Московский Государственный Университет
им. М.В. Ломоносова
Механико-математический факультет

На правах рукописи
УДК 519.237.5, 519.237.3

Суханова Екатерина Михайловна

МАТРИЧНОЗНАЧНЫЕ КОРРЕЛЯЦИОННЫЕ МЕРЫ И МНОГОМЕРНЫЕ ТЕСТЫ НЕЗАВИСИМОСТИ

Специальность: 01.01.05 – Теория вероятностей и математическая статистика

АВТОРЕФЕРАТ
диссертации на соискание ученой степени
кандидата физико-математических наук

Москва 2008

Работа выполнена в Московском Государственном Университете имени М. В. Ломоносова на кафедре теории вероятностей механико-математического факультета.

Научный руководитель: доктор физико-математических наук, профессор **Тюрин Юрий Николаевич**.

Официальные оппоненты: доктор физико-математических наук **Зубков Андрей Михайлович**, Математический Институт РАН имени В. А. Стеклова;

кандидат физико-математических наук **Уфимцев Михаил Валентинович**, Московский Государственный Университет имени М. В. Ломоносова.

Ведущая организация: Центральный Экономико-Математический Институт РАН.

Защита диссертации состоится 12 декабря 2008 года в 16 часов 40 минут на заседании диссертационного совета Д.501.001.85 при Московском Государственном Университете имени М. В. Ломоносова по адресу: 119991, ГСП-1, Москва, Ленинские Горы, Главное Здание МГУ, механико-математический факультет, аудитория 16-24.

С диссертацией можно ознакомиться в библиотеке механико-математического факультета МГУ (Главное здание, 14 этаж).

Автореферат разослан 11 ноября 2008 года.

Ученый секретарь диссертационного
совета Д.501.001.85 при МГУ,
доктор физико-математических наук,
профессор И. Н. Сергеев

1 Общая характеристика работы

Актуальность темы

Задача анализа статистической связи между признаками и, в частности, проверки статистической гипотезы о независимости двух случайных признаков часто встречается в прикладных исследованиях. Классический коэффициент корреляции Пирсона[1], обычно используемый для решения этой задачи, обладает тем недостатком, что он крайне ненадежен при наличии в данных грубых ошибок и при иных отклонениях модели распределения признаков от нормального. Альтернативными мерами взаимозависимости признаков служат непараметрические коэффициенты корреляций, построенные при помощи рангов и знаков. Это — популярные ранговые корреляции Спирмена[2], Кендэлла[3], квадрантная корреляция[4,5] и проч. Данная тематика хорошо освещена, например, в книгах Гаека, Шидака[6] и Кендэлла[7].

Непараметрические методы статистики — это комплекс методов статистической обработки данных, не требующих знания функционального вида генеральных распределений. Потеря информации, возникающая при переходе от точных значений наблюдений к их порядковым номерам (рангам) или знакам, компенсируется широкой применимостью методов и их устойчивостью по отношению к различного рода «выбросам», неточностям моделей и т.д. Поскольку ранговые методы базируются на упорядочении наблюдений, они используются, так же как и знаковые методы, только для вещественных данных. Для многомерных данных, когда результатом наблюдения над каждым объектом является несколько чисел (вектор), к сожалению, не существует естественного способа упорядочения и сравнения. Поэтому опыты многомерного обобщения ранговых и знаковых коэффициентов корреляций актуальны и оправданы.

Интерес к развитию методов многомерного непараметрического корреляционного анализа наблюдается на протяжении нескольких десятилетий вплоть до настоящего времени. Предпринимают много попыток получить адекватные результаты в данной области. Перечислим лишь некоторые из них в хронологическом порядке. Покоординатное ранжирование при постро-

[1] K. Pearson. “Mathematical Contributions to the Theory of Evolution: III. Regression, Heredity, and Panmixia”. — *Philosophical Transactions of the Royal Society of London*, Series A, Vol. 187, pp. 253-318, 1896.

[2] C. Spearman. “The Proof and Measurement of Association Between Two Things”. — *Amer. J. Psychology*, Vol. 15, pp. 72-101, 1904.

[3] M. G. Kendall. “A New Measure of Rank Correlation”. — *Biometrika*, Vol. 30, pp. 81-93, 1938.

[4] F. Mosteller. “On Some Useful ‘Inefficient’ Statistics”. — *Ann. Math. Statist.*, Vol. 17, pp. 377-408, 1946.

[5] N. Blomqvist. “On a Measure of Dependence Between Two Random Variables”. — *Ann. Math. Statist.*, Vol. 21, pp. 593-600, 1950.

[6] Я. Гаек, З. Шидак. *Теория ранговых критериев.* — М.: “Наука”, 1971.

[7] М. Кендэл. *Ранговые корреляции.* — М.: “Статистика”, 1975.

ении многомерного непараметрического критерия независимости применили Puri и Sen[8], но их тестовая статистика не удовлетворяет свойству аффинной инвариантности и, как следствие, ее эффективность зависит от ковариационной структуры наблюдений. Указанная статистика при специальном выборе функций меток служит обобщением квадрантной и спирменовской корреляций. С помощью так называемого углового расстояния между двумя многомерными наблюдениями — т.е. относительного количества гиперплоскостей, порожденных векторзначными данными, и разделяющих эти два наблюдения — Gieser и Randles[9] предложили многомерный вариант знакового квадрантного теста. Хотя полученный критерий аффинно инвариантен и асимптотически свободен от распределения, он весьма неудобен с вычислительной точки зрения. Воспользовавшись пространственным обобщением понятия знака, более практическую многомерную версию квадрантного теста недавно представили Taskinen, Kankainen, Oja[10]. Подобным образом многомерные версии критериев независимости Спирмена и Кендэлла определили Taskinen, Oja, Randles[11]. Общим недостатком упомянутых работ[9-11] можно назвать требование эллиптичности распределений многомерных признаков. Иные подходы к решению описанной задачи предлагали, среди прочего, Питербарг, Тюрин[12], Möttönen, Koshevoy, Oja, Tyurin[13] и Schmid, Schmidt[14].

