ЗАДАЧИ ИДЕНТИФИКАЦИИ СЕТЕВЫХ СТРУКТУР

Колданов Петр Александрович

Национальный исследовательский университет Высшая Школа Экономики
Лаборатория Алгоритмов и Технологий Анализа Сетевых Структур (ЛАТАС)
Нижний Новгород, Россия
pkoldanov@hse.ru

Саратов, 14-15 ноября 2019 года

Outline

Сетевая модель

Одно из основных направлений анализа сложных объектов заключается в построении и исследовании соответствующей сетевой модели.

- Простой полный взвешенный граф $G = (V, \Gamma), \Gamma = ||\gamma_{i,j}||$.
- Вершины графа - элементы объекта.
- веса ребер определяются мерой взаимодействия $\gamma_{i,j}$ между элементами.

При рассмотрении сложных объектов случайной природы, т.е. объектов, поведение элементов которых характеризуется случайными величинами, вершинам сетевой модели соответствуют случайные величины.

Примеры: рыночные сети, биологические сети.

Сетевые структуры

Сетевые структуры - невзвешенные подграфы сетевой модели.

- Отсеченный граф (TG) сетевой модели $G = (V, \Gamma)$ - подграф $G'(\gamma_0) = (V, E) : E = \{(i, j) : \gamma_{i,j} > \gamma_0\}$, где γ_0 - заданный порог.
- Максимальная клика - полный подграф отсеченного графа максимального размера.
- Максимальное независимое множество - пустой подграф отсеченного графа максимального размера.
- Максимальное остовное дерево (MST) сетевой модели $G = (V, \Gamma)$ - дерево (граф без циклов) $G' = (V, E') : E' \subset E; |E'| = |V| - 1;$ такое что $\sum_{(i,j)\in E'} \gamma_{i,j}$ максимальна.
- Редуцированный по вершинам граф сетевой модели $G = (V, \Gamma)$ - подграф $G' = (V', G') : V' \subset V, G' \subset G$

Описание проблемы

Проблема идентификации или выделения сетевой структуры заключается в выборе одного из многих решений о её составе по результатам наблюдений над функционированием сложной системы.

Обзор. Gaussian graphical model.

- Лауритзен (1990) - gaussian graphical model(concentration graph), частный коэффициент корреляции.
- Эдвардс (1995) - обзор статистических процедур.
- Андерсон (2003) - тест максимального правдоподобия.
- Дртон, Перлман (2004, 2007) - многошаговые статистические процедуры множественной проверки гипотез. Дртон (2017) - обзор современного состояния исследований.

Ограничения: асимптотика, нормальное распределение. Контроль FWER, другие свойства неисследованы.

Обзор. Multiple hypotheses testing

- Тьюки, Шеффе - 50-е годы.
- Габриель, Холм, Хочберг, Тамхан, Бенжамини - в конце 20-го столетия. Библиография Рао и Шварупчанд (2009) - 573 работы.

Контроль FWER, FDR, FDP.

- Бахадур, Леман - функция риска.
- Итон(1967), Спжтволл(1972), Коуэн, Сакровитц (2005, 2007) - задачи о математических ожиданиях нормально распределенных случайных величин.

Ограничения: независимые случайные величины или частный случай ковариационной матрицы - intraclass type.

Обзор. Фондовый рынок

- Mantegna(1999) - MST для сети фондового рынка. Коэффициент корреляции Пирсона между доходностями акций.
- Pardalos (2003) - TG для сети фондового рынка.
- В настоящее время около 3000 работ.
- Основная цель - построение сетевых структур численными алгоритмами и интерпретация полученных результатов.
- Проблема - неопределенность полученных результатов, связанная со случайным характером ограниченного числа наблюдений.
- Для изучения неопределенности и разработки методов её возможного уменьшения необходима математическая модель.
- Эллиптическая модель распределения доходностей.
- Актуальными становятся задачи построения методов идентификации сетевых структур, устойчивых в этом широком классе моделей.

Ограниченность полученных результатов

Таким образом, полученные результаты ограничены:

- исследованием нормального распределения;
- использованием в качестве меры связи коэффициента корреляции Пирсона или частного коэффициента корреляции,
- асимптотическим характером полученных результатов,
- контролем только вероятностей ошибок первого рода,
- отсутствием оценки неопределенности полученных результатов.

Возникающие вопросы

- Какая из популярных сетевых структур обладает меньшей статистической неопределенностью?
- Какими свойствами обладают используемые методы идентификации сетевых структур?
- Как строить процедуры, обладающие свойствами оптимальности и устойчивости?
- Какую меру близости целесообразно использовать при построении сетевой модели?

