begin{pmatrix}
0 & 1 & 0 & \ldots & 0 \\
1 & 0 & 0 & \ldots & 0 \\
0 & 0 & 0 & \ldots & 0 \\
\vdots & \ddots & \ddots & \ddots & \vdots \\
0 & 0 & 0 & \ldots & 0 \\
\end{pmatrix}
\]
the corresponding graph has only one edge \((1, 2)\) and the associated hypothesis \(H_{G_2}\) is
\[
H_{G_2} : p^{1,2} > p_0, p^{2,1} > p_0, p^{i,j} \leq p_0, \forall (i, j) \neq (1, 2), (i, j) \neq (2, 1), i \neq j
\]
For the matrix:
\[
G_3 = \begin{pmatrix}
0 & 1 & \ldots & 1 \\
1 & 0 & \ldots & 1 \\
\vdots & \ddots & \ddots & \vdots \\
1 & 1 & \ldots & 0 \\
\end{pmatrix}
\]
the corresponding graph is the complete graph and the associated hypothesis \(H_{G_3}\) is
\[
H_{G_3} : p^{i,j} > p_0, \forall i, j = 1, \ldots, N, \ i \neq j
\]
To solve the identification problem (3) one has to use multiple decision statistical procedures. A multiple decision statistical procedure \(\delta(x)\) is a map from the sample space \(R^{N \times n}\) to the decision space \(D = \{d_G, g \in \mathcal{G}\}\), where the decision \(d_G\) is the acceptance of the hypothesis \(H_G, G \in \mathcal{G}\).

### 4 Concept of optimality

In this Section we discuss the concept of optimality related to multiple decision statistical procedures. According to Wald (1950) the quality of the statistical procedure is defined by the risk function. Consider a statistical procedure \(\delta(x)\). Let \(S = (s_{i,j}), Q = (q_{i,j}), S, Q \in \mathcal{G}\). Denote by \(w(S, Q)\) the loss from the decision \(d_Q\) when the hypothesis \(H_S\) is true
\[
w(H_S; d_Q) = w(S, Q), \quad S, Q \in \mathcal{G}
\]
It is assumed that \(w(S, S) = 0, S \in \mathcal{G}\). The risk function \(Risk : \mathcal{G} \rightarrow \mathcal{R}\) is defined by
\[
Risk(S; \delta) = \sum_{Q \in \mathcal{G}} w(S, Q) P(\delta(x) = d_Q/H_S), \quad S \in \mathcal{G}
\]
where \(P(\delta(x) = d_Q/H_S)\) is the probability that the decision \(d_Q\) is taken while the true decision is \(d_S\). The problem of minimization of \(Risk\) is therefore a multiple criteria decision.
problem: the optimal procedure $\delta$ has to minimize $Risk(S, \delta)$ for any $S \in G$. In general, such a problem does not have a solution. Instead one can use Pareto optimal solutions. However, it is possible to obtain a solution if one imposes constraints on the procedures. One common constraint is the unbiasedness of the procedure. Following Lehmann and Romano (2005) we call the multiple decision procedure $\delta(x)$ $w$-unbiased if

$$\sum_{Q \in G} w(S, Q) P(\delta(x) = d_Q / H_S) \leq \sum_{Q \in G} w(S', Q) P(\delta(x) = d_Q / H_S), \ \forall S, S' \in G$$

(4)

"Thus $\delta$ is unbiased if on the average $\delta(x)$ comes closer to the correct decision than to any wrong one" (citation from Lehmann and Romano 2005, p. 13). Note that the unbiasedness depends on the loss function $w$.

For the market graph identification problem it is important to control the total loss from type I and type II errors. Let $a_{ij}$ be the loss from the false inclusion of the edge $(i, j)$ in the market graph and let $b_{ij}$ be the loss from the false non inclusion of the edge $(i, j)$ in the market graph, $i, j = 1, 2, \ldots, N; i \neq j$.