Большинство рассматриваемых в литературе многомерных вариантов ранговых и знаковых коэффициентов корреляций получены, исходя из интуитивных соображений, связанных с попыткой упорядочить и сравнить многомерные наблюдения. В диссертации предлагается более систематический подход. Сначала мы определяем понятие корреляции векторных случайных величин. Введение матричнозначной корреляционной меры также дополняет работу Тюрина[15], в которой совершенно по-новому излагается линейный многомерный статистический анализ с использованием матриц как обобщений чисел и заданием «матричного скалярного произведения». Матричная корреляция дает простой способ получить различные непараметрические многомерные корреляционные меры и построить с их помощью многомерные критерии

[8]M. L. Puri, P. K. Sen. *Nonparametric methods in Multivariate Analysis.* — N.Y.: Wiley, 1971.

[9]P. W. Gieser, R. H. Randles. “A Nonparametric Test of Independence Between Two Vectors”. — *J. Amer. Statist. Assoc.*, Vol. 92, pp. 561-567, 1997.

[10]S. Taskinen, A. Kankainen and H. Oja. “Sign Test of Independence Between Two Random Vectors”. — *Statist. Probab. Lett.*, Vol. 62, pp. 9-21, 2003.

[11]S. Taskinen, H. Oja and R. H. Randles. “Multivariate Nonparametric Tests of Independence”. — *J. Amer. Statist. Assoc.*, Vol. 100, pp. 916-925, 2005.

[12]В. И. Питербарг, Ю. Н. Тюрин. “Многомерные ранговые корреляции: гауссовское поле на прямом произведении сфер”. — *ТВП*, т. 45, сс. 236-250, 2000.

[13]J. Möttönen, G. Koshevoy, H. Oja and Y. Tyurin. “Multivariate Tests for Independence Based on Zonotopes”. Manuscript, 2005.

[14]F. Schmid, R. Schmidt. “Nonparametric Inference on Multivariate Versions of Blomqvist’s Beta and Related Measures of Tail Dependence”. — *Metrika*, Vol. 66, pp. 323-354, 2007.

[15]Ю. Н. Тюрин. “Многомерный анализ: геометрическая теория”. Манускрипт, 2008.

независимости.

Таким образом, тема диссертации представляется актуальной с теоретической точки зрения, и имеет практическую значимость.

Цель работы

Целью данной диссертации является расширение понятия коэффициента корреляции на случай многомерных величин, построение новых многомерных версий ранговых и знаковых корреляций и тестов независимости, исследование статистических свойств предложенных объектов и процедур.

Научная новизна

Основные результаты диссертации являются новыми и состоят в следующем:

1. Определена новая матричнозначная корреляционная мера и ее выборочный аналог для пары многомерных случайных признаков. Показано, что матричная корреляция в основных чертах повторяет свойства классического коэффициента корреляции с тем отличием, что роль чисел выполняют квадратные матрицы. В гауссовском случае найдено точное распределение выборочной матричной корреляции (при условии, что многомерные случайные признаки независимы) и асимптотическое распределение матричной корреляции при неограниченно растущем объеме выборки n. Также, с помощью матричной корреляции объединены многие понятия многомерного регрессионного и корреляционного анализа.

2. Предложены новые многомерные версии широко известных ранговых коэффициентов корреляций Спирмена, Кендэлла и знаковой квадрантной корреляции. Установлено, что выборочные ранговые и знаковые матричные корреляции (при некоторых слабых условиях регулярности) являются состоятельными $\sqrt{n}$-асимптотически гауссовским оценками своих теоретических аналогов.

3. Построено три новых многомерных непараметрических теста независимости на основе предложенных знаковых и ранговых матричных корреляций. Изучено асимптотическое поведение (при $n \to \infty$) тестовых статистик при гипотезе независимости и при близких альтернативах. Показано, что наши тесты аффинно инвариантны и асимптотически свободны от распределений (при гипотезе независимости). По сравнению с классическими процедурами новые тестовые статистики требуют более

слабых условий относительно моментов распределений признаков (достаточно существования конечных вторых моментов), они могут обладать большей асимптотической мощностью и при этом более устойчивы к «засорениям».

Методы исследования

В работе применяются общие методы теории вероятностей и математической статистики, математического и функционального анализа, а также элементы матричной алгебры. Широко используется теория U-статистик.

Теоретическая и практическая значимость

Работа носит теоретический характер, результаты диссертации расширяют совокупность многомерных статистических методов корреляционного анализа. Предложенные в диссертации критерии могут быть полезны для решения практических задач, связанных с изучением статистической зависимости двух многомерных признаков не очень больших размерностей ($\simeq 10$). Рекомендуется их использование в тех случаях, когда важно свойство аффинной инвариантности или распределение признаков имеет более «тяжелые хвосты» по сравнению с нормальным распределением.

Апробация работы

Основные результаты работы докладывались на Большом семинаре кафедры теории вероятностей в МГУ под руководством член-корр. РАН, профессора А. Н. Ширяева в 2008 году. Неоднократно делались доклады на семинаре «Непараметрическая Статистика и Временные Ряды» под руководством проф. Ю. Н. Тюрина, доц. М. В. Болдина и проф. В. Н. Тутубалина в МГУ в 2007 и 2008 годах. Также были сделаны презентации на нескольких конференциях: «Ломоносовских Чтениях», Москва, 2008, «Колмогоровских Чтениях», Ярославль, 2008, «Международной Конференции по Робастной Статистике» («International Conference on Robust Statistics»), Анталия, Турция, 2008, и на семинаре под руководством профессора Х. Ойа в Университете Тампере, Финляндия, 2008.

Публикации

По теме диссертации опубликовано 6 работ, список которых приведен в конце автореферата [1] - [6].

