УЧЕНЫЕ ЗАПИСКИ КАЗАНСКОГО УНИВЕРСИТЕТА.
СЕРИЯ ФИЗИКО-МАТЕМАТИЧЕСКИЕ НАУКИ

2018, Т. 160, кн. 2
С. 317–326

ISSN 2541-7746 (Print)
ISSN 2500-2198 (Online)

UDK 519.2

RISK FUNCTION AND OPTIMALITY OF STATISTICAL PROCEDURES FOR IDENTIFICATION OF NETWORK STRUCTURES

P.A. Koldanov

National Research University Higher School of Economics, Nizhny Novgorod, 603025 Russia

Abstract

Identification of network structures using the finite-size sample has been considered. The concepts of random variables network and network model, which is a complete weighted graph, have been introduced. Two types of network structures have been investigated: network structures with an arbitrary number of elements and network structures with a fixed number of elements of the network model. The problem of identification of network structures has been investigated as a multiple testing problem. The risk function of statistical procedures for identification of network structures can be represented as a linear combination of expected numbers of incorrectly included elements and incorrectly non-included elements. The sufficient conditions of optimality for statistical procedures for network structures identification with an arbitrary number of elements have been given. The concept of statistical uncertainty of statistical procedures for identification of network structures has been introduced.

Keywords: random variables network, network model, network structure, procedure for identification of network structure, additive loss function, risk function, unbiasedness, optimality, statistical uncertainty

Introduction

One approach to analyze a complex system with N elements is to consider the corresponding network model, which can be visualized as a complete weighted graph with N nodes [1]. Network model can be represented as a complete weighted graph $G = (V, E, \gamma)$, where nodes $V = \{1, 2, \ldots, N\}$ correspond to the elements of the system and weights $\gamma_{i,j}$ of edges $e_{i,j} \in E$ are given by measure γ of relation (dependence, association) between elements. In this paper, we focus on probabilistic networks models only. In probabilistic network models, nodes correspond to random variables. The Gaussian graphical model is a well-developed probabilistic network model [2]. Statistical procedures for selection (identification) of the Gaussian graphical model by observations were studied in [3–5]. The weak point of statistical procedures proposed in these works is control of type I errors only.

Another probabilistic network model is the network model of financial market. Every node of the network model corresponds to stock, and the weights of edges are given by the selected measure of dependence between stock returns. For financial market, the popular network structures are threshold graph [6] and maximum spanning tree [7].

The threshold graph is unweighted graph obtained from the network model by removing edges with weights less than or equal to the given threshold. The maximum spanning tree is a spanning tree of network model with the maximum sum of edges weights. There are many publications on calculation of such network structures and interpretation of obtained results. The statistical approach to threshold graph identification is proposed in [8].

The general problem statement of network structures identification is considered in the paper. The general approach to develop statistical procedures for network structure identification is discussed. The natural quality characteristic of these procedures is mean numbers of first- and second-kind errors, respectively. Two types of network structures identification problems are introduced: problems of network structure identification with an arbitrary number of elements from the network model; problems of network structure identification with a fixed number of elements from the network model. An example of the network structure identification problem with an arbitrary number of elements from the network model is the problem of threshold graph identification. An example of the network structure identification problem with a fixed number of elements from the network model is the problem of MST identification. It is shown in [9] that the risk function of statistical procedures for network structures identification of both types can be represented as a sum of mean numbers of first- and second-kind errors. In the paper, the sufficient conditions of optimality for statistical procedures for network structures identification with an arbitrary number of elements are given. The concept of statistical uncertainty of statistical procedures for network structures identification is introduced.

1. Basic definitions and problem statement

Let $X = (X_1, X_2, \ldots, X_N)$ be a random vector. It is assumed that density $f(x)$ of the vector X belongs to class $\{f(x,\theta); \theta \in \Omega\}$, where Ω is a parametric space. The partition of parametric space Ω by L regions $\Omega_i : i = 1, \ldots, L; \Omega_i \cap \Omega_j = \emptyset, i \neq j$ is defined and hypotheses $H_i : \theta \in \Omega_i, \Omega_i \subset \Omega, i = 1, \ldots, L$ are formulated. There is finite-size sample $x(1), x(2), \ldots, x(n)$ from sample space $\mathcal{X} = R^{N \times n}$.