Модель - сеть случайных величин

Сеть случайных величин - пара (X, γ):

- $X = (X_1, \ldots, X_N)-$ случайный вектор с плотностью $f(x, \theta), \theta \in \Omega$,
- $\gamma-$мера зависимости между двумя случайными величинами.

Рассматриваемые сети случайных величин

- Гауссовская сеть корреляций Пирсона: $X = (X_1, \dots, X_N)$ - $N(\mu, \Sigma)$, мера зависимости $\gamma^P_{i,j} = \rho_{i,j} = \frac{\sigma_{i,j}}{\sqrt{\sigma_{i,i}\sigma_{j,j}}}$
- Гауссовская сеть частных корреляций: $\gamma^{part}_{i,j} = |\rho^{i,j}| = \left|\frac{-\sigma^{i,j}}{\sqrt{\sigma^{i,i}\sigma^{j,j}}}\right|$
- Эллиптическая сеть вероятностей совпадения знаков: $X = (X_1, \dots, X_N)$ распределён с плотностью $f(x;\theta) = |\Lambda|^{-\frac{1}{2}} g\{(x-\mu)'\Lambda^{-1}(x-\mu)\}$, $X - EC(\mu, \Lambda, g)$ [1], мера зависимости $\gamma^{Sg}_{i,j} = p^{i,j} = P((X_i - \mu_i)(X_j - \mu_j) > 0)$.
- Эллиптическая сеть корреляций Пирсона.
- Эллиптическая сеть корреляций Кендалла $\gamma^{Kd}_{i,j} = 2P((X^1_i - X^2_i)(X^1_j - X^2_j) > 0) - 1$.
- Эллиптическая сеть корреляций Блумквиста - Краскала $\gamma^{Kr}_{i,j} = 2P((X_i - Med(X_i))(X_j - Med(X_j)) > 0) - 1$.

[1] $\theta = (\mu, \Lambda, g)$, $\mu \in R^N$, Λ - симметричная положительно определенная матрица, $g(x) \geq 0$, $\int_{-\infty}^{\infty} \dots \int_{-\infty}^{\infty} g(y'y)dy_1 \dots dy_N = 1$

Сеть случайных величин, сетевая модель, сетевые структуры

- Сеть случайных величин порождает сетевую модель.
- Разные сети случайных величин могут порождать одинаковые сетевые модели.
- Разные сетевые модели могут порождать одинаковые сетевые структуры

Общая постановка задачи идентификации.

- $\beta : \beta = 1, \dots, K$ - элемент сетевой модели $G = (V, \Gamma)$, $G' = (V, E)$-идентифицируемая сетевая структура, $\beta \in V, K = N$ или $\beta \in E, K = C_N^2$.
- пусть $h_\beta : \theta \in \omega_\beta$ - элемент β сетевой модели не принадлежит идентифицируемой структуре, $k_\beta : \theta \in \omega_\beta^{-1}$ - альтернатива h_β.
- $H_i : \theta \in \Omega_i$ $(i = 1, \dots, L)$ - элементы $\{i_1, i_2, \dots, i_M\} \subseteq \{1, 2, \dots, K\}$ принадлежат идентифицируемой сетевой структуре. M - число элементов идентифицируемой сетевой структуры, т.е.

$$H_i : \theta \in \Omega_i$$

где

$$\Omega_i = \left(\bigcap_{i_l \in \{i_1, \dots, i_M\}} \omega_{i_l}^{-1}\right) \cap \left(\bigcap_{i_s \in \{1, \dots, K\} - \{i_1, \dots, i_M\}} \omega_{i_s}\right) \qquad (1)$$

или

$$H_i = \left(\bigcap_{i_l \in \{i_1, \dots, i_M\}} k_{i_l}\right) \cap \left(\bigcap_{i_s \in \{1, \dots, K\} - \{i_1, \dots, i_M\}} h_{i_s}\right)$$

- Требуется построить процедуру δ выбора (по наблюдениям) одной из гипотез (1). Заметим, что $L \leq 2^{\frac{N(N-1)}{2}}$

Постановка задачи. Селекция по ребрам.

- (X, γ)-сеть случайных величин, $G = (V, \Gamma)$-порожденная сетевая модель.
- $G' = (V', E') : V' \subseteq V, E' \subseteq E$ - сетевая структура.
- X имеет распределение из класса $\mathcal{K} = \{(f(x, \theta), \theta \in \Omega)\}$.
- Пусть $S = (s_{i,j})$, $S \in \mathcal{G}$ - множество матриц смежности.
- $H_S : \theta \in \Omega_S$-гипотеза, что сетевая структура имеет матрицу смежности $S, S \in \mathcal{G}_1 \subseteq \mathcal{G}$.
- Наблюдения $X(t) = (X_1(t), \ldots, X_N(t)), t = 1, \ldots, n$-повторная выборка.