Структура диссертации

Диссертация состоит из введения, двух глав, списка обозначений и списка литературы, насчитывающего 77 наименований и организованного в алфавитном порядке. Результаты, полученные автором диссертации, оформлены в виде Теорем и Лемм; необходимые известные факты сформулированы в виде Утверждений, с указанием источника. Нумерация утверждений, лемм, теорем и формул начинается в каждой главе заново и состоит из двух чисел. Первое число относится к номеру главы, второе — к номеру утверждения (леммы, теоремы или формулы). Ссылки на работы других авторов сделаны по принципу «автор-дата». Общий объем работы составляет 115 страниц.

2 Краткое содержание диссертации

Диссертация посвящена матричнозначным корреляционным мерам, впервые предлагаемых в литературе в качестве многомерных аналогов числовых коэффициентов корреляций. Применение матричных корреляций продемонстрировано на примере проверки гипотезы о независимости двух многомерных признаков.

Первая глава диссертации состоит из восьми параграфов. В ней предлагается и изучается новое понятие — матричная корреляция.

В Разделе 1.1 приводится ряд полезных алгебраический понятий и фактов, на которые опирается основной материал настоящей работы. Мотивация и определение матричной корреляции обсуждаются в Разделе 1.2. Основанием для введения нового понятия послужило следующее соображение. Коэффициент корреляции двух одномерных случайных величин, по существу, являет собой нормированную ковариацию. Поэтому естественно в качестве основы многомерного обобщения корреляции взять ковариационную матрицу. Способ нормировки ковариационной матрицы дает матричный аналог неравенства Коши-Буняковского (см. Боровков[16], с.159).

Пусть даны две многомерные случайные величины $x \in \mathbb{R}^p$ и $y \in \mathbb{R}^q$. Их ковариационную матрицу обозначим через $\mathrm{Cov}(x, y)$, при этом матрицу $\mathrm{Var}\, x \equiv \mathrm{Cov}(x, x)$ будем называть дисперсионной. Символ $\mathbb{R}^p_q$ обозначает пространство вещественных $p \times q$-матриц; индексы, равные 1, опускаются. Для матричных неравенств, индуцированных положительной определенностью и полуопределенностью, будем использовать символы $\prec$ и $\preccurlyeq$. Относительно случайных векторов x, y мы предполагаем, что их дисперсионные матрицы существуют и невырождены.

[16] А. А. Боровков. *Математическая статистика.* — М.: “Наука”, 1984.

Определение. *Пусть* $x \in \mathbb{R}^p, y \in \mathbb{R}^q$ *имеют конечные* $\operatorname{Var} x, \operatorname{Var} y \succ 0$. *Матричной корреляцией случайных векторов* y *и* x *назовем*

$$\rho = \rho(y, x) = [\operatorname{Var} y]^{-1/2} \operatorname{Cov}(y, x)\ [\operatorname{Var} x]^{-1/2}. \tag{1}$$

Очевидным образом, для $p = q = 1$ формула (1) дает классический коэффициент корреляции. Заметим, что запись (1) корректна, поскольку положительно определенный квадратный корень из положительно определенных матриц $\operatorname{Var} x, \operatorname{Var} y$ существует и притом единственен. Матричный вариант неравенства Коши-Буняковского в принятых обозначениях имеет вид

$$\rho\rho' \preccurlyeq I,$$

где I — единичная матрица, $'$ — знак операции транспонирования.

Определение. *Для всякой матрицы* R *положительно полуопределенную матрицу* $|R| \equiv (RR')^{1/2}$ *назовем матричным модулем. Если* R *полного ранга по строкам, то* $|R|$ *невырожденна, и тогда величину* $\operatorname{sgn} R \equiv |R|^{-1}R$ *будем называть матричным знаком.*

Отметим, что матричный знак квадратной невырожденной матрицы R есть ортогональная матрица, и что разложение $R = |R| \operatorname{sgn} R$ — это полярное разложение матрицы R.

Перечислим элементарные свойства матричной корреляции ρ как меры связи, описанию которых посвящен Раздел 1.3:

(a) Нормировка: $0 \preccurlyeq |\rho| \preccurlyeq I$. Это равносильно тому, что все сингулярные числа матрицы ρ не превосходят 1.

(b) Независимость: Будем говорить, что случайные векторы y и x *некоррелированы*, если $\rho(y, x) = 0$. Таким образом, если y и x независимы, то они некоррелированы. В гауссовском случае понятия независимости и некоррелированности совпадают.

(c) Функциональная зависимость: Условие $|\rho| = I$ эквивалентно тому, что с вероятностью 1 $y = Kx + b$ для некоторой матрицы $K \in \mathbb{R}^q_p$ полного ранга по строкам и вектора $b \in \mathbb{R}^q$. При этом $\rho = \operatorname{sgn} \rho$ состоит из ортонормированных строк, и изменение значения x вдоль направления $[\operatorname{Var} x]^{1/2} e_i$, где $e_i \in \mathbb{R}^p$ — i-ая строчка матрицы ρ, приводит к увеличению i-ой компоненты признака y, $i = \overline{1, q}$.

(d) Свойство возрастания: При возрастании линейной зависимости $|\rho|$ увеличивается (в матричном смысле). Более того, среднеквадратическая погрешность линейной оценки y по x равна

$$\Delta^* = \inf_{K \in \mathbb{R}^q_p, b \in \mathbb{R}^q} \operatorname{Var}[y - Kx - b] = [\operatorname{Var} y]^{1/2}(I - |\rho|^2)[\operatorname{Var} y]^{1/2}.$$

(E) Эквивариантность/инвариантность: Для любых невырожденных матриц $K_1 \in \mathbb{R}^p_p$, $K_2 \in \mathbb{R}^q_q$ и векторов $b_1 \in \mathbb{R}^p$, $b_2 \in \mathbb{R}^q$, выполняется:

$$\widetilde{\rho} = \rho(K_2 y + b_2, K_1 x + b_1) = \operatorname{sgn}\{K_2[\operatorname{Var} y]^{1/2}\} \cdot \rho(y, x) \cdot \operatorname{sgn}'\{K_1[\operatorname{Var} x]^{1/2}\}.$$

В частности, если K_1, K_2 ортогональны, то $\widetilde{\rho} = K_2 \rho(y, x) K_1'$. Сингулярные числа матричной корреляции ρ при аффинных преобразованиях не меняются.
(F) Симметричность: $\rho(y, x) = \rho'(x, y)$.