The general problem is: *to construct the statistical procedure $\delta(x)$, which defines the partition of sample space $\mathcal{X}$ by L part $\mathcal{X} \to D, D = \{D_1, D_2, \ldots, D_L\}$*. Decision d_i : hypothesis H_i is true is accepted if $(x(1), x(2), \ldots, x(n)) \in D_i$.

In order to formulate the problems of network structures identification, the concept of random variables network is introduced.

Definition 1. The random variables network is a pair (X, γ), where $X = (X_1, \ldots, X_N)$ is a random vector and $\gamma = \{\gamma_{i,j} : i, j = 1, \ldots, N; i \neq j\}$ is a measure of dependence between random variables X_i, X_j.

The random variables network generates a network model which is complete weighted graph $G = (V, E, \gamma)$, where $V = \{1, 2, \ldots, N\}$ is a set of nodes corresponding to the random variables $X_1, X_2, \ldots, X_N$, and E is a set of edges with weights given by measure γ. In order to investigate the network model $G = (V, E, \gamma)$, it is clear that key structures of the corresponding graph should be identified.

The key structures satisfying the following definition are investigated in the paper.

Definition 2. The network structure of network model $G = (V, E, \gamma)$ is unweighted subgraph $G' = (V', E') : V' \subseteq V, E' \subseteq E$.

Two types of network structures are considered. The first type of network structures is that one with any number of elements from the network model. The threshold graph and the Gaussian graphical model are network structures of the first type.

Definition 3. The threshold graph (TG) of network model $G = (V, E, \gamma)$ is subgraph $G'(\gamma_0) = (V', E') : V' = V; E' \subseteq E, E' = \{(i,j) : \gamma_{i,j} > \gamma_0\}$, where γ_0 is some threshold.

The second type of network structures includes those of them with a fixed number of elements from the network model. The maximum spanning tree is a network structure

of the second type, because the maximum spanning tree must contain $N-1$ edges exactly.

Definition 4. Maximum spanning tree (MST) of network model $G=(V,E,\gamma)$ is a tree $G'=(V',E'):V'=V;\ E'\subset E;\ |E'|=|V|-1;$, such that $\sum\limits_{(i,j)\in E'}\gamma_{i,j}$ is maximum.

To provide more details, let us propose the following general formulation of the problem of network structures identification.

Let (X,γ) be a random variable network. Let the density of random vector X belong to $f(x)\in\{f(x,\theta):\theta\in\Omega\}$. Let $G=(V,E,\gamma)$ be a network model generated by random variable network (X,γ). Let $\beta\in E$ $(\beta=1,\dots,K,\ K=N(N-1)/2)$ be elements (edges) of network model $G=(V,E,\gamma)$. Let $G'=(V',E'):V'\subseteq V,\ E'\subseteq E$ be the network structure of interest, which must be defined by observations $x_i(t), i=1,\dots,N,\ t=1,\dots,n$. Let $h_\beta:\theta\in\omega_\beta$ be the hypothesis that element β of the network model does not belong to the network structure, $k_\beta:\theta\in\omega_\beta^{-1}$ be the alternative to h_β, $H_i:\theta\in\Omega_i; i=1,\dots,L$ be the hypothesis that elements $\{i_1,i_2,\dots,i_M\}$, $\{i_1,i_2,\dots,i_M\}\subseteq\{1,2,\dots,K\}$ belong to the network structure. Let M be the number of elements of the network structure. It is necessary to construct a statistical procedure to select one from the set of disjoint hypotheses:

$$\begin{aligned}&H_i:\theta\in\Omega_i,\\ &\text{where}\\ &\Omega_i=\Big(\bigcap_{i_l\in\{i_1,\dots,i_M\}}\omega_{i_l}^{-1}\Big)\bigcap\Big(\bigcap_{i_s\in\{1,\dots,K\}-\{i_1,\dots,i_M\}}\omega_{i_s}\Big)\\ &\text{or}\\ &H_i=\Big(\bigcap_{i_l\in\{i_1,\dots,i_M\}}k_{i_l}\Big)\bigcap\Big(\bigcap_{i_s\in\{1,\dots,K\}-\{i_1,\dots,i_M\}}h_{i_s}\Big).\end{aligned}\tag{1}$$

Depending on M, there are two types of problems:

- problems with an arbitrary number of elements of the network model $M\in\{0,1,\dots,C_N^2\}$
- problem with a fixed number M of elements of the network model

