УДК 519.21

ФУНКЦИЯ РИСКА СТАТИСТИЧЕСКИХ ПРОЦЕДУР ИДЕНТИФИКАЦИИ СЕТЕВЫХ СТРУКТУР[1]

Колданов П.А.
НИУ Высшая Школа экономики, г. Нижний Новгород

Поступила в редакцию 10.07.2017, после переработки 12.08.2017.

Рассматривается класс статистических задач идентификации сетевых структур по конечному объему наблюдений. Вводятся понятия сети случайных величин, сетевой модели, представимой в виде полного взвешенного графа. Рассматриваются два типа сетевых структур: сетевые структуры с заданным и произвольным числом элементов сетевой модели. Задачи идентификации сетевых структур рассматриваются как статистические задачи выбора одной из многих гипотез о составе сетевой структуры. Доказано, что функцию риска процедур идентификации сетевых структур можно представить как линейную комбинацию средних чисел ошибок неверного включения и невключения элемента сетевой модели в идентифицируемую структуру.

Ключевые слова: сеть случайных величин, сетевая модель, сетевая структура, процедура идентификации, ошибки первого и второго рода, аддитивная функция потерь, функция риска.

Вестник ТвГУ. Серия: Прикладная математика. 2017. № 3. С. 45–59.
https://doi.org/10.26456/vtpmk178

Введение

Один из основных методов анализа сложных объектов заключается в построении и исследовании соответствующей сетевой модели, которая допускает визуализацию простым полным взвешенным графом [1]. Под сетевой моделью понимается модель сложного объекта, состоящего из N связанных между собой элементов. В качестве элементов могут рассматриваться нейроны, люди, акции, гены и т.д. В качестве характеристики связей могут выступать произвольные меры взаимодействия между ними.

Сетевую модель G удобно представить простым полным взвешенным графом $G = (V, E, \gamma)$ где $V = \{1, 2, \ldots, N\}$ – множество элементов сложного объекта (вершин графа), а $E = \{(i, j) : i, j = 1, 2, \ldots, N\}$ – множество ребер с весами $\gamma_{i,j}$. По способу описания вершин можно выделить два типа сетевых моделей: детерминированные и вероятностные. В настоящей работе рассматриваются только вероятностные сетевые модели (probabilistic graphical model or graphical model), которые предполагают, что вершины графа характеризуются случайными величинами.

[1]Работа выполнена при финансовой поддержке Российского гуманитарного научного фонда (проект 15-32-01052).

Графическое представление связей между случайными величинами дает возможность факторизовать их совместное распределение, что приводит к упрощению вычислений. Эта идея использовалась еще в работах [2, 3]. В настоящее время такое представление находит применение в нейронных сетях и генетике [4, 5]. В этих работах совместное распределение случайных величин предполагается известным и основной целью является разработка вычислительных алгоритмов анализа больших массивов данных и их применение к решению задач выявления структуры зависимости элементов таких сетей. Хорошо изученной сетевой моделью, в которой вершины характеризуются случайными величинами, является Гауссовская графическая модель (Gaussian graphical model) [6]. В последние годы начато исследование статистических задач идентификации гауссовских графических моделей по наблюдениям [7–13]. В этих работах предлагаются статистические процедуры, контролирующие характеристики только ошибок первого рода (ложного включения ребра в сетевую структуру).

Другой сетевой моделью, в которой вершины описываются случайными величинами, является сетевая модель фондового рынка. Каждая вершина этой сетевой модели соответствует акции, а веса ребер определяются выбранной мерой зависимости между случайными величинами, характеризующими доходности акций. При этом популярными сетевыми структурами являются отсеченный граф [14] и максимальное остовное дерево [15]. Отсеченным графом принято называть граф без весов, полученный из сетевой модели удалением ребер с весами, меньшими заданного порога. Максимальным остовным деревом называют остовное дерево полного взвешенного графа с максимальной суммой весов входящих в него ребер. В настоящее время имеется значительное количество публикаций по вычислению таких структур по наблюдениям и интерпретации полученных результатов. В работе [16] предложен статистический подход к построению процедур идентификации отсеченного графа.

В настоящей работе рассматривается общая постановка задачи идентификации сетевых структур по наблюдениям. Естественной характеристикой таких процедур являются средние числа ошибок первого и второго (неверное невключение элемента в сетевую структуру) родов. Выделяется два типа задач. К первому типу относятся задачи идентификации сетевых структур с произвольным числом элементов сетевой модели. К такому типу относятся задачи идентификации отсеченного графа. Ко второму типу относятся задачи идентификации сетевых структур с заданным числом элементов сетевой модели. К такому типу относятся задачи идентификации максимального остовного дерева. Доказано, что функцию риска процедур идентификации сетевых структур первого и второго типов можно представить как сумму средних чисел ошибок первого и второго родов.

1. Основные понятия и формулировка статистических задач идентификации сетевых структур

Определение 1. *Сетью случайных величин будем называть пару* (X, γ), *где* $X = (X_1, \ldots, X_N)$ – *вектор случайных величин, а* $\gamma = \{\gamma_{i,j} : i, j = 1, \ldots, N; i \neq j\}$ *мера зависимости между случайными величинами* X_i, X_j.