Таким образом, матричная корреляция ρ является в некотором смысле направленной мерой (линейной) зависимости двух многомерных случайных признаков.

В Разделе 1.4 мы определяем выборочную версию матричной корреляции (1). Пусть даны n независимых реализаций пары p- и q-мерных случайных признаков $(x', y')'$,

$$\begin{pmatrix} \mathbf{x} \\ \mathbf{y} \end{pmatrix} = \begin{pmatrix} x_1 & \dots & x_n \\ y_1 & \dots & y_n \end{pmatrix}. \tag{2}$$

Как и в одномерном случае, для оценки ρ воспользуемся выборочными матрицами ковариаций $\mathbf{s}_{21} = \operatorname{ave}_i(x_i - \bar{x})(y_i - \bar{y})'$ и $\mathbf{s}_{22} = \operatorname{ave}_i(y_i - \bar{y})(y_i - \bar{y})'$, $\mathbf{s}_{11} = \operatorname{ave}_i(x_i - \bar{x})(x_i - \bar{x})'$. Здесь и далее символ ave_i обозначает усреднение по индексу $i = \overline{1, n}$.

Определение. *Для многомерной выборки* (2), *выборочной матричной корреляцией* $\mathbf{y}$ *и* $\mathbf{x}$ *будем называть величину*

$$\mathbf{r} = \mathbf{r}(\mathbf{y}, \mathbf{x}) = \mathbf{s}_{22}^{-1/2} \mathbf{s}_{21}\, \mathbf{s}_{11}^{-1/2}. \tag{3}$$

Опишем геометрический смысл определения. Положим сейчас $p = q$. В недавней работе Тюрин[17] предложил рассматривать $p \times n$-матрицы $\mathbf{x}$ и $\mathbf{y}$ как «векторы» размерности n, координатами которых являются p-столбцы. Обобщенное скалярное произведение так называемых $p \times n$-векторов $\mathbf{x}$ и $\mathbf{y}$ задается формулой:

$$\langle \mathbf{x}, \mathbf{y} \rangle = \sum_{i=1}^{n} x_i y_i'. \tag{4}$$

Ассоциированная с матричным скалярным произведением (4) длина $p \times n$-вектора $\mathbf{x}$ тогда определяется как $|\mathbf{x}| = \langle \mathbf{x}, \mathbf{x} \rangle^{1/2} = (\mathbf{x}\mathbf{x}')^{1/2}$.

Вспомним, что выборочная корреляция Пирсона ($p = q = 1$) представляет собой скалярное произведение нормированных остатков $\mathbf{x} - \bar{x}\mathbf{1}$ и $\mathbf{y} - \bar{y}\mathbf{1}$, где $\mathbf{1} = (1, \dots, 1) \in \mathbb{R}_n$. Отсюда

Определение. *Выборочной матричной корреляцией* $\mathbf{y}$ *и* $\mathbf{x}$ *назовем*

$$\begin{aligned} \mathbf{r}(\mathbf{y}, \mathbf{x}) &= \langle\, |\mathbf{x} - \bar{x}\mathbf{1}|^{-1}(\mathbf{y} - \bar{y}\mathbf{1}),\ |\mathbf{x} - \bar{x}\mathbf{1}|^{-1}(\mathbf{x} - \bar{x}\mathbf{1})\,\rangle \\ &= |\mathbf{y} - \bar{y}\mathbf{1}|^{-1}\langle \mathbf{y} - \bar{y}\mathbf{1},\ \mathbf{x} - \bar{x}\mathbf{1}\rangle |\mathbf{x} - \bar{x}\mathbf{1}|^{-1}. \end{aligned} \tag{5}$$

[17] Ю. Н. Тюрин. “Многомерный анализ: геометрическая теория”. Манускрипт, 2008.

Очевидно, что геометрическое определение (5) совпадает с (3). Более того, из двойственности понятий матрицы ковариаций в пространстве случайных векторов и матричного скалярного произведения (4) в выборочном пространстве вытекает, что свойства (A)-(F) теоретической матричной корреляции ρ также верны и для ее выборочного аналога $\mathbf{r}$.

Естественно, что $\mathbf{r}$ является сильно-состоятельной оценкой ρ, т.е. имеем сходимость матриц: $\mathbf{r}(\mathbf{y},\mathbf{x}) \stackrel{\text{п.н.}}{\to} \rho(y,x)$ при $n \to \infty$.

В Разделе 1.5 доказаны результаты о распределении выборочной матричной корреляции в гауссовском случае. Чтобы их сформулировать, необходимо напомнить некоторые понятия из многомерного анализа[18]. *Векторизацией* $vec\, A$ матрицы A называется вектор, составленный из столбцов A, последовательно записанных один под другим. Случайная $p \times q$-матрица $\mathbf{z}$ имеет *матричное нормальное* распределение $\mathbf{z} \backsim N_q^p(M,\Omega)$ с параметрами $M \in \mathbb{R}_q^p$ и $\Omega \in \mathbb{R}_{pq}^{pq}$, если распределение вектора $vec(\mathbf{z}') \backsim N^{pq}(vec(M'),\Omega)$ многомерное нормальное. Случайная $p \times p$-матрица $\mathbf{w}$ имеет *распределение Уишарта* с m степенями свободы $\mathbf{w} \backsim W_p(m)$, если $\mathbf{w} \stackrel{d}{=} \sum_{i=1}^m z_i z_i'$, где z_i — н. о. р. $N^p(0,I)$. На основе этих матричных распределений мы вводим

Определение. *Случайная $p \times q$-матрица $\mathbf{t}$ имеет матричное распределение Стьюдента с ν степенями свободы, $\nu \ge p$, пишем $\mathbf{t} \sim S_q^p(\nu)$, если*

$$\mathbf{t} \stackrel{d}{=} \sqrt{\nu}\, \mathbf{w}^{-1/2}\, \mathbf{z}, \quad \text{где } \mathbf{z} \perp\!\!\!\perp \mathbf{w},\ \mathbf{z} \sim N_q^p(0, I_{pq}),\ \mathbf{w} \sim W_p(\nu).$$

Матричное распределение Стьюдента ортогонально инвариантно, т.е. для $\mathbf{t} \sim S_q^p(\nu)$ и любых ортогональных матриц $O_1 \in \mathbb{R}_p^p, O_2 \in \mathbb{R}_q^q$ выполняется равенство по распределению $\mathbf{t} \stackrel{d}{=} O_1 \mathbf{t} O_2$.