Задача: построить нерандомизированную статистическую процедуру $\delta(x)$ выбора одной из попарно несовместных гипотез H_S, обладающую желательными свойствами.

Желательные свойства.

- оптимальность процедур идентификации сетевых структур в некотором классе.
- устойчивость, т.е. инвариантность функции риска процедур идентификации сетевых структур в некотором классе при произвольной функции потерь.

Два типа задач.

- задачи идентификации сетевой структуры, которая может содержать произвольное число элементов сетевой модели: отсеченный граф; граф концентраций; редуцированный по вершинам граф.
- задачи идентификации сетевой структуры, которая содержит заданное число элементов: максимальное остовное дерево $(M = N - 1)$; задачи выбора M наибольших (наименьших) элементов.

Процедуры идентификации сетевых структур с произвольным числом элементов. Отсеченный граф

- Пусть

$$\Phi(x) = \begin{pmatrix} 0 & \varphi_{1,2}(x) & \dots & \varphi_{1,N}(x) \\ \varphi_{2,1}(x) & 0 & \dots & \varphi_{2,N}(x) \\ \dots & \dots & \dots & \dots \\ \varphi_{N,1}(x) & \varphi_{N,2}(x) & \dots & 0 \end{pmatrix}$$

где $\varphi_{i,j}(x) = \begin{cases} 1, & x \in A_{i,j}^{-1} \\ 0, & x \in A_{i,j} \end{cases}$ – нерандомизированные тесты гипотез $h_{i,j}$ - ребро (i,j) не принадлежит TG.

$$\delta(x) = d_G \text{ если } \Phi(x) = G \tag{2}$$

где d_G - решение о том, что истинна гипотеза H_G.

Т.о. задача выбора одной из $2^{\frac{N(N-1)}{2}}$ гипотез сводится к задаче одновременной проверки $\frac{N(N-1)}{2}$ гипотез.

Процедуры идентификации сетевых структур с заданным числом элементов.

- Соотношение
$$\delta(x) = d_G \text{ если } \Phi(x) = G \ (2)$$
и в этом случае определяют общий вид процедур идентификации сетевых структур.
- Однако в этом случае должно выполняться условие совместности, которое имеет вид:

$$\sum_{(\kappa_{i\beta_{i_1}},\ldots,\kappa_{i\beta_{i_K}}):\kappa_{i\beta_{i_1}}=\ldots=\kappa_{i\beta_{i_M}}=-1;\kappa_{i\beta_{i_{M+1}}}=\ldots=\kappa_{i\beta_{i_K}}=1} P_\theta(x \in \bigcap_\beta A_\beta^{\kappa_{i\beta}}) = 1 \quad (3)$$

Ошибки 1-го и 2-го родов

Критерий качества процедур идентификации сетевых структур - число различных элементов двух матриц смежности: истинной ($g_{i,j}$) и выборочной ($\varphi_{i,j}$).
$Y_I(S,\delta) = \sum_{i<j} I(g_{i,j} = 0, \varphi_{i,j}(x) = 1) -$ число ошибочно проведенных ребер процедурой $\delta(x)$ с матрицей смежности ($\varphi_{i,j}(x)$) при идентификации сетевой структуры S с матрицей смежности ($g_{i,j}$).
$Y_{II}(S,\delta) = \sum_{i<j} I(g_{i,j} = 1, \varphi_{i,j}(x) = 0) -$ число ошибочно непроведенных ребер процедурой $\delta(x)$.
Характеристика процедуры $\delta(x)$ - вектор

$$(E_\theta Y_I(S,\delta), E_\theta Y_{II}(S,\delta)) \tag{4}$$

Качество процедур. Концепция Вальда

- $\delta(x) = d_Q$ - решение, что сетевая структура имеет матрицу смежности Q, $Q \in \mathcal{G}$.
- $w(H_S; d_Q) = w(S, Q)$ - потери от решения d_Q когда гипотеза H_S истинна, $w(S, S) = 0, S \in \mathcal{G}$.
- Функция риска статистической процедуры $\delta(x)$ определяется

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

$P_\theta(\delta(x) = d_Q)$ - вероятность принятия решения d_Q когда правильно решение d_S.

Аддитивность функции потерь

Естественно предположить, что $w(S, Q)$ тем больше, чем больше различаются матрицы S и Q.

$$w_{ij} = \sum_{\beta}(\epsilon_{ij\beta}a_{\beta} + \epsilon_{ji\beta}b_{\beta}) \tag{5}$$

a_{β}, b_{β}- потери от ложного отвержения и принятия гипотез h_{β}, w_{ij}− потери от принятия решения d_j при правильном решении d_i,

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

Функция риска.