Сеть случайных величин порождает сетевую модель, которая представляет собой простой полный неориентированный взвешенный граф без петель

$G = (V, E, \gamma)$, где $V = \{1, 2, \ldots, N\}$ – множество вершин, которые описываются случайными величинами $X_1, X_2, \ldots, X_N$, E – множество ребер с весами, заданными мерой γ. Изучение сетевых моделей $G = (V, E, \gamma)$ естественно свести к изучению ключевых характеристик соответствующих графов.

В настоящей работе исследуются характеристики графов, удовлетворяющие следующему определению.

Определение 2. *Сетевой структурой модели $G = (V, E, \gamma)$ называется подграф без весов $G' = (V', E') : V' \subseteq V, E' \subseteq E$.*

Можно выделить два типа сетевых структур. К первому типу отнесем сетевые структуры, которые могут содержать произвольное число элементов сетевой модели. К такому типу структур относится отсеченный граф, частным случаем которого является гауссовская графическая модель.

Определение 3. *Отсеченным графом (TG) сетевой модели $G = (V, E, \gamma)$ называется подграф $G'(\gamma_0) = (V', E') : V' = V; E' \subseteq E, E' = \{(i, j) : \gamma_{i,j} > \gamma_0\}$, где γ_0 – некоторый порог.*

Ко второму типу сетевых структур относятся сетевые структуры, которые содержат заданное число элементов сетевой модели. К такому типу структур относится максимальное остовное дерево.

Определение 4. *Максимальным остовным деревом (MST) сетевой модели $G = (V, E, \gamma)$ называется дерево (граф без циклов) $G' = (V', E') : V' = V; E' \subset E; |E'| = |V| - 1$; такое, что $\sum_{(i,j)\in E'} \gamma_{i,j}$ максимальна.*

В настоящей работе предлагается следующая общая статистическая формулировка задач идентификации сетевых структур:

Определение 5. *Пусть (X, γ) – сеть случайных величин. Плотность распределения вектора $X, f(x) \in \{f(x, \theta) : \theta \in \Omega\}$. Пусть $G = (V, E, \gamma)$ – порожденная сетью (X, γ) сетевая модель. Элементами сетевой модели $G = (V, E, \gamma)$ будем называть элементы множества E и обозначать через $\beta : \beta = 1, \ldots, K$, $K = \frac{N(N-1)}{2}$. Пусть $G' = (V', E') : V' \subseteq V, E' \subseteq E$ – сетевая структура, которую требуется идентифицировать по наблюдениям $x_i(t), i = 1, \ldots, N, t = 1, \ldots, n$. Задачу идентификации сетевой структуры по наблюдениям будем формулировать следующим образом: пусть $h_\beta : \theta \in \omega_\beta$ – гипотеза о том, что элемент β сетевой модели не принадлежит идентифицируемой структуре, $k_\beta : \theta \in \omega_\beta^{-1}$ – альтернатива h_β, $H_i : \theta \in \Omega_i; i = 1, \ldots, L$ – гипотеза о том, что элементы $\{i_1, i_2, \ldots, i_M\}, \{i_1, i_2, \ldots, i_M\} \subseteq \{1, 2, \ldots, K\}$ принадлежат идентифицируемой сетевой структуре. M – число элементов идентифицируемой сетевой структуры. Требуется построить статистическую процедуру выбора одной из гипотез*

$$
\begin{array}{l}
H_i : \theta \in \Omega_i, \\
\textit{где} \\
\Omega_i = (\bigcap_{i_l \in \{i_1, \ldots, i_M\}} \omega_{i_l}^{-1}) \cap (\bigcap_{i_s \in \{1, \ldots, K\} - \{i_1, \ldots, i_M\}} \omega_{i_s}) \\
\textit{или} \\
H_i = (\bigcap_{i_l \in \{i_1, \ldots, i_M\}} k_{i_l}) \cap (\bigcap_{i_s \in \{1, \ldots, K\} - \{i_1, \ldots, i_M\}} h_{i_s}).
\end{array} \tag{1}
$$

В зависимости от M можно выделить два типа задач:

– задачи с произвольным числом элементов сетевой модели – $M \in \{0, 1, \ldots, C_N^2\}$,

– задачи с заданным числом M элементов сетевой модели.

2. Общий вид процедур идентификации сетевых структур

Пусть $\varphi_\beta(x)$ – тесты проверки индивидуальных гипотез h_β против альтернатив k_β с областями принятия A_β и областями отвержения A_β^{-1} соответственно. Пусть $\delta(x)$ – статистическая процедура различения гипотез (5) с решениями d_i об истинности гипотезы $H_i, i = 1, \ldots, L$, D_i – область принятия H_i, т.е.

$$\begin{array}{l} \delta(x) = d_i, \text{ если } x \in D_i, \\ D_i \cap D_j = \emptyset, i \neq j, i, j = 1, \ldots, L; \bigcup_{i=1}^{L} D_i = \mathcal{X}, \\ \mathcal{X} - \text{выборочное пространство.} \end{array} \tag{2}$$

Любая процедура идентификации сетевой структуры с произвольным числом элементов сетевой модели может быть записана в виде:

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

где

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

Для задач идентификации сетевых структур с заданным числом M элементов сетевой модели необходимо выполнение условий совместности.