Let

$$l_{i,j}(S, Q) = \begin{cases} a_{ij}, & \text{if } s_{ij} = 0, q_{ij} = 1, \\ b_{ij}, & \text{if } s_{ij} = 1, q_{ij} = 0, \\ 0, & \text{else} \end{cases}$$

It is natural to define the total loss $w(S, Q)$ (additive loss function) as:

$$w(S, Q) = \sum_{i=1}^{N} \sum_{j=1}^{N} l_{i,j}(S, Q)$$

(5)

Thus, the total loss from the misclassification of $H_S$ is equal to the sum of losses from the misclassification of individual edges:

$$w(S, Q) = \sum_{\{i,j:s_{ij}=0; q_{ij}=1\}} a_{ij} + \sum_{\{i,j:s_{ij}=1; q_{ij}=0\}} b_{ij}$$

The risk minimization problem is therefore reduced to the risk minimization for the additive loss function (5).

5 Multiple decision procedures in the sign similarity network

In this Section we describe a class of multiple decision procedures in the sign similarity network, based on the simultaneous inference of individual edge tests. Any individual edge test can be reduced to a hypothesis testing problem:

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

(6)

Let $\varphi_{i,j}(x)$ be a test for individual hypothesis (6). More precisely, $\varphi_{i,j}(x) = 1$ means that the hypothesis $h_{ij}$ is rejected (the edge $(i, j)$ is included in the market graph), $\varphi_{i,j}(x) = 0$ means that $h_{ij}$ is accepted (the edge $(i, j)$ is not included in the market graph).

Let $\Phi(x)$ be the matrix

$$\Phi(x) = \begin{pmatrix} 1, & \varphi_{12}(x), & \ldots, & \varphi_{1N}(x) \\ \varphi_{21}(x), & 1, & \ldots, & \varphi_{2N}(x) \\ \vdots & \vdots & \ddots & \vdots \\ \varphi_{N1}(x), & \varphi_{N2}(x), & \ldots, & 1 \end{pmatrix}.$$
Any multiple decision statistical procedure \( \delta(x) \) based on the simultaneous inference of individual edge tests (6) can be written as

\[
\delta(x) = d_G, \quad \text{if } \Phi(x) = G
\] (8)

Consider now the sign similarity network. Let

\[
p_{0,0}^{i,j} = P(X_i \leq 0, X_j \leq 0), \quad p_{1,1}^{i,j} = P(X_i > 0, X_j > 0),
\]

\[
p_{1,0}^{i,j} = P(X_i > 0, X_j \leq 0), \quad p_{0,1}^{i,j} = P(X_i \leq 0, X_j > 0)
\]

One has \( p^{i,j} = p_{0,0}^{i,j} + p_{1,1}^{i,j} \). Define

\[
u_k(t) = \begin{cases} 0, & x_k(t) \leq 0 \\ 1, & x_k(t) > 0 \end{cases}
\]

\( k = 1, 2, \ldots, N \). Let us introduce the statistics

\[
T_{1,1}^{i,j} = \sum_{t=1}^{n} u_t(t) u_j(t); \quad T_{0,0}^{i,j} = \sum_{t=1}^{n} (1 - u_t(t))(1 - u_j(t));
\]

\[
V_{i,j} = T_{1,1}^{i,j} + T_{0,0}^{i,j}
\]

To construct a multiple decision procedure we use the following individual edge tests:

\[
\phi_{i,j}^{Sg}(x_i, x_j) = \begin{cases} 0, & V_{i,j} \leq c_{i,j} \\ 1, & V_{i,j} > c_{i,j} \end{cases}
\] (10)

where for a given significance level \( \alpha_{i,j} \), the constant \( c_{i,j} \) is defined as the smallest integer such that:

\[
\sum_{k=c_{i,j}}^{n} \frac{n!}{k!(n-k)!} (p_0)^k (1 - p_0)^{n-k} \leq \alpha_{i,j}
\] (11)