Очевидно, что $S_1^1(\nu)$ есть обычное (одномерное) распределение Стьюдента. Кроме того, $S_1^p(\nu)$ — это один из встречающихся в литературе многомерных вариантов стьюдентова распределения, определяемый плотностью

$$\frac{\Gamma(\frac{\nu+p}{2})}{(\nu\pi)^{p/2}\Gamma(\frac{\nu}{2})}\Big(1 + \frac{\|x\|^2}{\nu}\Big)^{-(\nu+p)/2} \tag{6}$$

для $x \in \mathbb{R}^p$, где $\|\cdot\|$ — евклидовская норма. Другие многомерные версии можно найти в книге авторов Kotz, Nadarajah[19].

Следующий результат можно использовать для проверки гипотезы о независимости двух гауссовских векторов.

Теорема 1.1. *Пусть $\{(x_i', y_i')'\}_{i=1}^n$ суть n независимых реализаций пары $x \in \mathbb{R}^p, y \in \mathbb{R}^q$, имеющей совместное нормальное распределение с $\rho(y,x) = 0$,*

[18] M. Bilodeau, D. Brenner. *Theory of Multivariate Statistics.* — N.Y.: Springer-Verlag, 1999.

[19] S. Kotz, S. Nadarajah. *Multivariate t Distributions and their Applications.* — Cambridge Univ. Press, 2004.

и $\mathbf{r}$ *— выборочная версия* ρ*. Пусть* $n \geq p+q+1$*. Тогда*

$$|\,\mathbf{r}'(I_q - \mathbf{r}\mathbf{r}')^{-1/2}\,|^2 \stackrel{d}{=} |\mathbf{t}'|^2/(n-1-p),$$

где величина $\mathbf{t} \sim S_p^q(n-1-p)$*.*

Теорема 1.1 обобщает известное свойство пирсоновской корреляции r: для выборки $\{(x_i', y_i')'\}$ из двумерного нормального распределения с $\rho = 0$,

$$\frac{r}{\sqrt{1-r^2}} \stackrel{d}{=} t_{n-2}/\sqrt{n-2},$$

где случайная величина t_{n-2} имеет распределение Стьюдента с $n-2$ степенями свободы.

Далее нам нужны следующие определения. *Коммутационная матрица* $K_{s,t} \in \mathbb{R}^{st}_{st}$ — это матрица блочного вида, (i,j)-ый блок которой размера $t \times s$ с единицей на месте (j,i) и остальными нулями; $i = \overline{1,s}$, $j = \overline{1,t}$. *Кронекеровское произведение* $\otimes$ матриц $A \in \mathbb{R}^p_q$ и $B \in \mathbb{R}^r_s$ задается соотношением $A \otimes B = (a_{ij}B) \in \mathbb{R}^{pr}_{qs}$. *Кронекеровская сумма* $\oplus$ двух квадратных матриц $A \in \mathbb{R}^p_p$ и $B \in \mathbb{R}^q_q$ определяется формулой $A \oplus B = A \otimes I_q + I_p \otimes B$.

Теорема 1.2. *Пусть выборка* $\{(x_i', y_i')'\}_{i=1}^n$ *получена из совместного нормального распределения с матричной корреляцией* $\rho(y,x) = 0$ *и дисперсионными матрицами* $\operatorname{Var} x = I_p, \operatorname{Var} y = I_q$*. Тогда для* $\mathbf{r}$ *— выборочной версии* ρ *при* $n \to \infty$ *выполняется*

$$n^{1/2}(\mathbf{r} - \rho) \xrightarrow{d} N_p^q(0, \Omega), \text{ где}$$

$$4\Omega = 2(I_q - \rho\rho') \otimes (I_p - \rho'\rho) + ((I_q - \rho\rho') \oplus (I_p - \rho'\rho))(I_{qp} - K_{q,p}\rho' \otimes \rho).$$

В случае $p = q = 1$ Теорема 1.2 дает хорошо известный результат о распределении коэффициента корреляции Пирсона: для выборки из двумерного нормального распределения

$$n^{1/2}(r - \rho) \xrightarrow{d} N(0, (1-\rho^2)^2).$$

Итак, наши результаты, наряду с работой Тюрина[20], свидетельствуют о том, что обобщение чисел до квадратных матриц имеет естественное приложение в области многомерного статистического анализа. В дополнение к этому в диссертации, в Разделах 1.6 и 1.7, мы определяем матричное корреляционное отношение и матричную частную корреляцию, и обсуждаем их дальнейшее возможное применение.

Завершающий Раздел 1.8 описывает связь матричной корреляции с другими понятиями многомерного корреляционного и регрессионного анализа.

[20] Ю. Н. Тюрин. "Многомерный анализ: геометрическая теория". Манускрипт, 2008.

Чтобы избежать путаницы в терминологии особо отметим, что известная в литературе «корреляционная матрица» совершенно отлична от нашей матричной корреляции.

ОПРЕДЕЛЕНИЕ. *Для случайного вектора* $z = (z^1, \dots, z^l)'$, *компоненты которого имеют невырожденные дисперсии, корреляционной матрицей называется* $C(z) = (\rho(z^i, z^j)) \in \mathbb{R}_l^l$.