Теорема 1:

- *пусть функция потерь аддитивна и задается (5). Тогда функция риска статистической процедуры δ идентификации сетевой структуры с произвольным числом элементов имеет вид:*

$$R(S,\theta,\delta) = \sum_{\beta=1}^{K} r(h_\beta,\varphi_\beta) \tag{6}$$

где $r(h_\beta,\varphi_\beta)$- функция риска тестов φ_β проверки h_β,

- *если $a_\beta = a, b_\beta = b, \forall\beta = 1,\dots,K$, то*

$$R(S,\theta,\delta) = aE_\theta(Y_I(S,\delta)) + bE_\theta(Y_{II}(S,\delta)) \tag{7}$$

где $Y_I(S,\delta)-$ число элементов, неправильно включенных (число ошибок $1-$го рода) процедурой δ в сетевую структуру S; $Y_{II}(S,\delta)-$ число элементов, неправильно невключенных (число ошибок $2-$го рода) процедурой δ в сетевую структуру S.

Функция риска.

Теорема 2: *Пусть*

- *семейство тестов φ_β индивидуальных гипотез h_β совместно с задачей различения гипотез H_i, т.е. выполняется условие (3);*
- *функция потерь аддитивна и задается (5); Тогда функция риска статистической процедуры δ для задач идентификации сетевых структур с заданным числом элементов имеет вид:*

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

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

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

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

где $Y_I(S,\delta)$, $Y_{II}(S,\delta)$- число ошибок $1-$ го, $2-$ го рода, допущенных процедурой δ при идентификации сетевой структуры S.

Статистическая неопределенность

- Статистическая неопределенность процедуры δ идентификации сетевой структуры S по выборке объема n -

$$R(S,\theta,\delta,n) = aE_\theta(Y_I(S,\delta,n)) + bE_\theta(Y_{II}(S,\delta,n)) \quad \forall\theta \in \Omega_S \tag{10}$$

- Процедура δ_1 идентификации структуры S предпочтительней, чем процедура δ_2 идентификации структуры S, если

$$R(S,\theta,\delta_1,n) < R(S,\theta,\delta_2,n) \quad \forall\theta \in \Omega; \forall n \tag{11}$$

- **Процедура δ_1 идентификации структуры S_1 предпочтительней в области $\Omega_1 \subset \Omega$, чем процедура δ_2 идентификации структуры S_2, если**

$$R(S_1,\theta,\delta_1,n) < R(S_2,\theta,\delta_2,n) \quad \forall\theta \in \Omega_1; \forall n \tag{12}$$

Результаты статистического моделирования.

Исследована статистическая неопределенность некоторых сетевых структур. При уровне статистической неопределенности ≤ 0.1.

- Клики максимального веса - $n = 150$ при всех порогах.
- MG - $n = 300$ при всех порогах.
- MIS минимального веса - $n = 700$ при всех порогах.
- MST - $n > 10000$

Статистическая неопределенность отсеченного графа и связанных с ним структур значительно меньше статистической неопределенности максимального остовного дерева и связанных с ним структур

Свойства оптимальности применяемых процедур идентификации сетевых структур.

- процедура $\delta^\star \in C$ оптимальна в классе C, если

$$R(\theta,\delta^\star) \le R(\theta,\delta)$$
$$\forall\theta\in\Omega, \forall\delta\in C \qquad (13)$$

- аддитивная функция потерь, процедуры идентификации сетевых структур с произвольным числом элементов:

$$aE_\theta(Y_I(S,\delta^\star)) + bE_\theta(Y_{II}(S,\delta^\star)) \le aE_\theta(Y_I(S,\delta)) + bE_\theta(Y_{II}(S,\delta))$$
$$\forall\theta\in\Omega_S, \forall S\in\mathcal{G}, \forall\delta\in C \qquad (14)$$

- аддитивная функция потерь, процедуры идентификации сетевых структур с заданным числом элементов:

$$E_\theta(Y_I(S,\delta^\star)) \le E_\theta(Y_I(S,\delta))$$
$$\forall\theta\in\Omega_S, \forall S\in\mathcal{G}_1\subset\mathcal{G}, \forall\delta\in C \qquad (15)$$

Условие несмещенности.