Let \( \Phi^{Sg}(x) \) be the matrix

\[
\Phi^{Sg}(x) = \begin{pmatrix}
1, & \phi_{12}^{Sg}(x), & \ldots, & \phi_{1N}^{Sg}(x) \\
& \phi_{21}^{Sg}(x), & 1, & \ldots, & \phi_{2N}^{Sg}(x) \\
& & \cdots & \cdots & \ldots \\
& & & \phi_{Ni}^{Sg}(x), & \phi_{N1}^{Sg}(x), & \ldots, & 1
\end{pmatrix},
\] (12)

where \( \phi_{ij}^{Sg}(x) \) is defined by (10)–(11). Now we can define our multiple decision statistical procedure for the market graph identification

\[
\delta^{Sg}(x) = d_G, \quad \text{if } \Phi^{Sg}(x) = G
\] (13)

### 6 Optimality of the decision procedure \( \delta^{Sg} \)

The main result of the paper is the following theorem.

**Theorem 1** Let the loss function \( w \) be additive, let the individual test \( \phi_{i,j} \) depend only on \( u_i(t), u_j(t) \) for any \( i, j \), and let the following symmetry conditions be satisfied

\[
p_{11}^{ij} = p_{00}^{ij}, \quad p_{10}^{ij} = p_{01}^{ij}, \quad \forall i, j = 1, 2, \ldots, N
\] (14)
Then, for the statistical procedure $\delta^{Sg}$ defined by (10)–(13) for the market graph identification in the sign similarity network, one has $\text{Risk}(S, \delta^{Sg}) \leq \text{Risk}(S, \delta)$ for any adjacency matrix $S$ and any $w$-unbiased statistical procedure $\delta$.

**Proof** We prove optimality in three steps. First, we prove that under the symmetry conditions (14), each individual test (10) is uniformly most powerful (UMP) in the class of tests based on $u_i(t)$, $u_j(t)$ for the individual hypothesis testing

$$h_{i,j} : p_{i,j} \leq p_0 \quad \text{vs} \quad k_{i,j} : p_{i,j} > p_0 \quad (15)$$

By symmetry the individual hypothesis (15) can be written as:

$$h_{i,j} : p_{00} \leq p_0 \quad \text{vs} \quad k_{i,j} : p_{i,j} > p_0 \quad (16)$$

Let $p_{00} = T_{00}^{i,j}$, $p_{01} = T_{01}^{i,j}$, $p_{10} = T_{10}^{i,j}$, $p_{11} = T_{11}^{i,j}$, $T_{00} = T_{00}^{i,j}$, $T_{01} = T_{01}^{i,j}$, $T_{10} = T_{10}^{i,j}$, $T_{11} = T_{11}^{i,j}$. One has

$$T_{11} + T_{10} + T_{01} + T_{00} = n;
$$

Symmetry implies

$$p_{00} + p_{10} = \frac{1}{2}$$

Let $t_{11}, t_{10}, t_{01}, t_{00}$ be non-negative integers with $t_{11} + t_{10} + t_{01} + t_{00} = n$ and $C = n!/(t_{11}! t_{10}! t_{01}! t_{00}!)$. One has

$$P(T_{11} = t_{11}; T_{10} = t_{10}; T_{01} = t_{01}; T_{00} = t_{00}) = C p_{00}^{t_{00}} p_{01}^{t_{01}} p_{10}^{t_{10}} p_{11}^{t_{11}} = C \exp \left\{ (t_{11} + t_{00}) \ln \frac{p_{00}}{p_{10}} \right\}$$

where $C_1 = C (1/2 - p_{00})^n$.

Then, the hypotheses (16) are equivalent to the hypotheses:

$$h'_{i,j} : \ln \left( \frac{p_{00}}{1/2 - p_{00}} \right) \leq \ln \left( \frac{p_0}{1 - p_0} \right) \quad \text{vs} \quad k'_{i,j} : \ln \left( \frac{p_{00}}{1/2 - p_{00}} \right) > \ln \left( \frac{p_0}{1 - p_0} \right) \quad (17)$$

For $p_{00} = p_0/2$ the random variable $V = T_{11} + T_{00}$ has the binomial distribution $B(n, p_0)$. Therefore, the critical value for the test (10) is defined from (11). According to Lehmann and Romano (2005), Ch. 3, corollary 3.4.1 the test (10) is uniformly most powerful (UMP) at the level $\alpha_{i,j}$ for hypothesis testing (17).