Пусть случайные векторы x, y имеют невырожденные дисперсионные матрицы и дисперсии компонент векторов тоже невырождены. Пусть $z = (x', y')'$ и $\mathrm{Diag}(\mathrm{Var}\, x)$ — диагональная матрица, имеющая на диагонали те же элементы, что и $\mathrm{Var}\, x$. Тогда корреляционная матрица $C(z)$ состоит из ковариационных и дисперсионных матриц векторов $[\mathrm{Diag}(\mathrm{Var}\, x)]^{-1/2}x$, $[\mathrm{Diag}(\mathrm{Var}\, y)]^{-1/2}y$, с коррелированными координатами. А матричная корреляция есть ковариационная матрица векторов $[\mathrm{Var}\, x]^{-1/2}x$, $[\mathrm{Var}\, y]^{-1/2}y$, нормированных так, что координаты каждого из них *не*коррелированы между собой.

В случае $p > q = 1$, матричный модуль $|\rho|$ есть вещественное число. Его называют *коэффициентом множественной корреляции*. Это максимально возможный коэффициент корреляции между $y \in \mathbb{R}$ и линейной комбинацией компонент $x \in \mathbb{R}^p$: $|\rho(y, x)| = \sup\{\rho(y, h'x) : h \in \mathbb{R}^p\}$.

В случае $p \geq q > 1$ большей размерности, $|\rho|$ теряет свойство инвариантности при линейных преобразованиях (см. свойство (E)). Однако, неизменными остаются сингулярные значения $\rho_1, \dots, \rho_q$ матричной корреляции:

$$\rho(y, x) = G(D_\rho, 0)H',$$

где $D_\rho = \mathrm{diag}\{\rho_1, \dots, \rho_q\}$ и $H \in \mathbb{R}_p^p$, $G \in \mathbb{R}_q^q$ — ортогональные матрицы. Числа ρ_i суть *канонические корреляции* случайных величин $x \in \mathbb{R}^p$, $y \in \mathbb{R}^q$, а компоненты случайных векторов $H'[\mathrm{Var}\, x]^{-1/2}x$ и $G'[\mathrm{Var}\, y]^{-1/2}y$ являются *каноническими величинами*. Впервые эти понятия употребил Hotelling[21].

Далее, рассмотрим многомерную линейную регрессионную модель:

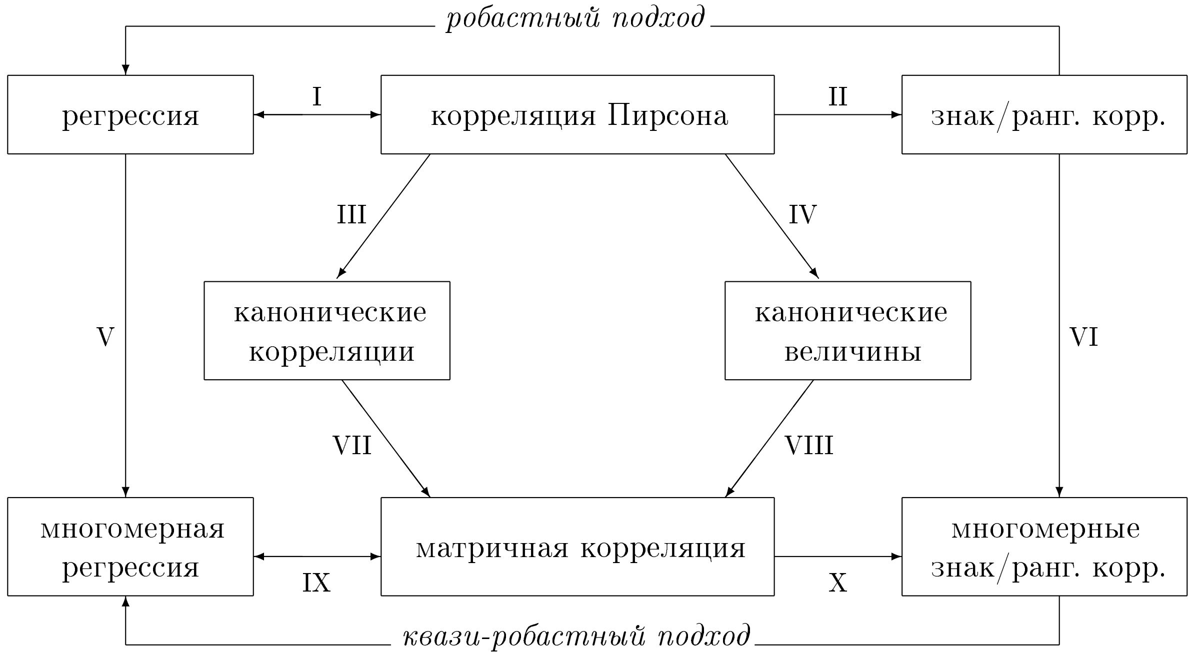
$$\mathbf{y} = K\mathbf{x} + \mathbf{e},$$

где $K \in \mathbb{R}_p^q$ — неизвестный коэффициент регрессии, $\mathbf{x} = (x_1, \dots, x_n) \in \mathbb{R}_n^p$, $\mathbf{y} = (y_1, \dots, y_n) \in \mathbb{R}_n^q$ — экспериментальные данные. Столбцы матрицы ошибок $\mathbf{e} = (\varepsilon_1, \dots, \varepsilon_n) \in \mathbb{R}_n^q$ суть н. о. р. случайные векторы. Как и в одномерной теории, поиск параметра K из условия некоррелированности фактора $\mathbf{x}$ и остатков $\mathbf{y} - K\mathbf{x}$, т.е. как решения уравнения $\mathbf{r}(\mathbf{x},\ \mathbf{y} - K\mathbf{x}) = 0$, приводит к известной оценке наименьших квадратов.

Резюмируем результаты Главы 1. Канонические корреляции, и коэффициент множественной корреляции как частный случай, являются известным

[21] H. Hotelling. “Relation Between Two Sets of Variables”. — *Biometrika*, Vol. 28, pp. 321-377, 1936.