- Статистическая процедура $\delta(x)$ называется W-несмещенной, если

$$E_\theta w(\theta, \delta(x)) \leq E_\theta w(\theta', \delta(x)) \quad \forall \theta, \theta' \in \Omega \tag{16}$$

- Задачи идентификации сетевых структур с произвольным числом элементов при аддитивной функции потерь:

$$aE_\theta(Y_I(S,\delta)) + bE_\theta(Y_{II}(S,\delta)) \leq aE_\theta(Y_I(S',\delta)) + bE_\theta(Y_{II}(S',\delta)) \quad \forall \theta \in \Omega_S, \forall S, S' \in \mathcal{G}, \forall \delta \in C \tag{17}$$

- Задачи идентификации сетевых структур с заданным числом элементов при аддитивной функции потерь:

$$E_\theta(Y_I(S,\delta)) \leq E_\theta(Y_I(S',\delta)) \quad \forall \theta \in \Omega_S, \forall S, S' \in \mathcal{G}_1 \subset \mathcal{G}, \forall \delta \in C \tag{18}$$

Оптимальность процедуры идентификации гауссовской графической модели.

- Построен РНМ в классе несмещенных тест проверки гипотезы $h_{i,j} : \rho^{i,j} = 0$ против альтернативы $k_{i,j} : \rho^{i,j} \neq 0$. Такой тест имеет вид:

$$\varphi_{i,j} = \begin{cases} 0, & 2q - 1 < \frac{as_{i,j} - b/2}{\sqrt{b^2/4 + ac}} < 1 - 2q \\ 1, & \text{иначе} \end{cases} \qquad (19)$$

  где $q-$квантиль бета-распределения $F_{Be}(q(\alpha_{ij})) = \frac{\alpha_{ij}}{2}$, $a, b, c-$ коэффициенты в представлении $\det(s_{k,l})$ полиномом второй степени от $s_{i,j}$, $\det(s_{k,l}) = -as_{i,j}^2 + bs_{i,j} + c$, $||s_{k,l}||-$ выборочная ковариационная матрица.

Оптимальность процедуры идентификации гауссовской графической модели.

- стандартный тест проверки гипотезы об условной независимости в многомерном нормальном распределении имеет вид:

$$\varphi_{i,j} = \begin{cases} 0, & |r^{i,j}| \leq c_{i,j} \\ 1, & |r^{i,j}| > c_{i,j} \end{cases} \tag{20}$$

где $c_{i,j}$ является $(1-\alpha/2)$-квантилем распределения с плотностью

$$f(x) = \frac{1}{\sqrt{\pi}} \frac{\Gamma(n-N+1)/2)}{\Gamma((n-N)/2)} (1-x^2)^{(n-N-2)/2}, \quad -1 \leq x \leq 1$$

Оптимальность процедуры идентификации гауссовской графической модели.

- Теорема 3. *Тест (20) эквивалентен РНМ в классе несмещенных тесту (19) проверки гипотезы* $\rho^{i,j} = 0$ *против альтернативы* $\rho^{i,j} \neq 0$.
- Теорема 4. *Статистическая процедура*

$$\delta(x) = d_G \text{ если } \Phi(x) = G$$

где $\Phi(x) = ||\varphi_{i,j}(x)||$, $\varphi_{i,j}(x)$ *имеют вид (20) или (19), является оптимальной в классе несмещенных процедурой идентификации графа концентраций при аддитивной функции потерь.*

Оптимальность процедуры идентификации редуцированного по вершинам графа.

Индивидуальные гипотезы $h_i : \mu_i \leq \mu_0, i = 1, \ldots, N$, где $(\mu_1, \mu_2, \ldots, \mu_N)$- вектор математических ожиданий нормального распределения $N(\mu, \Sigma)$.

Теорема 5. *Вектор $(X_1, \ldots, X_N)$ - $N(\mu, \Sigma)$ с неизвестным μ и известными элементами $\sigma_{i,i}$ матрицы Σ, функция потерь W является аддитивной.*

- *Процедура*

$$\delta(x) = (\varphi_1(x), \varphi_2(x), \ldots, \varphi_N(x)) \,. \tag{21}$$

основанная на индивидуальных тестах (22)

$$\varphi_i(x) = \begin{cases} 1, & \sqrt{n}\frac{\overline{x_i} - \mu_0}{\sqrt{\sigma_{ii}}} > c \\ 0, & \textit{иначе} \end{cases} \tag{22}$$

является оптимальной в классе W-несмещенных процедурой идентификации редуцированного по вершинам графа.

- *Функция риска процедуры (21) не зависит от корреляционной матрицы.*

Свойства процедуры идентификации отсеченного графа.

Введен класс $\mathcal{D}$ статистических процедур $\delta(x)$:

1. $\delta(x)$ - инвариантные статистические процедуры по отношению к группе $G^{c,d}: y = cx + d$.
2. $R(S, \theta, \delta)$ непрерывна по θ.
3. элементы $\varphi_{i,j}(x)$ матрицы $\Phi(x)$ зависят только от $x_i(t), x_j(t)$.