Second, we prove that the statistical procedure (13) is $w$-unbiased. For any two-decision test for hypothesis testing (15) the risk function can be written as:

$$\text{Risk} = R(s_{i,j}, \varphi_{i,j}) = \begin{cases} a_{i,j} P(\varphi_{i,j}(x) = 1/p_{i,j}) + 0 \text{ if } s_{i,j} = 0(p_{i,j} \leq p_0) \\ b_{i,j} P(\varphi_{i,j}(x) = 0/p_{i,j}) + 0 \text{ if } s_{i,j} = 1(p_{i,j} > p_0) \end{cases}$$

Note that by the general principle, the UMP test (10) is $w$-unbiased (Lehmann and Romano 2005, Ch. 4). More precisely, one has

$$a_{i,j} P(\varphi_{i,j}(x) = 1/p_{i,j}) \leq b_{i,j} P\left( \frac{\varphi_{i,j}(x)}{p_{i,j}} \right) \quad \text{if } p_{i,j} \leq p_0$$

$$a_{i,j} P\left( \frac{\varphi_{i,j}(x)}{p_{i,j}} \right) \geq b_{i,j} P\left( \frac{\varphi_{i,j}(x)}{p_{i,j}} \right) \quad \text{if } p_{i,j} > p_0$$

$\textcircled{S}$ Springer
which is equivalent to
\[ R(s_{i,j}, \varphi_{i,j}^g) \leq R(s'_{i,j}, \varphi_{i,j}^g), \quad \forall s_{i,j}, s'_{i,j} \] (18)

This relation implies
\[ P(\varphi_{i,j}^g(x) = 1/p^i,j = p_0) = \alpha_{i,j} = \frac{b_{i,j}}{a_{i,j} + b_{i,j}} \]

This implies that the test \( \varphi_{i,j}^g \) has significance level \( \alpha_{i,j} = b_{i,j}/(a_{i,j} + b_{i,j}) \).

For the loss function (5) and any multiple decision statistical procedure \( \delta \) one has
\[
R(H_S, \delta) = \sum_{Q \in \mathcal{G}} \left( \sum_{i,j:s_{i,j} = 0; q_{i,j} = 1} a_{i,j} + \sum_{i,j:s_{i,j} = 1; q_{i,j} = 0} b_{i,j} \right) P(x \in D_Q/H_S)
\]
\[
= \sum_{s_{i,j} = 0} a_{i,j} P(\varphi_{i,j}(x) = 1/H_S) + \sum_{s_{i,j} = 1} b_{i,j} P(\varphi_{i,j}(x) = 0/H_S) \] (19)

Therefore:
\[
R(H_S, \delta) = \sum_{i=1}^{N} \sum_{j=1}^{N} R(s_{i,j}; \varphi_{i,j}) \] (20)

From (18) one has
\[
\sum_{Q \in \mathcal{G}} w(S, Q) P(\delta^S_S(x) = d_Q/H_S) \leq \sum_{Q \in \mathcal{G}} w(S', Q) P(\delta^S_S(x) = d_Q/H_S), \quad \forall S, S' \in \mathcal{G} \] (21)

Thus, the multiple testing statistical procedure \( \delta_S^S \) is unbiased.