2. Statistical procedures for network structure identification

Let $\varphi_\beta(x)$ be the tests for testing individual hypotheses h_β versus k_β. Let A_β be the acceptance region of test $\varphi_\beta(x)$ and A_β^{-1} be the rejection region of test $\varphi_\beta(x)$, respectively. Let $\delta(x)$ be the statistical procedure for problem (1), where d_i is the decision that hypothesis $H_i, i=1,\dots,L$ is true, and D_i be the acceptance region of hypothesis H_i

$$\begin{aligned}&\delta(x)=d_i,\ \text{if } x\in D_i,\\ &D_i\bigcap D_j=\emptyset,\quad i\neq j,\quad i,j=1,\dots,L;\quad \bigcup_{i=1}^{L}D_i=\mathcal{X},\\ &\text{where } \mathcal{X} \text{ is a sample space.}\end{aligned}\tag{2}$$

According to the results of [10], any procedure for network structure identification with an arbitrary number of elements from the network model can be written in the following form:

$$D_i=\bigcap_{\beta=1}^{K}A_\beta^{\kappa_{i\beta}},\tag{3}$$

where

$$\kappa_{i\beta} = \begin{cases} 1, & \Omega_i \bigcap \omega_\beta \neq \emptyset, \\ -1, & \Omega_i \bigcap \omega_\beta = \emptyset. \end{cases} \tag{4}$$

For statistical procedures for network structure identification with fixed number M of elements from the network model, the condition of compatibility must be satisfied, which can be written as:

Definition 5. Set of tests $\varphi_\beta(x), \beta = 1, \ldots, M$ is compatible with decision space of procedure $\delta(x)$ (2) if

$$\sum_{\substack{(\kappa_{i\beta_{i_1}}, \ldots, \kappa_{i\beta_{i_K}}): \\ \kappa_{i\beta_{i_1}} = \ldots = \kappa_{i\beta_{i_M}} = -1; \\ \kappa_{i\beta_{i_{M+1}}} = \ldots = \kappa_{i\beta_{i_K}} = 1}} P(x \in \bigcap_\beta A_\beta^{\kappa_{i\beta}}) = 1. \tag{5}$$

If the set of tests $\varphi_\beta(x), \beta = 1, \ldots, M$ is compatible with the decision space of procedure $\delta(x)$, then there is one-to-one correspondence between procedure $\delta(x)$ (2) and the set of tests $\varphi_\beta(x), \beta = 1, \ldots, M$ [10]. Such correspondence has the form:

$$D_i = \bigcap_{\beta=1}^{K} A_\beta^{\kappa_{i\beta}}, \quad A_\beta = \bigcup_{i:\kappa_{i\beta}=1} D_i, \quad A_\beta^{-1} = \bigcup_{i:\kappa_{i\beta}=-1} D_i \tag{6}$$

In the case of compatible set of tests $\varphi_\beta(x)$, relations (6) define the statistical procedures for network structure identification.

3. Risk function of statistical procedures for network structure identification

Let $w(H_i; d_j) = w_{ij}$ be the loss from decision d_j when hypothesis H_i is true. Let us assume that the loss from the correct decision is equal to zero, $w_{ii} = 0 \ \forall\, i = 1, \ldots, L$. According to [11], the quality of any statistical procedure $\delta(x)$ is characterized by the risk function

$$R(H_i, \theta; \delta) = \sum_{j=1}^{L} w_{ij} P_\theta(\delta(x) = d_j), \quad \theta \in \Omega_i, \quad i = 1, \ldots, L,$$

where $P_\theta(\delta(x) = d_j)$ is the probability of decision d_j.

Let a_β, b_β be the loss from the first- and second-kind errors for testing of individual hypotheses h_β. Consider loss function w_{ij} of the following form

$$w_{ij} = \sum_\beta (\epsilon_{ij\beta} a_\beta + \epsilon_{ji\beta} b_\beta), \tag{7}$$

where

$$\epsilon_{ij\beta} = \begin{cases} 1, & \text{if } \ \kappa_{i\beta} = 1, \quad \kappa_{j\beta} = -1, \\ 0, & \text{otherwise,} \end{cases}$$

$\kappa_{i\beta}$ defined by (4).

The following theorems [9] characterize the risk function for the problem of network structure identification.

Theorem 1. *Let the loss function be defined by (7). Then the risk function of the statistical procedure for the problem of identification of the network structure with an arbitrary number of elements is:*

$$R(H_i, \theta, \delta) = \sum_{\beta=1}^{K} r(h_\beta, \varphi_\beta), \tag{8}$$

where $r(h_\beta, \varphi_\beta)$ *is the loss function of test* φ_β.