Теорема 6. *Пусть (X, γ) - Гауссовская сеть корреляций Пирсона, функция потерь аддитивна (5) и потери $a_{i,j}, b_{i,j}$ связаны с уровнем значимости $\alpha_{i,j}$ соотношением $a_{i,j} = 1 - \alpha_{i,j}; b_{i,j} = \alpha_{i,j}$. Тогда статистическая процедура $\delta(x) = d_G$ если $\Phi(x) = G$ где элементы матрицы Φ имеют вид*

$$\varphi_{i,j}(x_i, x_j) = \begin{cases} 1, & \frac{r_{i,j} - \rho_0}{\sqrt{1 - r_{i,j}^2}} > c_{i,j} \\ 0, & \frac{r_{i,j} - \rho_0}{\sqrt{1 - r_{i,j}^2}} \leq c_{i,j} \end{cases} \tag{23}$$

является оптимальной статистической процедурой идентификации отсеченного графа в классе $\mathcal{D}$.

Сеть случайных величин, сетевая модель, сетевые структуры

- Сеть случайных величин порождает сетевую модель.
- Разные сети случайных величин могут порождать одинаковые сетевые модели.
- Разные сетевые модели могут порождать одинаковые сетевые структуры
- Рассмотрим вопрос выбора меры зависмости в сети случайных величин.

Эллиптическая сеть вероятностей совпадения знаков.

- Новый класс моделей сетей случайных величин - класс эллиптических сетей вероятностей совпадения знаков (X, γ^{Sg}).
- $X = (X_1, \ldots, X_N)$ имеет эллиптическое распределение с плотностью

$$f(x) = |\Lambda|^{-\frac{1}{2}} g\{(x-\mu)'\Lambda^{-1}(x-\mu)\} \tag{24}$$

- мера зависимости имеет вид:

$$\gamma^{Sg}_{i,j} = p^{i,j} = P\{(X_i - \mu_i)(X_j - \mu_j) > 0\} \tag{25}$$

Эквивалентность сетевых структур в эллиптической сети корреляции Пирсона.

- Класс $\mathcal{K}$ распределений вектора X такой, что при фиксированной γ сетевые модели, порожденные $(X^{(1)}, \gamma)$, $(X^{(2)}, \gamma)$ совпадают, т.е.

$$\gamma(X_i^{(1)}, X_j^{(1)}) = \gamma(X_i^{(2)}, X_j^{(2)}), \forall X^{(1)}, X^{(2)} \in \mathcal{K}, \forall i, j = 1, \ldots, N$$

- Для всех распределений из $\mathcal{K}$ сетевые структуры также совпадают.
- В классе эллиптических распределений выделяется подкласс $\mathcal{K}(\Lambda)$ распределений с фиксированной матрицей Λ. Так как $\gamma_{i,j}^P = \lambda_{i,j}/\sqrt{\lambda_{i,i}\lambda_{j,j}}$, $\Lambda = (\lambda_{i,j})$, то сетевые модели, порожденные сетями случайных величин $(X, \gamma^P), X \in \mathcal{K}(\Lambda)$ совпадают.

Эквивалентность сетевых структур в эллиптической сети вероятностей совпадения знаков.

- Теорема 7: *Если случайный вектор* $X = (X_i, X_j)$ *имеет плотность*

$$f(x_i, x_j) = \begin{vmatrix} a_{i,i} & a_{i,j} \\ a_{i,j} & a_{j,j} \end{vmatrix}^{\frac{1}{2}} g(a_{i,i}x_1^2 + 2a_{i,j}x_1x_2 + a_{j,j}x_2^2)$$

то вероятность совпадения знаков $\gamma_{i,j}^{Sg}$ *не зависит от* g.

- Т.е. сетевые модели, порожденные сетями случайных величин $(X, \gamma^{Sg}), X \in \mathcal{K}(\Lambda)$ совпадают.

Эквивалентность сетевых структур при разных мерах

- Теорема 8: *Если случайный вектор* $X = (X_1, X_2, \ldots, X_N)$ *имеет распределение из класса* $EC(\mu, \Lambda, g)$, *то отсеченный граф в сети корреляции Пирсона с порогом* ρ_0 *совпадает с отсеченным графом в знаковой сети с порогом* $p_0 = \frac{1}{2} + \frac{1}{\pi}\arcsin(\rho_0)$.
- Теорема 9: *Если случайный вектор* $X = (X_1, X_2, \ldots, X_N)$ *имеет распределение из класса* $EC(\mu, \Lambda, g)$, *то максимальное остовное дерево в сети корреляции Пирсона совпадает с максимальным остовным деревом в сети корреляции знаков.*

Эти теоремы показывают, что в классе эллиптических распределений $EC(\mu, \Lambda, g)$ сетевые структуры в сети Пирсона и знаковой сети определяются матрицей Λ, не зависят от g и находятся во взаимно-однозначном соответствии.