Third, we prove that the procedure (13) is optimal in the class of unbiased statistical procedures for the market graph identification in the sign similarity network. Let \( \delta(x) \) be another unbiased statistical procedure for the market graph identification in the sign similarity network. Then \( \delta(x) \) generates a partition of the sample space \( R^{N \times n} \) into \( L \) parts:
\[
D_G = \left\{ x \in R^{N \times n} : \delta(x) = G \right\}; \quad \bigcup_{G \in \mathcal{G}} D_G = R^{N \times n}
\]

Define
\[
A_{i,j} = \bigcup_{G: \varphi_{i,j}(x) = 0} D_G
\]
\[
\overline{A}_{i,j} = \bigcup_{G: \varphi_{i,j}(x) = 1} D_G \] (22)

and
\[
\varphi_{i,j}(x) = \begin{cases} 0, & x \in A_{i,j} \\ 1, & x \notin A_{i,j} \end{cases} \] (23)

The tests (23) are the tests for individual hypotheses testing (15). Since the procedure \( \delta(x) \) is unbiased, one has
\[
\sum_{Q \in \mathcal{G}} w(S, Q) P(\delta(x) = d_Q/H_S) \leq \sum_{Q \in \mathcal{G}} w(S', Q) P(\delta(x) = d_Q/H_S), \quad \forall S, S' \in \mathcal{G}
\]
Consider the hypotheses $H_S$ and $H'_S$ which differ only in two components $s_{i,j} \neq s'_{i,j}$; $s'_{j,i} \neq s_{j,i}$. Taking into account the unbiasedness of the procedure $\delta$ and the structure of the loss function (5) one has $R(s_{i,j}, \varphi_{i,j}) \leq R(s'_{i,j}, \varphi_{i,j})$. This means that two decision tests (23) are unbiased. Therefore

$$P(\varphi_{i,j} = 1/p_0) = \alpha_{i,j} = \frac{b_{i,j}}{a_{i,j} + b_{i,j}}.$$ 

Since we are restricted to the tests based only on $u_i(t)$, $u_j(t)$ and the test $\varphi_{i,j}^{Sg}$ is UMP among tests of this class at the significance level $\alpha_{i,j}$, for any test $\varphi_{i,j}$ based only on $u_i(t)$, $u_j(t)$ one has:

$$R(s_{i,j}, \varphi_{i,j}^{Sg}) \leq R(s_{i,j}, \varphi_{i,j}).$$

From (20) one has

$$R(H_S, \delta^{Sg}) \leq R(H_S, \delta)$$

for any adjacency matrix $S$. The optimality of the multiple testing statistical procedure $\delta^{Sg}$ has been proved.

\[\square\]

## 7 Numerical comparison of identification procedures in sign similarity and Pearson correlation networks

In this Section we compare the behavior of risk functions for two multiple decision procedures, i.e. the identification procedure in the sign similarity network (10)–(13) and the popular identification procedure in the Pearson correlation network. The optimality of the identification procedure in the Pearson correlation network with the Gaussian distribution was studied in Koldanov et al. (2013). This procedure can be presented as follows. Define sample Pearson correlations

$$r_{ij} = \frac{\sum_t x_i(t)x_j(t)}{\sqrt{\sum_t x_i(t)^2 \sum_t x_j(t)^2}}$$

Let $\rho_0$ be the threshold. We use the following individual edge tests:

$$\varphi_{i,j}^P(x_i, x_j) = \begin{cases} 0, & z_{i,j} \leq c_{i,j} \\ 1, & z_{i,j} > c_{i,j} \end{cases} \quad (24)$$

where $z_{i,j}$ is the Fisher transformation of $r_{i,j}$

$$z_{i,j} = \sqrt{n} \left( \frac{1}{2} \ln \left( \frac{1 + r_{i,j}}{1 - r_{i,j}} \right) - \frac{1}{2} \ln \left( \frac{1 + \rho_0}{1 - \rho_0} \right) \right)$$

$c_{i,j}$ is the $(1 - \alpha_{i,j})$-quantile of the standard normal distribution $N(0, 1)$, $\alpha_{i,j}$ is the given significance level.