Определение 6. *Семейство тестов $\varphi_\beta(x)$ будем называть совместным с пространством решений процедуры $\delta(x)$(2), если*

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

Если семейство тестов $\varphi_\beta(x)$ совместно с пространством решений процедуры $\delta(x)$, то можно установить взаимно-однозначное соответствие между процедурой $\delta(x)$(2) и этим семейством [17]. Такое соответствие имеет вид:

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

В случае совместного семейства тестов $\varphi_\beta(x)$ соотношения (6) определяют общий вид статистических процедур идентификации сетевых структур.

3. Функция риска статистических процедур идентификации сетевых структур

Пусть $w(H_i; d_j) = w_{ij}$ – потери от принятия решения d_j при истинности гипотезы H_i. Предположим, что потери от правильного решения равны нулю, т.е. $w_{ii} = 0, \forall i = 1, \ldots, L$. Будем характеризовать качество статистической процедуры $\delta(x)$ функцией риска

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

где $P_\theta(\delta(x) = d_j)$ – вероятность принятия решения d_j.

Будем считать, что функция потерь w_{ij} задается соотношениями

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

где

$$\epsilon_{ij\beta} = \begin{cases} 1, & \text{если } \kappa_{i\beta} = 1, \kappa_{j\beta} = -1, \\ 0, & \text{иначе.} \end{cases}$$

$\kappa_{i\beta}$ определены в (4).

Для задач идентификации сетевых структур с произвольным числом элементов справедлива следующая теорема

Теорема 1. *Пусть функция потерь задается (7). Тогда справедливо следующее соотношение:*

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

где $r(h_\beta, \varphi_\beta)$ – функция риска теста φ_β.

Если $a_\beta = a, b_\beta = b, \forall \beta = 1, \ldots, K$, то:

$$R(H_i, \theta, \delta) = a E_\theta\{Y_I(H_i, \delta)\} + b E_\theta\{Y_{II}(H_i, \delta)\}, \tag{9}$$

где $Y_I(H_i, \delta)$ – число элементов, неправильно включенных (число ошибок первого рода) процедурой δ при истинности H_i; $Y_{II}(H_i, \delta)$ – число элементов, неправильно не включенных (число ошибок второго рода) процедурой δ при истинности H_i.

Доказательство. Для $\forall \theta \in \Omega_i$ функция риска процедуры δ может быть представ-

лена в виде:

$$R(H_i,\theta,\delta)=\sum_{j\neq i}w_{ij}P_\theta(\delta=d_j)=\sum_{j\neq i}\sum_\beta(a_\beta\epsilon_{ij\beta}+b_\beta\epsilon_{ji\beta})P_\theta(x\in D_j)=$$

$$=\sum_\beta\left[a_\beta\sum_{j\neq i}\epsilon_{ij\beta}P_\theta(x\in D_j)+b_\beta\sum_{j\neq i}\epsilon_{ji\beta}P_\theta(x\in D_j)\right]=$$

$$=\sum_\beta\left[a_\beta\sum_{\substack{j\neq i;\\ \kappa_{i\beta}=1;\\ \kappa_{j\beta}=-1}}P_\theta(x\in D_j)+b_\beta\sum_{\substack{j\neq i;\\ \kappa_{i\beta}=-1;\\ \kappa_{j\beta}=1}}P_\theta(x\in D_j)\right]= \tag{10}$$

$$=\sum_\beta\left[a_\beta P_\theta(x\in\bigcup_{\substack{j\neq i;\\ \kappa_{i\beta}=1;\\ \kappa_{j\beta}=-1;}}D_j)+b_\beta P_\theta(x\in\bigcup_{\substack{j\neq i;\\ \kappa_{i\beta}=-1;\\ \kappa_{j\beta}=1;}}D_j)\right]=$$

$$=\sum_\beta\left[a_\beta P_\theta(x\in\bigcup_{j:\kappa_{j\beta}=-1}D_j)+b_\beta P_\theta(x\in\bigcup_{j:\kappa_{j\beta}=1}D_j)\right]=$$

$$=\sum_\beta\left[a_\beta P_\theta(x\in A_\beta^{-1})+b_\beta P_\theta(x\in A_\beta)\right].$$

Так как

$$r(h_\beta,\varphi_\beta)=\begin{cases}a_\beta P_\theta(x\in A_\beta^{-1}), & \theta\in\omega_\beta,\\ b_\beta P_\theta(x\in A_\beta), & \theta\in\omega_\beta^{-1},\end{cases}$$

то (8) доказано.