In the case $a_\beta = a, b_\beta = b, \forall \beta = 1, \ldots, K$, *one has:*

$$R(H_i, \theta, \delta) = aE_\theta\{Y_I(H_i, \delta)\} + bE_\theta\{Y_{II}(H_i, \delta)\}, \tag{9}$$

where $Y_I(H_i, \delta)$ *is the number of erroneously included elements (the number of first-kind errors) by procedure* δ *if hypothesis* H_i *is true,* $Y_{II}(H_i, \delta)$ *is the number or erroneously non-included elements (the number of second-kind errors) by procedure* δ *if hypothesis* H_i *is true.*

Theorem 2. *Let*

- *the set of tests* φ_β *for testing individual hypotheses* h_β *be compatible with the decision space of statistical procedure* δ *for testing hypotheses* H_i;
- *the loss function be additive and defined by (7). Then the risk function of statistical procedure* δ *for the problem of identification of the network structure with a fixed number of elements has the form:*

$$R(H_i, \theta, \delta) = \sum_{\beta=1}^{K} r(h_\beta, \varphi_\beta), \tag{10}$$

where $r(h_\beta, \varphi_\beta)$ *is the risk function of test* φ_β.

- *If* $a_\beta = a,\ b_\beta = b,\ \beta = 1, \ldots, K$ *then the risk function of statistical procedure* δ *for the problem of identification of the network structure with a fixed number of elements has the form:*

$$R(H_i, \theta, \delta) = (a+b)E_\theta(Y_I(H_i, \delta)) = (a+b)E_\theta(Y_{II}(H_i, \delta)), \tag{11}$$

where $Y_I(H_i, \delta)$ *is the number of first-kind errors,* $Y_{II}(H_i, \delta)$ *is the number of second-kind errors of procedure* δ *when hypothesis* H_i *is true.*

Note that theorem 1 is a simple result of [10]. On the other hand, theorem 2 is new and corresponds to a generalization of the result of [10].

4. Sufficient conditions to optimality of statistical procedure for identification of network structure with arbitrary number of elements

Consider the set $\mathcal{G}$ 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 $G \in \mathcal{G}$ represent the adjacency matrices of all simple undirected graphs with N nodes. The total number of matrices in $\mathcal{G}$ is equal to $L = 2^M$ with $M = N(N-1)/2$. The problem of identification of the network structure with an arbitrary number of elements can be formulated as a multiple decision problem of the selection of one hypothesis from the set of L hypotheses:

$$H_G : \gamma_{ij} \le \gamma_0, \quad \text{if } g_{i,j} = 0, \quad \gamma_{ij} > \gamma_0, \quad \text{if } g_{i,j} = 1; \quad i \ne j. \tag{12}$$

Let $\beta = (i, j)$. Let individual tests for individual edge hypotheses:

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

have the form:

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

where c_{ij} is defined from:

$$P_{\gamma_0}(T_{ij} > c_{ij}) = \alpha_{ij} \tag{13}$$

and α_{ij} is the given significance level.

According to (3), the multiple statistical procedure for identification of the network structure with an arbitrary number of elements has the form

$$\Phi(x) = \begin{pmatrix} 1, & \varphi_{12}(x), & \dots, & \varphi_{1N}(x) \\ \varphi_{21}(x), & 1, & \dots, & \varphi_{2N}(x) \\ \dots & \dots & \dots & \dots \\ \varphi_{N1}(x), & \varphi_{N2}(x), & \dots, & 1 \end{pmatrix}. \tag{14}$$

Let us define the multiple statistical procedure for network structure identification

$$\delta(x) = d_G, \quad \text{iff} \quad \Phi(x) = G. \tag{15}$$

Let $S = (s_{i,j})$, $Q = (q_{i,j})$, $S, Q \in \mathcal{G}$. Denote by $w(S, Q)$ the loss from decision d_Q when hypothesis H_S is true

$$w(H_S; d_Q) = w(S, Q), \quad S, Q \in \mathcal{G}.$$

The risk function is defined by

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

where $P_\theta(\delta(x) = d_Q$ is the probability that decision d_Q is taken, while the true decision is $d_S : \theta \in \Omega_S$, Ω_S with $\theta = ||\gamma_{ij}||$, such that hypothesis H_S is true. According to [10], the multiple decision procedure $\delta(x)$ is w-unbiased if

$$\sum_{Q \in \mathcal{G}} w(S, Q) P_\theta(\delta(x) = d_Q) \le \sum_{Q \in \mathcal{G}} w(S', Q) P_\theta(\delta(x) = d_Q) \quad \forall\, S,\, S' \in \mathcal{G},\ \theta \in \Omega_S. \tag{16}$$