Let $\Phi^P(x)$ be the matrix

$$\Phi^P(x) = \begin{pmatrix} 1, & \varphi_{12}^P(x), \ldots, \varphi_{1N}^P(x) \\ \varphi_{21}^P(x), & 1, \ldots, \varphi_{2N}^P(x) \\ \vdots & \vdots & \ddots & \vdots \\ \varphi_{N1}^P(x), \varphi_{N2}^P(x), \ldots, 1 \end{pmatrix}. \quad (25)$$
where $\phi_{ij}^P(x)$ is defined by (24). Define the following multiple statistical procedure

$$\delta^P(x) = d_G, \text{ iff } \Phi^P(x) = G \quad (26)$$

It is proved in Koldanov et al. (2013) that $\delta^P$ is optimal in the Pearson correlation Gaussian network in the class of $w$-unbiased procedures, which have individual edge $(i, j)$ tests based on observations $x_i(t), x_j(t) : i, j = 1, \ldots, N$.

To compare the risk functions, we use two types of multivariate distributions, the multivariate Gaussian and the Student distributions. The multivariate Gaussian distribution is a popular model for stock returns, and the multivariate Student distribution represents the case of stock returns distribution with heavy tails. To make a correct comparison we use three types of correlation matrices:

1. $\Sigma_1 = \text{diag}(1, 1, \ldots, 1)$. Random variables $X_i, X_j$ are not correlated.
2. $\Sigma_2$ is a correlations matrix taken from the real stock market.
3. $\Sigma_3 = (\sigma_{i,j})$, with $\sigma_{i,i} = 1, \sigma_{i,j} = 0.9, i \neq j$. Random variables $X_i, X_j$ are highly correlated.

We choose the following parameter values: $N = 30, n = 400$, the significance level for all individual tests $\alpha = 0.5, \alpha = 0.1$. The matrix $\Sigma_2$ is calculated from the Dow-Jones index stocks for the year 2013.

The most interesting is the behavior of risk function as a function of the threshold. For the Pearson correlation network, the value of the threshold $\rho_0$ belongs to the interval $[-1, 1]$. For the sign similarity network, the value of the threshold $p_0$ belong to the interval $[0, 1]$. To make a correct comparison we use the following transformation formula from the Pearson correlation to the probability of sign coincidence:

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

This formula is known for the Gaussian distribution (Kramer 1962, Ch. 21), and it can be proved for the Student distribution, too Kalyagin et al. (2017).

The results of numerical experiments are presented in Figs. 1, 2, 3, 4, 5 and 6. Figures 1 and 2 present the behavior of the risk function of $\delta^{Sg}$ and $\delta^P$ as a function of the threshold for the correlation matrix $\Sigma_1$, and $\alpha = 0.5$ (Fig. 1), $\alpha = 0.1$ (Fig. 2).

One can see that both procedures control the risk function for the Gaussian distribution. Namely, with the change of $\alpha$ from 0.5 to 0.1 the maximum value of the risk function for

![Fig. 1 Risk functions $\text{Risk}(p_0)$ for the matrix $\Sigma_1$ and $\alpha = 0.5$. Left Gaussian distribution. Right Student distribution. Solid line Pearson correlation network. Dashed line sign similarity network. The horizontal axe represents the value of $p_0$.](image)
both procedures drops approximately from 120 to 40. In contrast, for the Student distribution the maximum value of the risk function for the procedure $\delta^S_g$ drops, but the maximum value of the risk function for the procedure $\delta^P$ stays the same. This phenomenon is confirmed for the matrix $\Sigma_2$ by comparing Figs. 3 and 4 and for the matrix $\Sigma_3$ by comparing Figs. 5 and 6. These results show that the procedure $\delta^S_g$ controls the risk function for both distributions
Fig. 5 Risk functions $\text{Risk}(p_0)$ for the matrix $\Sigma_3$ and $\alpha = 0.5$. Left Gaussian distribution. Right Student distribution. Solid line Pearson correlation network. Dashed line sign similarity network. The horizontal axe represents the value of $p_0$.