В предположении $a_\beta=a;b_\beta=b,\forall\beta=1,\ldots,K$ риск процедуры δ равен

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

где $Y_I(H_i,\delta)$ – число ошибок первого рода, $Y_{II}(H_i,\delta)$ – число ошибок второго рода, допущенных процедурой δ. □

Для задач идентификации сетевых структур с заданным числом элементов имеет место следующая теорема:

Теорема 2. *Пусть*

- *семейство тестов φ_β индивидуальных гипотез h_β совместно с пространством решений процедуры различения гипотез H_i;*
- *функция потерь аддитивна и задается (7) Тогда функция риска статистической процедуры δ для задач идентификации сетевых структур с заданным числом элементов имеет вид:*

$$R(H_i,\theta,\delta)=\sum_{\beta=1}^{K}r(h_\beta,\varphi_\beta), \tag{11}$$

где $r(h_\beta,\varphi_\beta)$ – функция риска теста φ_β.

– *Если* $a_\beta = a, b_\beta = b, \beta = 1, \ldots, K$, *то функция риска статистической процедуры* δ *для задач идентификации сетевых структур с заданным числом элементов имеет вид:*

$$R(H_i, \theta, \delta) = (a+b)E_\theta(Y_I(H_i, \delta)) = (a+b)E_\theta(Y_{II}(H_i, \delta)), \tag{12}$$

где $Y_I(H_i, \delta)$ *– число ошибок* $1-$*го рода,* $Y_{II}(H_i, \delta)$ *– число ошибок* $2-$*го рода, допущенных процедурой* δ *при истинности* H_i.

Доказательство. 1. Покажем, что $\forall i, j = 1, \ldots, L$ параметрические области $\Omega_i \neq \Omega_j$ гипотез $H_i : \theta \in \Omega_i, H_j : \theta \in \Omega_j$ отличаются не менее, чем на две индивидуальные гипотезы h_β.

$$\forall i, j = 1, \ldots, L : \Omega_i \neq \Omega_j \Rightarrow \exists \beta_0 : \kappa_{i\beta_0} \neq \kappa_{j\beta_0}.$$

Пусть $\kappa_{i\beta_0} = 1, \kappa_{j\beta_0} = -1$. Так как число элементов в сетевой сруктуре фиксировано, то

$$\exists \beta' : \kappa_{i\beta'} = -1; \kappa_{j\beta'} = 1.$$

Заметим, что в общем случае $\forall i, j = 1, \ldots, L : \Omega_i, \Omega_j$ отличаются друг от друга на четное число областей индивидуальных гипотез и альтернатив.

2. Пусть $\varphi_\beta = \begin{cases} 1, & x \in A_\beta^{-1}, \\ 0, & x \in A_\beta \end{cases}$ тесты проверки индивидуальных гипотез h_β против альтернатив k_β. Рассмотрим правило

$$\delta = \begin{cases} d_1, & x \in D_1 = \bigcap_\beta A_\beta^{\kappa_{1\beta}}, \\ d_2, & x \in D_2 = \bigcap_\beta A_\beta^{\kappa_{2\beta}}, \\ \ldots, & \ldots, \\ d_L, & x \in D_L = \bigcap_\beta A_\beta^{\kappa_{L\beta}}. \end{cases} \tag{13}$$

Условие совместности семейства тестов φ_β и правила δ означают:

$$\sum_{(\kappa_{i\beta_1}, \kappa_{i\beta_2}, \ldots, \kappa_{i\beta_K}) : \kappa_{i\beta_{i_1}} = \ldots = \kappa_{i\beta_{i_M}} = -1, \kappa_{i\beta_{i_{M+1}}} = \ldots = \kappa_{i\beta_{i_N}} = 1} P_\theta(\bigcap_\beta A_\beta^{\kappa_{i\beta}}) = 1. \tag{14}$$

Следовательно,

$$P_\theta(\bigcap_\beta A_\beta^{\kappa_{i\beta}}) = 0,$$

если количество $\kappa_{i\beta} = -1$ не равно M. Поэтому альтернативная запись условия совместности для задач этого типа имеет вид:

$$P_\theta(\bigcap_\beta A_\beta^{\kappa_{i\beta}}) = 0 \text{ если } \sum_{\beta : \kappa_{i\beta} = -1} \kappa_{i\beta} \neq -M. \tag{15}$$

3. Пусть при истинности гипотезы $H_i : \theta \in \Omega_i$ принято решение d_j о том, что истинна гипотеза H_j. В соответствии с условием аддитивности функции потерь (7)

$$w_{ij} = \sum_{\beta=1}^{K} (\epsilon_{ij\beta} a_\beta + \epsilon_{ji\beta} b_\beta) = \left(\sum_{\substack{\beta : \kappa_{i\beta} = 1 \\ \kappa_{j\beta} = -1}} a_\beta + \sum_{\substack{\beta : \kappa_{i\beta} = -1 \\ \kappa_{j\beta} = 1}} b_\beta \right), \tag{16}$$

где a_β, b_β – потери от ложного отвержения, принятия гипотез h_β, соответственно. Из пункта 1 доказательства следует, что число $\beta : \kappa_{i\beta} = 1, \kappa_{j\beta} = -1$ равно числу $\beta : \kappa_{i\beta} = -1, \kappa_{j\beta} = 1$, т.е. число слагаемых в первой сумме (16) равно числу слагаемых во второй сумме (16).