Let $a_{i,j}$ be the loss from the false inclusion of edge (i, j) in the network structure, and let $b_{i,j}$ be the loss from the false non-inclusion of edge (i, j) in the network structure, $i, j = 1, 2, \dots, N$, $i \ne j$.

Then additive loss function (7) can be written as

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

It means that the loss from the misclassification of H_S is equal to the sum of losses from the misclassification of individual edges.

Theorem 3. *Let the loss function be additive and tests $\varphi_{ij}(x)$ be uniformly most powerful in the class of unbiased (UMPU) levels α_{ij} tests. Then statistical procedure* (15) *is optimal in the class of unbiased statistical procedures for identification of the network structure with an arbitrary number of elements if* $\alpha_{ij} = \dfrac{b_{ij}}{a_{ij} + b_{ij}}$.

Proof. First, we prove that statistical procedure δ is unbiased. Individual tests $\varphi_{ij}(x)$ are unbiased, then $r(s_{i,j}, \varphi_{ij}(x)) \le r(s'_{i,j}, \varphi_{ij}(x))$ for any $s_{i,j}$, $s'_{i,j} \in \{0,1\}$, i, $j = 1, \ldots, N$.

The loss function is additive, then, according to theorem 1, the risk function of statistical procedure δ can be written as $R(H_S, \theta, \delta) = \sum\limits_{i,\,j=1}^{N} r(s_{i,j}, \varphi_{ij})$. Therefore, $\forall S$, $S' \in \mathcal{G}$, $\theta \in \Omega_S$

$$\sum_Q w(S,Q) P_\theta(\delta(x) = d_Q) \le \sum_Q w(S',Q) P_\theta(\delta(x) = d_Q).$$

Then $\delta(x)$ is unbiased.

Now we should prove that statistical procedure δ is optimal in the class of unbiased statistical procedures. Let $\delta'(x)$ be any other unbiased procedure. Then $\delta'(x)$ defines the partition of the sample space by L parts $D_G = \{x : \delta'(x) = G\}$. Let $A_{i,j} = \bigcup\limits_{G:g_{i,j}=0} D_G$, $A_{i,j}^{-1} = \bigcup\limits_{G:g_{i,j}=1} D_G$. Define

$$\varphi'_{i,j} = \begin{cases} 0, & x \in A_{i,j}, \\ 1, & x \in A_{i,j}^{-1}. \end{cases}$$

Tests $\varphi'_{i,j}$ are used to test individual hypotheses, $h_{i,j}$ – elements (i,j) do not belong to network structure S. Then, according to theorem 1, the risk function of statistical procedure δ' can be written as

$$R(H_S, \theta, \delta') = \sum_{i,j=1}^{N} r(s_{i,j}, \varphi'_{ij}).$$

Since statistical procedure $\delta'(x)$ is unbiased, then

$$\sum_Q w(S,Q) P_\theta(\delta'(x) = d_Q) \le \sum_Q w(S',Q) P_\theta(\delta'(x) = d_Q) \quad \forall\, S,\, S' \in \mathcal{G},\ \theta \in \Omega_S. \quad (17)$$

Since network structure S has an arbitrary number of elements, there exists network structure S', such that

$$\exists i, j : s_{i,j} \ne s'_{i,j}, s_{j,i} \ne s'_{j,i} \quad \forall\, (k,l) \ne (i,j), \quad (k,l) \ne (j,i), \quad s_{k,l} = s'_{k,l}.$$

Then, (17) has the form:

$$r(s_{i,j}, \varphi'_{i,j}(x)) \le r(s'_{i,j}, \varphi'_{ij}(x)).$$

Hence, tests $\varphi'_{i,j}$ are unbiased.

However, tests $\varphi_{i,j}(x)$ are UMPU, then $r(s_{i,j}, \varphi_{i,j}(x)) \le r(s_{i,j}, \varphi'_{i,j}(x))$.