Fig. 6 Risk functions $\text{Risk}(p_0)$ for the matrix $\Sigma_3$ and $\alpha = 0.1$. Left Gaussian distribution. Right Student distribution. Solid line Pearson correlation network. Dashed line sign similarity network. The horizontal axe represents the value of $p_0$.

while the procedure $\delta^p$ does not. This gives an advantage to the procedure $\delta^{SS}$ for multivariate Student distributions for a small value of $\alpha$.

8 Concluding remarks

In this paper we introduced and studied a class of statistical procedures with a high reliability for the market graph identification in the sign similarity network. A theoretical study was conducted within the framework of multiple decision theory. The optimality of the multiple decision procedure $\delta^{SS}$ was proved under the following assumptions: additivity of loss functions, unbiasedness of procedures, sign symmetry conditions and known expectations $E(X_i), j = 1, 2, \ldots, N$. Additivity of the loss function and unbiasedness of procedures are appropriate for the reviewed problems. Sign symmetry conditions are satisfied for a large class of distributions used in financial analyses, in particular for elliptically contoured distributions. This class includes multivariate Gaussian and Student distributions with heavy tails. The practical advantage of the constructed procedures with respect to traditional ones is the high reliability of identification in a larger class of distributions.

Sign similarity networks constitute a promising model for performing stock market data mining. Sign similarity is easy to interpret. It is connected with Pearson correlation, and statis-
tical procedures based on sign similarity are distribution-free in the large class of elliptically contoured distributions (Kalyagin et al. 2017). It is known that this class of distributions is appropriate for portfolio optimization (Gupta et al. 2013). Likelihood robust optimization for portfolio selection is discussed in Wang et al. (2016). The authors use a discretization of the return distributions for their optimization. The main problem is that the resulting algorithm is computationally expensive. Numerical experiments show that the use of the sign of return distributions can significantly reduce the computational complexity and can make the algorithm practical. This can become a subject for further research.

Acknowledgements The work has been conducted at the Laboratory of Algorithms and Technologies for Network Analysis of the National Research University Higher School of Economics. V. A. Kalyagin and A. P. Koldanov are partially supported by RFFI Grant 14-01-00807, and P. A. Koldanov is partially supported by RFHR Grant 15-32-01052.

References

Bautin, G. A., Kalyagin, V. A., & Koldanov, A. P. (2013). Comparative analysis of two similarity measures for the market graph construction. In Proceedings in mathematics and statistics (Vol. 59, pp. 29–41). Springer.

Bautin, G. A., Kalyagin, V. A., Koldanov, A. P., Koldanov, P. A., & Pardalos, P. M. (2013). Simple measure of similarity for the market graph construction. Computational Management Science, 10, 105–124.

Boginsky, V., Butenko, S., & Pardalos, P. M. (2003). On structural properties of the market graph. In A. Nagurney (Ed.), Innovations in financial and economic networks (pp. 29–45). Northampton: Edward Elgar Publishing Inc.

Boginsky, V., Butenko, S., & Pardalos, P. M. (2004). Network model of massive data sets. Computer Science and Information Systems, 1, 75–89.

Boginsky, V., Butenko, S., & Pardalos, P. M. (2005). Statistical analysis of financial networks. Journal of Computational Statistics and Data Analysis, 48(2), 431–443.

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

Boginski, V., Butenko, S., Shirokikh, O., Trukhanov, S., & Lafuente, J. G. (2014). A network-based data mining approach to portfolio selection via weighted clique relaxations. Annals of Operations Research, 216, 23–34.

Cesarone, F., Scozzari, A., & Tardella, F. (2015). A new method for mean-variance portfolio optimization with cardinality constraints. Annals of Operations Research, 215, 213–234.

Emmert-Streib, F., & Dehmer, M. (2010). Identifying critical financial networks of the DJIA: Towards a network based index. Complexity, 16(1), 24–33.

Garas, F., & Argyakis, P. (2007). Correlation study of the Athens stock exchange. Physica A, 380, 399–410.