Рис. 1: Связь матричной корреляции с различными понятиями многомерного анализа. Описание схемы см. в тексте, с. 10-11.



обобщением классического коэффициента корреляции [Рис. 1:III]. Существенное отличие такого многомерного аналога от своего «родителя» заключается в утрате информации о направлении связи. Поэтому наряду с каноническими корреляциями, для изучения направлений многомерной зависимости, также рассматривают канонические величины [Рис. 1:IV]. Предложенная матричная корреляция (1), имея сингулярными числами канонические корреляции, позволяет делать выводы как о силе связи [Рис. 1:VII], так и о ее направлениях [Рис. 1:VIII]. Как следствие, матричная корреляция зависима от выбора системы координат. Тесная связь корреляции Пирсона с линейной регрессией [Рис. 1:I] повторяется для многомерного случая [Рис. 1:IX]. «Унаследованным» недостатком матричной корреляции является необходимость совместного распределения данных по нормальному закону, так как в общем случае понятия независимости и некоррелированности не совпадают. Построение на основе корреляции Пирсона непараметрических знаковых и ранговых коэффициентов корреляций [Рис. 1:II] и дальнейшее их использование для поиска робастных решений в задачах линейной регрессии можно провести также в многомерном случае [Рис. 1:X]. Этому вопросу посвящена вторая глава.

Глава 2 диссертационной работы состоит из восьми параграфов. В ней мы определяем матричнозначные версии популярных ранговых коэффициентов корреляций Спирмена, Кендэлла и квадрантной корреляции, и строим на их основе многомерные аффинно инвариантные непараметрические тесты независимости.

В Разделе 2.1 дается краткий обзор одномерного случая. В Разделе 2.2 представлены векторные обобщения рангов и знаков. Пусть $\mathbf{x} = (x_1, \dots, x_n)$ — n независимых реализаций p-мерной случайной величины с произвольным распределением F. Положим $I = \{i_1, \dots, i_p\}$ для упорядоченного набора индексов $1 \le i_1 < \dots < i_p \le n$, и пусть $\mathbf{x}_I = \{x_{i_1}, \dots, x_{i_p}\}$. *Аффинно-эквивариантная медиана Ойа*[22], будем обозначать ее $\widehat{\mu}_{\mathbf{x}} \in \mathbb{R}^p$, минимизирует целевую функцию

$$V(x; \mathbf{x}) := \operatorname*{ave}_I \triangle_p(x, \mathbf{x}_I), \quad x \in \mathbb{R}^p,$$

где усреднение ave_I берется по всевозможным $\binom{n}{p}$ наборам индексов I, и величина

$$\triangle_p(x, \mathbf{x}_I) := \mathrm{abs}\Big\{\det \begin{pmatrix} 1 & 1 & \dots & 1 \\ x & x_{i_1} & \dots & x_{i_p} \end{pmatrix}\Big\},$$

деленная на $p!$, равна абсолютному объему p-симплекса с вершинами $x, \mathbf{x}_I$. Положим $J = \{j_1, \dots, j_{p-1}\}$, $1 \le j_1 < \dots < j_{p-1} \le n$. Обобщение модуля, соответствующее медиане Ойа, задается в виде

$$V_0(x; \mathbf{x}) := \operatorname*{ave}_J \triangle_p(x, \mathbf{x}_J, 0).$$

Определение. *Многомерной знаковой и центрированной ранговой функциями $x \in \mathbb{R}^p$ называются, соответственно, градиенты:*

$$S_0(x; \mathbf{x}) := \nabla_x V_0(x; \mathbf{x}), \qquad R\,(x; \mathbf{x}) := \nabla_x V(x; \mathbf{x}). \tag{7}$$

Считаем, что вектор частных производных ∇_x является столбцом. Эмпирическими знаками (относительно медианы Ойа) и рангами наблюдений называются векторные величины $\widehat{S}_{i\mathbf{x}} \equiv S_0(x_i - \widehat{\mu}_{\mathbf{x}}; \mathbf{x} - \widehat{\mu}_{\mathbf{x}}\mathbf{1})$, $R_{i\mathbf{x}} \equiv R(x_i; \mathbf{x})$.

Для эмпирических знаков и рангов $\mathrm{ave}_i\, \widehat{S}_{i\mathbf{x}} = \mathrm{ave}_i\, R_{i\mathbf{x}} = 0$, и выполняется свойство аффинной эквивариантности *в том смысле*, что для знаков $\widehat{S}^*_{i\mathbf{x}}$ и рангов $R^*_{i\mathbf{x}}$, построенных по преобразованным наблюдениям $\{x^*_i = Ax_i + b\}$ с невырожденной матрицей $A \in \mathbb{R}^p_p$ и $b \in \mathbb{R}^p$,

$$\widehat{S}^*_{i\mathbf{x}} = |\det A|(A^{-1})'\,\widehat{S}_{i\mathbf{x}}, \qquad R^*_{i\mathbf{x}} = |\det A|(A^{-1})'\,R_{i\mathbf{x}}. \tag{8}$$

Теоретические аналоги выглядят следующим образом. *Медиана Ойа* $\mu = \mu(F)$ минимизирует ожидаемую функцию объема $V(x; F) := \mathrm{E}_F \triangle_p(x, \mathbf{x}_I)$, где математическое ожидание берется по независимым p наблюдениям с распределением F, перечисленным в $\mathbf{x}_I = \{x_{i_1}, \dots, x_{i_p}\}$.

Определение. *Теоретическими центрированными знаковой и ранговой функциями $x \in \mathbb{R}^p$ называются, соответственно,*

$$S_F(x; \mu) := \mathrm{E}_F \nabla_x [\triangle_p(x, \mathbf{x}_J, \mu)], \qquad R_F(x) := \mathrm{E}_F \nabla_x [\triangle_p(x, \mathbf{x}_I)], \tag{9}$$

[22] H. Oja. “Descriptive Statistics for Multivariate Distributions”. — *Stat. Prob. Lett.*, Vol. 1, pp. 327-332, 1983.