4. Пусть $\theta \in \Omega_i$. Тогда $\exists \beta' : \kappa_{i\beta'} = 1$, т.е. элемент β' не принадлежит сетевой структуре. Функция риска правила δ при $\theta \in \Omega_i$ имеет вид:

$$R(H_i, \theta, \delta) = \sum_{j=1}^{L} w_{ij} P_\theta(\delta = d_j | H_i) =$$

$$= \sum_{j=1}^{L} \left(\sum_{\substack{\beta : \kappa_{i\beta} = 1 \\ \kappa_{j\beta} = -1}} a_\beta + \sum_{\substack{\beta : \kappa_{i\beta} = -1 \\ \kappa_{j\beta} = 1}} b_\beta \right) P_\theta(\textstyle\bigcap_\beta A_\beta^{\kappa_{j\beta}} | H_i) =$$

$$= \sum_{\substack{\beta : \kappa_{i\beta} = 1 \\ \kappa_{j\beta} = -1}} \sum_{j=1}^{L} a_\beta P_\theta(\textstyle\bigcap_\beta A_\beta^{\kappa_{j\beta}} | H_i) +$$
$$+ \sum_{\substack{\beta : \kappa_{i\beta} = -1 \\ \kappa_{j\beta} = 1}} \sum_{j=1}^{L} b_\beta P_\theta(\textstyle\bigcap_\beta A_\beta^{\kappa_{j\beta}} | H_i) =$$

$$= a_{\beta'} \sum_{j=1}^{L} P_\theta(A_{\beta'}^{-1} \textstyle\bigcap_{\beta \neq \beta'} A_\beta^{\kappa_{j\beta}} | H_i) +$$
$$+ \sum_{\substack{\beta \neq \beta' \\ \kappa_{i\beta} = 1 \\ \kappa_{j\beta} = -1}} \sum_{j=1}^{L} a_\beta P_\theta(\textstyle\bigcap_\beta A_\beta^{\kappa_{j\beta}} | H_i) +$$
$$+ \sum_{\substack{\beta : \kappa_{i\beta} = -1 \\ \kappa_{j\beta} = 1}} \sum_{j=1}^{L} b_\beta P_\theta(\textstyle\bigcap_\beta A_\beta^{\kappa_{j\beta}} | H_i).$$

Заметим, что в соответствии с (15)

$$P_\theta(A_{\beta'}^{-1} \cap \bigcap_{\beta \neq \beta'} A_\beta^{\eta_{i\beta}}) = 0,$$

если

$$\sum_{\beta : \eta_{i\beta} = -1} \eta_{i\beta} \neq -M + 1.$$

Поэтому

$$\sum_{j=1}^{L} P_\theta(\bigcap_\beta A_\beta^{\kappa_{j\beta}} | H_i) = P(A_{\beta'}^{-1}).$$

Следовательно при $\kappa_{i\beta'} = 1$ функция риска имеет вид

$$R(H_i, \theta, \delta) = a_{\beta'} P_\theta(A_{\beta'}^{-1}) +$$
$$+ \sum_{\substack{\beta \neq \beta' \\ \kappa_{i\beta} = 1 \\ \kappa_{j\beta} = -1}} \sum_{j=1}^{L} a_\beta P_\theta(\textstyle\bigcap_\beta A_\beta^{\kappa_{j\beta}} | H_i) +$$
$$+ \sum_{\substack{\beta : \kappa_{i\beta} = -1 \\ \kappa_{j\beta} = 1}} \sum_{j=1}^{L} b_\beta P_\theta(\textstyle\bigcap_\beta A_\beta^{\kappa_{j\beta}} | H_i). \qquad (17)$$

Применяя аналогичные рассуждения ко второй и третьей сумме (17), получаем справедливость (11) для $\forall\theta \in \Omega$.

При $a_\beta = a, b_\beta = b$ имеем

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

Из пунктов 1, 3 следует, что $Y_I(H_i, \delta) = Y_{II}(H_i, \delta)$, что доказывает справедливость (12). □

Пример 1. Пусть $N = 3$, $M = 2$ и гипотезы, которые необходимо различить, характеризуются областями:

$$\begin{array}{l} \Omega_1 = \omega_1 \cap \omega_2^{-1} \cap \omega_3^{-1}, \\ \Omega_2 = \omega_1^{-1} \cap \omega_2 \cap \omega_3^{-1}, \\ \Omega_3 = \omega_1^{-1} \cap \omega_2^{-1} \cap \omega_3, \end{array} \tag{18}$$