Gunawardena, A. D. A., Meyer, R. R., Dougan, W. L., Monaghan, P. E., & ChotonBasu, P. E. M. (2012). Optimal selection of an independent set of cliques in a market graph. In: International proceedings of economics development and research (Vol. 29, p. 281285).

Gupta, F. K., Varga, T., & Bodnar, T. (2013). Elliptically contoured models in statistics and portfolio theory. New York: Springer.

Hero, A., & Rajaratnam, B. (2012). Hub discovery in partial correlation graphs. IEEE Transactions on Information Theory, 58(9), 6064–6078.

Huang, W. Q., Zhuang, X. T., & Yao, S. A. (2009). A network analysis of the Chinese stock market. Physica A, 388, 2956–2964.

Huffner, F., Komusiewicz, C., Moser, H., & Niedermeier, R. (2008). Enumerating isolated cliques in synthetic and financial networks. In Combinatorial optimization and applications, lecture notes in computer science (Vol. 5165, pp. 405–416).

Kalyagin, V. A., Koldanov, A. P., & Koldanov, P. A. (2017). Robust identification in random variables networks. Journal of Statistical Planning and Inference, 181(2017), 30–40.

Kenett, D. Y., Tumminello, M., Madi, A., Gur-Gershgoren, G., Mantegna, R. N., & Ben-Jacob, E. (2010). Dominating clasp of the financial sector revealed by partial correlation analysis of the stock market. PLoS ONE, 5(12), e15032. doi:10.1371/journal.pone.0015032.
Koldanov, A. P., Koldanov, P. A., Kalyagin, V. A., & Pardalos, P. M. (2013). Statistical procedures for the market graph construction. *Computational Statistics and Data Analysis, 68*, 17–29.

Kramer, H. (1962). *Mathematical methods of statistics* (9th ed.). Princeton: Princeton University Press.

Lehmann, E. L. (1957). A theory of some multiple decision procedures 1. *Annals of Mathematical Statistics, 28*, 1–25.

Lehmann, E. L., & Romano, J. P. (2005). *Testing statistical hypotheses*. New York: Springer.

Mantegna, R. N. (1999). Hierarchical structure in financial markets. *European Physical Journal, Series B, 11*, 93–97.

Namaki, A., Shirazi, A. H., Raei, R., & Jafari, G. R. (2011). Network analysis of a financial market based on genuine correlation and threshold method. *Physica A, 390*, 3835–3841.

Onella, J.-P., Kaski, K., & Kertesz, J. (2004). Clustering and information in correlation based financial networks. *The European Physical Journal B-Condensed Matter and Complex Systems, 38*(2), 353–362.

Shirokikh, J., Pastukhov, G., Boginski, V., & Butenko, S. (2013). Computational study of the US stock market evolution: A rank correlation-based network model. *Computational Management Science, 10*(2–3), 81–103.

Tse, C. K., Liu, J., & Lau, F. C. M. (2010). A network perspective of the stock market. *Journal of Empirical Finance, 17*, 659–667.

Tumminello, M., Coronello, C., Lillo, F., Micciche, S., & Mantegna, R. (2007). Spanning trees and bootstrap reliability estimation in correlation-based network. *International Journal of Bifurcation and Chaos, 17*, 2319–2329.

Vizgunov, A. N., Goldengorin, B., Kalyagin, V. A., Koldanov, A. P., Koldanov, P. A., & Pardalos, P. M. (2014). Network approach for the Russian stock market. *Computational Management Science, 11*, 45–55.

Wald, A. (1950). *Statistical decision function*. New York: Wiley.

Wang, G. J., Chi, X., Han, F., & Sun, B. (2012). Similarity measure and topology evolution of foreign exchange markets using dynamic time warping method: Evidence from minimal spanning tree. *Physica A: Statistical Mechanics and its Applications, 391*(16), 4136–4146.

Wang, Z., Glynn, P. W., & Ye, Y. (2016). Likelihood robust optimization for data-driven problems. *Computational Management Science, 13*, 241–261.