Процедура идентификации TG, основанная на частоте совпадения знаков

$$h^{Sg}_{i,j} : p^{i,j} \leq p_0 \text{ против альтернатив } k^{Sg}_{i,j} : p^{i,j} > p_0$$

$$\varphi^{Sg}_{i,j} = \begin{cases} 1, & v^{Sg}_{i,j} > c_{i,j} \\ 0, & v^{Sg}_{i,j} \leq c_{i,j} \end{cases} \tag{26}$$

$$v^{Sg}_{i,j} = T^{i,j}_{0,0} + T^{i,j}_{1,1} \tag{27}$$

$$T^{i,j}_{0,0} = \sum_{t=1}^{n} (1 - u_i(t))(1 - u_j(t)),\ T^{i,j}_{1,1} = \sum_{t=1}^{n} u_i(t)u_j(t)$$

$$u_k(t) = \begin{cases} 1, & x_k(t) > 0 \\ 0, & x_k(t) \leq 0 \end{cases} \tag{28}$$

$$\sum_{k=c_{i,j}}^{n} \frac{n!}{k!(n-k)!} (p_0)^k (1 - p_0)^{n-k} \leq \alpha_{i,j} \tag{29}$$

Процедура идентификации TG, основанная на частоте совпадения знаков

Процедура идентификации TG, основанная на частоте совпадения знаков, имеет вид

$$\delta^{Sg}(x) = d_G, \quad \text{если} \quad \Phi^{Sg}(x) = G \tag{30}$$

где

$$\Phi^{Sg}(x) = \begin{pmatrix} 0, & \varphi_{12}^{Sg}(x), & \dots, & \varphi_{1N}^{Sg}(x) \\ \varphi_{21}^{Sg}(x), & 0, & \dots, & \varphi_{2N}^{Sg}(x) \\ \dots & \dots & \dots & \dots \\ \varphi_{N1}^{Sg}(x), & \varphi_{N2}^{Sg}(x), & \dots, & 0 \end{pmatrix}. \tag{31}$$

Устойчивость процедуры δ^{Sg}

- *Статистическая процедура δ идентификации сетевой структуры S в сетевой модели $G = (V, \Gamma)$, порожденной сетью случайных величин $(X, \gamma) : X \in EC(\mu, \Lambda, g)$, имеет инвариантную функцию риска (устойчива) в классе $\mathcal{K}(\Lambda)$, если функция риска $R(S, \theta, \delta), \theta = (\mu, \Lambda, g)$ не зависит от g.*
- Теорема 10: *Пусть случайный вектор $(X_1, \dots, X_N)$ имеет распределение $EC(\mu, \Lambda, g), \theta = (\mu, \Lambda, g)$ с известным μ. Тогда функции риска процедур идентификации отсеченного графа $R(S, \theta, \delta^{SS})$, $R(S, \theta, \delta^{H})$, $R(S, \theta, \delta^{Hg})$ (δ^{SS}-одношаговая процедура, δ^{H}-процедура Холма, δ^{Hg}- процедура Хочберга, основанные на статистиках $V_{i,j}^{Sg}$) и функция риска процедуры Краскала идентификации максимального остовного дерева определяются матрицей Λ и не зависят от функции g при любой функции потерь $w(S, Q)$.*

Неустойчивость процедур идентификации сетевых структур, основанных на выборочном коэффициенте корреляции Пирсона.

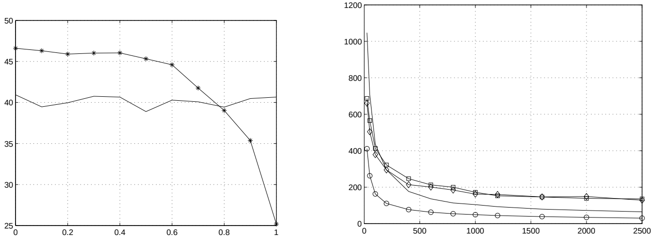

Figure: 2: Функция риска для TG, $\rho_0 = 0.64$. Слева - $n = 400$, линия со звездой - δ^P, линия - δ^S, по горизонтали - параметр смеси ϵ; справа: окружность - $\epsilon = 1$, δ^P; ромб - $\epsilon = 0,5$, δ^P; квадрат - $\epsilon = 0$, δ^P, линия - δ^S, по горизонтали - число наблюдений. Смесь распределений нормального с весом ϵ и Стьюдента с 3 степенями свободы с весом $1 - \epsilon$.