где ω_β соответствует гипотезе h_β о том, что элемент β не принадлежит идентифицируемой структуре. Имеем:

$$\begin{array}{lll} \kappa_{1,1} = 1; & \kappa_{1,2} = -1; & \kappa_{1,3} = -1; \\ \kappa_{2,1} = -1; & \kappa_{2,2} = 1; & \kappa_{2,3} = -1; \\ \kappa_{3,1} = -1; & \kappa_{3,2} = -1; & \kappa_{3,3} = 1. \end{array} \tag{19}$$

Пусть φ_β – тесты проверки индивидуальных гипотез h_β, с областями принятия A_β и областями отвержения A_β^{-1}. Построим правило с тремя решениями

$$\begin{array}{l} D_1 = A_1 \cap A_2^{-1} \cap A_3^{-1}, \\ D_2 = A_1^{-1} \cap A_2 \cap A_3^{-1}, \\ D_3 = A_1^{-1} \cap A_2^{-1} \cap A_3. \end{array} \tag{20}$$

Условие совместности тестов $\varphi_\beta, \beta = 1, 2, 3$ и задачи (18) означает, что

$$\begin{array}{l} P_\theta(A_1^{\kappa_{i,1}} \cap A_2^{\kappa_{i,2}} \cap A_3^{\kappa_{i,3}}) = 0, \\ \text{если } (\kappa_{i,1}, \kappa_{i,2}, \kappa_{i,3}) \neq (1, -1, -1), \\ \text{или } (\kappa_{i,1}, \kappa_{i,2}, \kappa_{i,3}) \neq (-1, 1, -1), \\ \text{или } (\kappa_{i,1}, \kappa_{i,2}, \kappa_{i,3}) \neq (-1, -1, 1). \end{array} \tag{21}$$

Пусть $\delta(x)$ – правило с тремя решениями

$$\delta(x) = \begin{cases} d_1, & x \in D_1 = A_1 \cap A_2^{-1} \cap A_3^{-1}, \\ d_2, & x \in D_2 = A_1^{-1} \cap A_2 \cap A_3^{-1}, \\ d_3, & x \in D_3 = A_1^{-1} \cap A_2^{-1} \cap A_3. \end{cases} \tag{22}$$

Покажем, что функция риска правила (22) имеет вид:

$$R(\theta, \delta) = r(\theta, \varphi_1) + r(\theta, \varphi_2) + r(\theta, \varphi_3). \tag{23}$$

Рассмотрим

$$\begin{array}{c} P_\theta(A_1^{-1}) = P_\theta(A_1^{-1} \cap (A_2 \cup A_2^{-1}) \cap (A_3 \cup A_3^{-1})) = \\ = P_\theta(A_1^{-1} \cap A_2 \cap A_3) + P_\theta(A_1^{-1} \cap A_2 \cap A_3^{-1}) + \end{array}$$

$$+P_\theta(A_1^{-1} \cap A_2^{-1} \cap A_3) + P_\theta(A_1^{-1} \cap A_2^{-1} \cap A_3^{-1}).$$

Из условия (21) получаем, что

$$P_\theta(A_1^{-1}) = P_\theta(D_2 \cup D_3) = P_\theta(\bigcup_{i:\kappa_{i\beta}=-1} D_i).$$

Аналогично получаем, что

$$P_\theta(A_\beta^{-1}) = P_\theta(\bigcup_{i:\kappa_{i\beta}=-1} D_i).$$

Пусть $\theta \in \Omega_1$. Тогда $R(\theta, \delta) = w_{12}P_\theta(x \in D_2) + w_{13}P_\theta(x \in D_3)$. Так как

$$\begin{array}{lll} \epsilon_{121} = 1; & \epsilon_{211} = 0; & \epsilon_{122} = 0; \\ \epsilon_{212} = 1; & \epsilon_{123} = 0; & \epsilon_{213} = 0; \end{array}$$

то

$$w_{12} = a_1 + b_2; w_{13} = a_1 + b_3.$$

Следовательно,

$$R(\theta, \delta) = (a_1 + b_2)P_\theta(x \in D_2) + (a_1 + b_3)P_\theta(x \in D_3) =$$

$$= a_1P_\theta(x \in D_2 \cup D_3) + b_2P_\theta(x \in A_1^{-1} \cap A_2 \cap A_3^{-1}) + b_3P_\theta(x \in A_1^{-1} \cap A_2^{-1} \cap A_3) =$$

$$= a_1P_\theta(A_1^{-1}) + b_2P_\theta((A_1^{-1} \cap A_2 \cap A_3^{-1}) \cup (A_1 \cap A_2 \cap A_3^{-1}) \cup (A_2 \cap A_3)) +$$

$$+b_3P_\theta((A_1^{-1} \cap A_2^{-1} \cap A_3) \cup (A_1 \cap A_2^{-1} \cap A_3) \cup (A_2 \cap A_3)) =$$

$$= a_1P_\theta(A_1^{-1}) + b_2P_\theta(A_2) + b_3P_\theta(A_3) =$$

$$= r(\theta, \varphi_1) + r(\theta, \varphi_2) + r(\theta, \varphi_3) \text{ при } \theta \in \Omega_1.$$

При $a_i = a, b_i = b, i = 1, 2, 3$ имеем

$$R(\theta, \delta) = (a + b)P_\theta(x \in D_2 \cup D_3) = (a + b)P_\theta(x \in A_1^{-1}) = (a + b)E(X_1), \forall\theta \in \Omega_1.$$

Заключение

Теорема 1 дает строгое доказательство соотношения (9), использовавшегося в работах [17–21] при решении задач множественной проверки гипотез, в которых произвольное число индивидуальных гипотез может быть истинно. Теорема 2 доказывает, что подобное соотношение справедливо также для задач множественной проверки гипотез, в которых истинными могут быть только заданное число индивидуальных гипотез.

