Robust Identification of Subgraphs in a Complete Weighted Graph Associated with a Set of Random Variables

Kalyagin V.A., Koldanov A.P., and Koldanov P.A.

National Research University Higher School of Economics,
Laboratory of Algorithms and Technologies for Network Analysis (LATNA),
Nizhny Novgorod, Rodionova 136, 603155 Russia
vkalyagin@hse.ru

Abstract. A class of distribution free multiple decision statistical procedures is proposed for threshold graph identification in a complete weighted graph associated with a set of random variables (random variables network). The decision procedures are based on simultaneous application of sign statistics. It is proved that single step, step down Holm and step up Hochberg statistical procedures for threshold graph identification are distribution free in sign similarity network in the class of elliptically contoured distributions.

1 Introduction

Network model of complete weighted graph associated with a set of random variables is useful in biological and financial applications. Biological applications are mostly related with probabilistic graphical models [5], weighted correlation networks [3] and others. Financial applications are related with market network analysis [7], [2]. In this paper we consider a model which we call *random variables networks*. Random variables network is a pair (X, γ), where $X = (X_1, X_2, \ldots, X_N)$ is a random vector and γ is a measure of association of random variables. For Gaussian graphical model vector X has a multivariate Gaussian distribution and γ is the partial correlation. For market network model X_i is an attribute of stock i (return, volume, price and at.) and γ is the Pearson correlation (in most cases). Main goal of network analysis is to identify a network structures containing a key information about network. Popular network structures studied in the literature are concentration graph in Gaussian graphical models, and, minimum spanning tree, planar maximally filtered graph, threshold (market) graph, cliques and independent sets in market network analysis.

Random variable network is as complete weighted graph where the nodes are associated with random variables and weight of edge are given by a measure of association between them. Threshold graph is a subgraph of random variable network. An edge is included in *threshold graph* iff its weight is larger than a given threshold. According to the choice of measure of association one get different correlation networks and threshold

In: A. Kononov et al. (eds.): DOOR 2016, Vladivostok, Russia, published at http://ceur-ws.org

graphs. Threshold graph identification problem is to identify the threshold graph from observations. In this paper we study threshold graph identification problem in *sign similarity network* and compare it with identification problem in Pearson correlation network.

On our study of threshold graph identification problem we use a multiple decision statistical approach. The decision procedures considered in this paper are based on simultaneous application of sign tests. Three popular multiple statistical procedures are investigated: single step multiple decision procedure, step down Holm multiple testing procedure and step up Hochberg multiple testing procedure. The quality of the procedures is measured by risk function, and in particular FWER (Family Wise Error Rate). Our main result is: considered multiple decision procedures for threshold graph identification are robust (distribution free) in sign similarity network in the class of elliptically contoured distributions. Moreover it is shown that these procedures can be adapted for robust threshold graph identification in Pearson correlation network. This result gives a theoretical foundations for practical threshold graph identification algorithms in the case where the distribution of the vector X is unknown.

2 Basic definitions and notations

Let $X = (X_1, X_2, \ldots, X_N)$ be a random vector. Consider a complete weighted graph associated with X. Nodes of the graph are random variables X_i, $i = 1, \ldots, N$ and weight of edge (i, j) is given by some measure of association $\gamma(X_i, X_j)$ between them. One popular measure of association is Pearson correlation $\gamma^P_{i,j}$. Pearson correlation generates a Pearson correlation network. In this paper we study a sign similarity network, where the measure of association is given by the probability of sign coincidence $\gamma^S_{i,j} = P((X_i - E(X_i))(X_j - E(X_j)) > 0)$. This measure of association has a simple interpretation and was shown to be appropriate in market network analysis [1].

In this paper we assume that distribution of the random vector X belong to the class of elliptically contoured distributions (ECD) with density functions:

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

where Λ is positive definite matrix, $g(y) \geq 0$. Multivariate Gaussian and Student distributions are a particular cases of ECD.

Threshold graph is constructed as follows: the edge between two vertices i and j is included in the threshold graph, iff $\gamma_{i,j} > \gamma_0$ (where γ_0 is a threshold). For a given threshold γ_0 the threshold graph is defined by its adjacency matrix $S = (s_{i,j})$, where $s_{i,j} = 0$ if $\gamma^{i,j} \leq \gamma_0$ and $s_{i,j} = 1$ if $\gamma^{i,j} > \gamma_0$, $s_{i,i} = 0$, $i, j = 1, 2, \ldots, N$.

Let $x(t)$ be a sample of the size n from distribution of the random vector X:

$$x(t) = (x_1(t), x_2(t), \ldots, x_N(t)),\ t = 1, 2, \ldots, n$$

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 adjacency matrices of all simple undirected graphs with N vertices. Total number of matrices in $\mathcal{G}$ equals to $L = 2^M$ with $M = N(N-1)/2$.