Резюме: вероятностная модель

- популярность корреляции Пирсона, как меры зависимости в сетевом анализе.
- можно рассмотреть новый класс сетевых моделей, в которых мера зависимости определяется вероятностью совпадения знаков.
- эквивалентность сетевых структур в знаковой сети и в сети корреляций Пирсона в классе эллиптических распределений.
- Какую меру выбирать при построении сетевых моделей?
- Ответ: в классе эллиптических распределений сети случайных величин с мерами близости, такими как Пирсона, Кендалла, Блумквиста-Краскала и вероятности совпадения знаков, сетевые модели эквивалентны.
- простота интепретации - $\gamma_{ij}^{P} = 0,1 \leftrightarrow \gamma_{ij}^{Sg} = 0,53; \gamma_{ij}^{P} = 0,6 \leftrightarrow \gamma_{ij}^{Sg} = 0,705$

Какую статистическую процедуру использовать?

- **гауссовская модель или близкие к ней - выборочный коэффициент корреляции Пирсона.**
- **негауссовские эллиптические модели - частота совпадения знаков.**
- **частота совпадения знаков может быть использована для идентификации сетевых структур в эллиптической сети корреляции Пирсона и наоборот.**

Применение к задачам анализа фондовых рынков

- проверка гипотезы симметрии совместного распределения доходностей акций фондового рынка.
- построение оптимального портфеля инвестиций.

Эффективный фронт портфелей акций

4-х ступенчатая процедура построения портфеля инвестиций

- идентификация редуцированного по вершинам графа (по отношению Шарпа).
- идентификация редуцированного по ребрам графа.
- идентификация максимального независимого множества.
- построение портфеля инвестиций на акциях максимального независимого множества.

Спасибо за внимание

Публикации

- Kalyagin V. A., Koldanov A. P., Koldanov P., Pardalos P. M. Loss function, unbiasedness, and optimality of Gaussian graphical model selection // Journal of Statistical Planning and Inference. 2019. Vol. 201. P. 32-39.
- Koldanov P.A. Invariance Properties of Statistical Procedures for Network Structures Identification. Springer Proceedings in Mathematics & Statistics, 2018, vol 247, 289-297.
- Kalyagin V. A., Koldanov A. P., Koldanov P., Pardalos P. M. Optimal decision for the market graph identification problem in a sign similarity network // Annals of Operations Research, 2018, vol. 266, 313-327.

Публикации

- Колданов П. А. Функция риска статистических процедур идентификации сетевых структур // Вестник Тверского государственного университета. Серия: Прикладная математика. —2017. —№ 3. — С. 45-59.
- P. A. Koldanov, A. P. Koldanov, V. A. Kalyagin, P. M. Pardalos. Uniformly most powerful unbiased test for conditional independence in gaussian graphical model //Statistics & Probability Letters. —2017. —Vol.122 —P. 90–95
- V. A. Kalyagin, A. P. Koldanov, P. A. Koldanov Robust identification in random variables networks/ // Journal of Statistical Planning and Inference. —2017. —Vol. 181. —P. 30–40.
- П. А. Колданов, , А.П. Колданов, В.А. Калягин, П.М. Пардалос Статистические процедуры идентификации сетевых структур фондовых рынков // Журнал Новой Экономической Ассоциации. —2017. —Т.3. —№35. —С.33-52

Публикации

- P. A. Koldanov, N. N. Lozgacheva Multiple testing of sign symmetry for stock return distributions/ // International Journal of Theoretical and Applied Finance. —2016. —Vol.19, issue 8 —P. 1650049–1–1650049–14.
- P. A. Koldanov, V. A. Kalyagin, A. P. Koldanov, V. A. Zamaraev. Market Graph and Markowitz Model/ // Optimization in Science and Engineering (In Honor of the 60th Birthday of Panos M. Pardalos). Springer Science, Business Media. —2014. —P. 301–313.
- V. A. Kalyagin, A. P. Koldanov, P. A. Koldanov et al Measures of uncertainty in market network analysis/ // Physica A: Statistical Mechanics and its Applications. —2014. —Vol. 413, issue 1. —P. 59–70.

Публикации

- P. A. Koldanov, G. A. Bautin Multiple decision problem for stock selection in market network/ // Learning and Intelligent Optimization, Lecture Notes in Computer Science. —2014. —Vol. 8426. —P. 98–110.
- G. A. Bautin, V. A. Kalyagin, P. A. Koldanov et al. Simple measure of similarity for the market graph construction // Computational Management Science. —2013. —Vol.10. —P. 105–124.
- A. P. Koldanov, P. A. Koldanov, V. A. Kalyagin, P. M. Pardalos. Statistical procedures for the market graph construction// Computational Statistics & Data Analysis. —2013. —Vol.68 —P. 17–29.