Список литературы

[1] Jordan M.I. Graphical models // Statistical Science. 2004. Vol. 19. Pp. 140–155. https://doi.org/10.1214/088342304000000026

[2] Gibbs W. Elementary Principles of Statistical Mechanics. NewHaven, Connecticut: Yale University press, 1902.

[3] Wright S. The method of path coefficients // Annals of Mathematical Statistics. 1934. Vol. 5. Pp. 161–215. https://doi.org/10.1214/aoms/1177732676

[4] Koller D., Friedman N. Probabilistic Graphical Models. Massachusetts: MIT Press, 2009.

[5] Lauritzen S.L., Sheehan N.A. Graphical models for genetic analyses // Statistical Science. 2003. Vol. 18, № 4. Pp. 489–514. https://doi.org/10.1214/ss/1081443232

[6] Lauritzen S.L. Graphical Models. Oxford university press, 1996.

[7] Dempster A.P. Covariance selection // Biometrics. 1972. Vol. 28. Pp. 157–175.

[8] Wermuth N. Analogies between multiplicative models in contingency tables and covariance selection // Biometrics. 1976. Vol. 32. Pp. 95–108.

[9] Edwards D. Introduction to Graphical Modeling. New York: Springer-Verlag, 2000.

[10] Anderson T.W. An Introduction to Multivariate Statistical Analysis. 3-d edition. New York: Wiley-Interscience, 2003.

[11] Drton M., Perlman M. Model selection for Gaussian concentration graph // Biometrika. 2004. Vol. 91, № 3. Pp. 591–602. https://doi.org/10.1093/biomet/91.3.591

[12] Drton M., Perlman M. Multiple testing and error control in Gaussian graphical model selection // Statistical Science. 2007. Vol. 22, № 3. Pp. 430–449. https://doi.org/10.1214/088342307000000113

[13] Liu W. Gaussian graphical model estimation with false discovery rate control // The Annals of Statistics. 2013. Vol. 41, № 6. Pp. 2948–2978. https://doi.org/10.1214/13-AOS1169

[14] Boginski V., Butenko S., Pardalos Panos M. On structural properties of the market graph // Innovations in Financial and Economic Networks. 2003. Pp. 29–45.

[15] Mantegna R.N. Hierarchical structure in financial markets // The European Physical Journal B-Condensed Matter and Complex Systems. 1999. Vol. 11, № 1. Pp. 193–197. https://doi.org/10.1007/s100510050929

[16] Koldanov A.P., Koldanov P.A., Kalyagin V.A. Statistical procedures for the market graph construction // Computational Statistics & Data Analysis. 2013. Vol. 68. Pp. 17–29.

[17] Lehmann E.L. A theory of some multiple decision problems, I // The Annals of Mathematical Statistics. 1957. Vol. 28, № 1. Pp. 1–25. https://doi.org/10.1214/aoms/1177707034

[18] Lehmann E.L. A theory of some multiple decision problems. II // The Annals of Mathematical Statistics. 1957. Vol. 28, № 3. Pp. 547–572. https://doi.org/10.1214/aoms/1177706873

[19] Cohen A., Sackrowitz H.B. Monotonicity properties of multiple endpoint testing procedures // Journal of Statistical Planning and Inference. 2004. Vol. 125. Pp. 17–30. https://doi.org/10.1016/j.jspi.2003.10.008

[20] Cohen A., Sackrowitz H.B. Decision theory results for one-sided multiple comparison procedures // The Annals of Statistics. 2005. Vol. 33, № 1. Pp. 126–144. https://doi.org/10.1214/009053604000000968

[21] Genovese C., Wasserman L. Operating characteristics and extensions of the false discovery rate procedure // Royal Statistical Society. 2002. Vol. 64, № 3. Pp. 499–517. https://doi.org/10.1111/1467-9868.00347

Образец цитирования

Колданов П.А. Функция риска статистических процедур идентификации сетевых структур // Вестник ТвГУ. Серия: Прикладная математика. 2017. № 3. С. 45–59. https://doi.org/10.26456/vtpmk178

Сведения об авторах

1. **Колданов Петр Александрович**
доцент кафедры прикладной математики и информатики НИУ ВШЭ в Нижнем Новгороде.

Россия, 603005, г. Нижний Новгород, ул. Большая Печерская, д. 25/12.
E-mail: pkoldanov@hse.ru.