Therefore, $R(H_S, \theta, \delta) \le R(H_S, \theta, \delta')$. □

Note that theorem 3 is based on the general ideas of [10]. Nevertheless, the restriction of the problem of identification of the network structure with an arbitrary number of elements allows to give a simpler proof.

Multiple testing procedures for Gaussian graphical model (GGM) identification. Let us consider random variables network (X, γ), where vector $X = (X_1, X_2, \ldots, X_N)$ has multivariate normal distribution $N(\mu, \Sigma)$ and measure $\gamma_{i,j} = |\rho^{i,j}|$ is the absolute value of partial correlation coefficient $\rho^{i,j}$.

The individual hypotheses for the problem of GGM identification have the form:

$$h_{ij} : \rho^{i,j} = 0 \quad vs \quad k_{ij} : \rho^{i,j} \neq 0. \tag{18}$$

According to [12], UMPU tests for testing individual hypotheses (18) are:

$$\varphi_{i,j}^{opt} = \begin{cases} 0, & |r^{i,j}| < 1 - 2c_{\alpha/2}^{\beta}, \\ 1, & |r^{i,j}| > 1 - 2c_{\alpha/2}^{\beta}, \end{cases} \tag{19}$$

where $c_{\alpha/2}^{\beta}$ is the $\alpha/2$-quantile of Beta distribution $\mathrm{Be}\left(\frac{n-N}{2}, \frac{n-N}{2}\right)$. Let us define the multiple statistical procedure for concentration graph identification

$$\delta^{opt}(x) = d_G, \quad \text{iff} \quad \Phi^{\mathrm{opt}}(x) = G, \tag{20}$$

where

$$\Phi^{opt}(x) = \begin{pmatrix} 0, & \varphi_{1,2}^{\mathrm{opt}}(x), & \dots, & \varphi_{1,N}^{\mathrm{opt}}(x) \\ \varphi_{2,1}^{\mathrm{opt}}(x), & 0, & \dots, & \varphi_{2,N}^{\mathrm{opt}}(x) \\ \dots & \dots & \dots & \dots \\ \varphi_{N,1}^{\mathrm{opt}}(x), & \varphi_{N,2}^{\mathrm{opt}}(x), & \dots, & 0 \end{pmatrix}. \tag{21}$$

According to theorem 3, it is easy to prove the following

Theorem 4. *Multiple-decision statistical procedure (20) is optimal in the class of unbiased statistical procedures for GGM identification under the additive loss function.*

5. Statistical uncertainty

Theorems 1, 2 allow to introduce the unique measure of uncertainty for the statistical procedures of network structures identification.

Definition 6. Value $R(S, \theta, \delta, n)$ will be called the statistical uncertainty of procedure δ for network structure S identification under n observations and distribution of vector X with $\theta \in \Omega_S$.

Definition 7. Statistical procedure δ_1 of network structure S_1 identification has a smaller statistical uncertainty for $\Omega_1 \subset \Omega$ than statistical procedure δ_2 of network structure S_2 identification if

$$R(S_1, \theta, \delta_1, n) \leq R(S_2, \theta, \delta_2, n), \forall n, \forall \theta \in \Omega_1$$

If $a = \dfrac{1}{2M_1}$, $b = \dfrac{1}{2M_2}$, where M_i is the maximum number of type i errors $(i = 1, 2)$, then the measure of statistical uncertainty is equal to the average number of erroneous decisions of procedure δ. The experimental results from [13] show that the uncertainty of the statistical procedure for threshold graph identification is much smaller than the uncertainty of the statistical procedure for maximum spanning tree identification.

Conclusions

The general approach to identification of network structures is proposed in the paper. In contrast to the known approach [4, 5], our approach allows to pay attention to both types of errors, as well as to investigate the properties of optimality and to compare different network structures by statistical uncertainty of their identification procedures.

Acknowledgments. This work was supported in part by the Laboratory of Algorithms and Technologies for Network Analysis of National Research University Higher School of Economics and by the Russian Foundation for Basic Research (project no. 18-07-00524).