3 Multiple decision framework

Threshold graph identification problem is to identify the threshold graph from observations. The problem can be formulated as a multiple decision problem of selecting one from a set of L hypotheses:

$$H_G : \gamma_{i,j} \leq \gamma_0, \text{ if } g_{i,j} = 0, \quad \gamma_{i,j} > \gamma_0, \text{ if } g_{i,j} = 1; \quad i \neq j \tag{2}$$

Multiple decision statistical procedure δ for threshold graph identification 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 hypothesis H_G, $G \in \mathcal{G}$.

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 $w(S,S) = 0, S \in \mathcal{G}$. According to general decision theory [8] the quality of statistical procedure δ is measured by the risk function. Let $f_X(x)$ be the density function for the random vector X. Risk function is then defined by

$$R(f_X; \delta) = \sum_{Q \in \mathcal{G}} w(S,Q) P_X(\delta(x) = d_Q / H_S),$$

where $w(f_X; \delta(x)) = w(S,Q)$ if $f_X \in H_S, \delta(x) = d_Q$.

In multiple hypotheses testing [6], there are different way to measure errors: per-comparison error rate (PCER), per-family error rate (PFER), family wise error rate (FWER), generalized family wise error rate (GFWER), false discovery rate (FDR). These errors can be considered as risk for appropriate choice of losses. For example if the loss $w(S,Q)$ takes two values zero and one: $w(S,Q) = 1$ if there is at least one incorrect inclusion of edge in the threshold graph. Risk function in this case is equal to the probability of at least one type I error (FWER, Family Wise Error Rate): In multiple decision theory [4], the losses are supposed to be additive. It means that the loss from misclassification of H_S is equal to the sum of losses from misclassification of individual hypotheses.

Consider the set of individual hypotheses:

$$h_{i,j} : \gamma_{i,j} \leq \gamma_0 \quad \text{vs} \quad k_{i,j} : \gamma_{i,j} > \gamma_0 \ (i,j = 1, \ldots, N; i \neq j).$$

we shall assume that tests for the individual hypotheses are available. For Pearson correlation network we use a well known correlation test. For sign similarity network we construct a uniformly most powerful tests for individual hypotheses testing. Using these tests one can construct a different multiple testing statistical procedures largely used in the literature: single step procedure δ^S (all tests are applied simultaneously), Holm step down procedure δ^H (at each step either one individual hypothesis $h_{i,j}$ is rejected or all remaining hypotheses are accepted) or Hochberg step up procedure δ^{Sg} (at each step either one individual hypothesis $h_{i,j}$ is accepted or all remaining hypotheses are rejected).

4 Robustness of statistical procedures for threshold graph identification in sign similarity network

We prove that single step, step down Holm, and step up Hochberg multiple testing procedures for threshold graph identification in sign similarity network are distribution free for any loss function.

Theorem. Let random vector $(X_1, \dots, X_N)$ has elliptically contoured distribution with density $f(x; 0, \Lambda)$. Then for single step, Holm, Hochberg identification statistical procedures the probabilities $P(\delta(x) = d_Q/H_S)$, $Q, S \in \mathcal{G}$ are defined by the matrix Λ and does not depend on the function g.

Corollary. Let random vector $(X_1, \dots, X_N)$ has elliptically contoured distribution with density $f(x; 0, \Lambda)$. Then the risk functions $R(f_X; \delta^S)$, $R(f_X; \delta^H)$, $R(f_X; \delta^{Hg})$ are defined by the matrix Λ and does not depend on the function g for any loss function $w(S, Q)$.

In particular for the loss function $w(S, Q)$ such that $w(S, Q) = 1$ if there is at least one incorrect inclusion of edge in the threshold graph, and $w(S, Q) = 0$ otherwise the risk is equal to FWER (Family Wise Error Rate). Therefore the FWER of single step, Holm, and Hochberg statistical procedures are defined by the matrix Λ and does not depend on the function g. The same is true for other type of errors, PCER, PFER, GFWER, FDR and for risk function with additive losses. Using this result it is possible to construct single step, Holm and Hochberg distribution free statistical procedures for threshold graph identification in Pearson correlation network too.

Acknowledgement: this work is partly supported by RFFI 14-01-00807 grant.

References

1. Bautin G.A., Kalyagin V.A., Koldanov A.P., Koldanov P.A., Pardalos P.M. Simple measure of similarity for the market graph construction Computational Management Science, 10, 105-124 (2013).
2. Boginski V., Butenko S., Pardalos P.M.: Statistical analysis of financial networks, Computational Statistics and Data Analysis. 48 (2), 431–443 (2005).
3. Horvath S. Weighted Network Analysis: Application in Genomics and Systems Biology, Springer book, 2011.
4. Hochberg, Y. and Tamhane, A. C. Multiple Comparison Procedures, John Wiley and Sons, Inc., Hoboken, NJ, USA, 1987.
5. Koller D. Friedman N. Probabilistic Graphical Models, MIT Press, 2009.
6. Lehmann E.L., Romano J.P.: Testing statistical hypotheses. Springer, New York, (2005).
7. 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.
8. Wald A.: Statistical Decision Function. John Wiley and Sons, New York (1950).