RISK FUNCTION OF STATISTICAL PROCEDURES FOR NETWORK STRUCTURES IDENTIFICATION

Koldanov Petr Alexandrovich
Associated professor at Applied Mathematics and Informatics department,
NRU Higher School of Economics
Russia, 603005, Nizhniy Novgorod, 25/12 B.-Pecherskay.
E-mail: pkoldanov@hse.ru

Received 10.07.2017, revised 12.08.2017.

Problems of network structures identification by sample of fixed size are considered. The concepts of random variables network and network model as complete weighted graph are introduced. Two types of network structures are introduced: network structures with fixed and arbitrary numbers of network model elements. Problems of network structures identification as multiple decision problems are considered. It is proved that risk functions of statistical procedures for network structures identification can be considered as linear combination of mean numbers of errors of incorrect inclusions and exclusions elements from network model to network structure.

Keywords: random variables network, network model, network structure, identification procedure, first kind errors, second kind errors, additive loss, risk function.

Citation

Koldanov P.A. Risk function of statistical procedures for network structures identification. *Vestnik TvGU. Seriya: Prikladnaya Matematika* [Herald of Tver State University. Series: Applied Mathematics], 2017, no. 3, pp. 45–59. (in Russian). https://doi.org/10.26456/vtpmk178

References

[1] Jordan M.I. Graphical models. *Statistical Science*, 2004, vol. 19, pp. 140–155. https://doi.org/10.1214/088342304000000026

[2] Gibbs W. *Elementary Principles of Statistical Mechanics*. Yale University press, NewHaven, Connecticut, 1902.

[3] Wright S. The method of path coefficients. *Annals of Mathematical Statistics*, 1934, vol. 5, pp. 161–215. https://doi.org/10.1214/aoms/1177732676

[4] Koller D., Friedman N. *Probabilistic Graphical Models*. MIT Press, Massachusetts, 2009.

[5] Lauritzen S.L., Sheehan N.A. Graphical models for genetic analyses. *Statistical Science*, 2003, vol. 18(4), pp. 489–514. https://doi.org/10.1214/ss/1081443232

[6] Lauritzen S.L. *Graphical Models*. Oxford university press, 1996.

[7] Dempster A.P. Covariance selection. *Biometrics*, 1972, vol. 28, pp. 157–175.

[8] Wermuth N. Analogies between multiplicative models in contingency tables and covariance selection. *Biometrics*, 1976, vol. 32, pp. 95–108.

[9] Edwards D. *Introduction to Graphical Modeling*. Springer-Verlag New York, Inc., 2000.

[10] Anderson T.W. *An Introduction to Multivariate Statistical Analysis*. 3-d edition. Wiley-Interscience, New York, 2003.

[11] Drton M., Perlman M. Model selection for Gaussian concentration graph. *Biometrika*, 2004, vol. 91(3), pp. 591–602. https://doi.org/10.1093/biomet/91.3.591

[12] Drton M., Perlman M. Multiple testing and error control in Gaussian graphical model selection. *Statistical Science*, 2007, vol. 22(3), pp. 430–449. https://doi.org/10.1214/088342307000000113

[13] Liu W. Gaussian graphical model estimation with false discovery rate control. *The Annals of Statistics*, 2013, vol. 41(6), pp. 2948–2978. https://doi.org/10.1214/13-AOS1169

[14] Boginski V., Butenko S., Pardalos Panos M. On structural properties of the market graph. *Innovations in Financial and Economic Networks*, 2003, pp. 29–45.

[15] Mantegna R.N. Hierarchical structure in financial markets. *The European Physical Journal B-Condensed Matter and Complex Systems*, 1999, vol. 11(1), pp. 193–197. https://doi.org/10.1007/s100510050929

[16] Koldanov A.P., Koldanov P.A., Kalyagin V.A. Statistical procedures for the market graph construction. *Computational Statistics & Data Analysis*, 2013, vol. 68, pp. 17–29.

[17] Lehmann E.L. A theory of some multiple decision problems, I. *The Annals of Mathematical Statistics*, 1957, vol. 28(1), pp. 1–25. https://doi.org/10.1214/aoms/1177707034

[18] Lehmann E.L. A theory of some multiple decision problems. II. *The Annals of Mathematical Statistics*, 1957, vol. 28(3), pp. 547–572. https://doi.org/10.1214/aoms/1177706873

[19] Cohen A., Sackrowitz H.B. Monotonicity properties of multiple endpoint testing procedures. *Journal of Statistical Planning and Inference*, 2004, vol. 125, pp. 17–30. https://doi.org/10.1016/j.jspi.2003.10.008

[20] Cohen A., Sackrowitz H.B. Decision theory results for one-sided multiple comparison procedures. *The Annals of Statistics*, 2005, vol. 33(1), pp. 126–144. https://doi.org/10.1214/009053604000000968

[21] Genovese C., Wasserman L. Operating characteristics and extensions of the false discovery rate procedure. *Royal Statistical Society*, 2002, vol. 64(3), pp. 499–517. https://doi.org/10.1111/1467-9868.00347