References

1. Jordan M.I. Graphical models. *Stat. Sci.*, 2004, vol. 19, no. 1, pp. 140–155. doi: 10.1214/088342304000000026.
2. Lauritzen S.L. *Graphical Models*. Oxford, Oxford Univ. Press, 1996. 298 p.
3. Anderson T.W. *An Introduction to Multivariate Statistical Analysis*. New York, John Wiley & Sons, 2003. 752 p.
4. Drton M., Perlman M.D. Model selection for Gaussian concentration graph. *Biometrika*, 2004, vol. 91, no. 3, pp. 591–602. doi: 10.1093/biomet/91.3.591.
5. Drton M., Perlman M. Multiple testing and error control in Gaussian graphical model selection. *Stat. Sci.*, 2008, vol. 22, no. 3, pp. 430–449. doi: 10.1214/088342307000000113.
6. Boginski V., Butenko S., Pardalos P.M. On structural properties of the market graph. In: *Innovations in Financial and Economic Networks*. Cheltenham, Edward Elgar Publ., 2003, pp. 29–45.
7. Mantegna R.N. Hierarchical structure in financial markets. *Eur. Phys. J. B*, 1999, vol. 11, no. 1, pp. 193–197. doi: 10.1007/s100510050929.
8. Koldanov A.P., Koldanov P.A. , Kalyagin V.A., Pardalos P.M. Statistical procedures for the market graph construction. *Comput. Stat. Data Anal.*, 2013, vol. 68, pp. 17–29. doi: 10.1016/j.csda.2013.06.005.
9. Koldanov P.A. Risk function of statistical procedures for network structures identification. *Vestn. TVGU. Ser. Prikl. Mat.*, 2017, no. 3, pp. 45–59. doi: 10.26456/vtpmk178. (In Russian)
10. Lehmann E.L. A theory of some multiple decision problems, I. *Ann. Math. Stat.*, 1957, vol. 28, no. 1, pp. 1–25.
11. Wald A. *Statistical Decision Functions*. New York, John Wiley & Sons, 1950. 179 p.
12. Koldanov P., Koldanov A., Kalyagin V., Pardalos P.M. Uniformly most powerful unbiased test for conditional independence in Gaussian graphical model. *Stat. Probab. Lett.*, 2017, vol. 122, pp. 90–95. doi: 10.1016/j.spl.2016.11.003.
13. Kalyagin V.A., Koldanov A.P., Koldanov P.A., Pardalos P.M., Zamaraevand V.A. Measures of uncertainty in market network analysis, *Phys. A*, 2014, vol. 413, no. 1, pp. 59–70. doi: 10.1016/j.physa.2014.06.054.

Recieved
October 10, 2017

Koldanov Petr Alexandrovich, Candidate of Technical Sciences
National Research University Higher School of Economics
ul. B. Pecherskaya 2, Nizhny Novgorod, 603025 Russia
E-mail: *pkoldanov@hse.ru*

УДК 517.2

Функция риска и оптимальность статистических процедур определения сетевых структур

П.А. Колданов

Национальный исследовательский университет Высшая школа экономики, г. Нижний Новгород, 603025, Россия

Аннотация

Исследуется проблема определения сетевой структуры на основе конечной выборки. Приводятся понятия сети из случайных величин и сетевой модели. Рассматривается два типа сети: сетевые структуры с произвольным набором элементов и сетевые структуры с фиксированным количеством элементов сетевой модели. Определение сетевой структуры рассматривается как проблема множественного тестирования. Функция риска таких процедур может быть представлена как линейная комбинация числа неверно включённых в сеть и ошибочно не включённых в сеть элементов. Приводятся достаточные условия оптимальности статистических процедур для определения сетевых структур с произвольным количеством элементов. Рассматривается концепция неопределённости статистических процедур определения сетевой структуры.

Ключевые слова: сеть случайных величин, сетевая структура, сетевая модель, процедура определения сетевой структуры, аддитивная функция потерь, функция риска, несмещённость, оптимальность, статистическая неопределённость

Поступила в редакцию
10.10.17

Колданов Петр Александрович, кандидат технических наук, доцент кафедры прикладной математики и информатики

Национальный исследовательский университет Высшая школа экономики
ул. Б. Печерская, д. 2, г. Нижний Новгород, 603025, Россия
E-mail: *pkoldanov@hse.ru*,

⟨ ***For citation*:** Koldanov P.A. Risk function and optimality of statistical procedures for identification of network structures. *Uchenye Zapiski Kazanskogo Universiteta. Seriya Fiziko-Matematicheskie Nauki*, 2018, vol. 160, no. 2, pp. 317–326. ⟩

⟨ ***Для цитирования*:** *Koldanov P.A.* Risk function and optimality of statistical procedures for identification of network structures // Учен. зап. Казан. ун-та. Сер. Физ.-матем. науки. – 2018. – Т. 160, кн. 2. – С. 317–326. ⟩