где $\mu = \mu(F)$ — медиана Ойа, и математическое ожидание берется по независимым величинам, перечисленным в $\mathbf{x}_J$ и $\mathbf{x}_I$, соответственно.

Отметим, что для F с конечными первыми моментами функции (9) существуют и равномерно ограничены. См., например, обзорную работу Oja[23].

В Разделе 2.3 мы определяем знаковые и ранговые матричные корреляции, применяя понятие матричной корреляции к аффинно-эквивариантным знакам наблюдений, рангам и знакам попарных разностей. Обратимся к многомерной выборке (2). Положим $x_{ij} := x_i - x_j$, $y_{ij} := y_i - y_j$, и совокупности таких разностей обозначим ${}^0\mathbf{x} := \{x_{ij}\}$, ${}^0\mathbf{y} := \{y_{ij}\}$, $1 \le i < j \le n$. Действуя с элементами введенных множеств как с обычными наблюдениями, можно вычислить эмпирические знаки

$$S_{ij\mathbf{x}} \equiv S_0(x_{ij}; {}^0\mathbf{x}) \qquad \text{и} \qquad S_{ij\mathbf{y}} \equiv S_0(y_{ij}; {}^0\mathbf{y}), \tag{10}$$

используя знаковую функцию S_0 из (7). Естественно, построенные таким образом знаки также аффинно-эквивариантны (в смысле (8)).

Определение. *Пусть $R_{i\mathbf{x}}$, $\widehat{S}_{i\mathbf{x}}$ и $S_{ij\mathbf{x}}$ обозначают аффинно-эквивариантные ранговые и знаковые векторы из* (7) *и* (10), *и аналогично для* $\mathbf{y}$. *Матричными версиями коэффициентов корреляций Спирмена, Кендэлла и квадрантной корреляции будем называть, соответственно,*

$$\begin{aligned}
\mathbf{r}_S &= (\operatorname*{ave}_i R_{i\mathbf{x}} R'_{i\mathbf{x}})^{-1/2} \operatorname*{ave}_i R_{i\mathbf{x}} R'_{i\mathbf{y}} \ (\operatorname*{ave}_i R_{i\mathbf{y}} R'_{i\mathbf{y}})^{-1/2}, \\
\mathbf{r}_K &= (\operatorname*{ave}_{i,j,k} S_{ij\mathbf{x}} S'_{ik\mathbf{x}})^{-1/2} \operatorname*{ave}_{i<j} S_{ij\mathbf{x}} S'_{ij\mathbf{y}} \ (\operatorname*{ave}_{i,j,k} S_{ij\mathbf{y}} S'_{ik\mathbf{y}})^{-1/2}, \\
\mathbf{r}_Q &= (\operatorname*{ave}_i \widehat{S}_{i\mathbf{x}} \widehat{S}'_{i\mathbf{x}})^{-1/2} \operatorname*{ave}_i \widehat{S}_{i\mathbf{x}} \widehat{S}'_{i\mathbf{y}} \ (\operatorname*{ave}_i \widehat{S}_{i\mathbf{y}} \widehat{S}'_{i\mathbf{y}})^{-1/2}.
\end{aligned} \tag{11}$$

Лемма 2.1. *Сингулярные числа матриц $\mathbf{r}_S$, $\mathbf{r}_K$, $\mathbf{r}_Q$ инварианты относительно группы аффинных преобразований $\{x_i \to A_1 x_i + b_1\}$, $\{y_i \to A_2 y_i + b_2\}$ с невырожденными матрицами A_1, A_2 и векторами b_1, b_2.*

Перейдем к определению теоретических аналогов матричных корреляций (11). Мы будем предполагать, что распределения изучаемых признаков симметричны в смысле следующего

Определение. *Распределение $x \sim F$ из $\mathbb{R}^p$ называется симметричным, если существует вектор $\theta \in \mathbb{R}^p$, называемый центром симметрии, для которого выполняется равенство по распределению $x - \theta \stackrel{d}{=} -(x - \theta)$.*

Для симметричного F с центром θ из аффинной эквивариантности медианы Ойа немедленно получаем, что $\mu(F) = \theta$. Следовательно, в этом случае центрированная знаковая функция (9) для F определена однозначно, и мы будем кратко писать $S_F(x) \equiv S_F(x; \mu(F))$.

[23] H. Oja. “Affine Invariant Multivariate Sign and Rank Tests and Corresponding Estimates: a Review”. — *Scand. J. Statist.* (invited paper), Vol. 26, pp. 319-343, 1999.

Обозначим симметричное (с нулевым центром) распределение разности $x_{12} \equiv x_1 - x_2$ двух независимых случайных величин с одним и тем же распределением F через ${}^0F$ — оно называется *симметризацией* F. Для него также можно построить теоретическую знаковую функцию (9). По определению,

$$S_{{}^0F}(x) = \mathrm{E}\nabla_x[\triangle_p(x, x_{12}, x_{34}, \dots, x_{2p-3,2p-2}, 0)], \tag{12}$$

где математическое ожидание берется по независимым величинам x_{ij} с распределением ${}^0F$.

Определение. *Пусть распределение $(x', y')' \sim H$ с маргинальными $x \sim F$, $y \sim G$ имеет конечные первые моменты. Пусть ранговые и знаковые функции $R_F, S_F, S_{{}^0F}$ задаются формулами* (9), (12), *и аналогично для G. Теоретической матричной корреляцией Спирмена будем называть*

$$\rho_S(H) = \left[\mathrm{E}_F R_F(x) R'_F(x)\right]^{-1/2} \mathrm{E}_H R_F(x) R'_G(y) \ \left[\mathrm{E}_G R_G(y) R'_G(y)\right]^{-1/2},$$

Кендэлла —

$$\rho_K(H) = \left[\mathrm{E}_F S_{{}^0F}(x_{12}) S'_{{}^0F}(x_{13})\right]^{-1/2} \mathrm{E}_H S_{{}^0F}(x_{12}) S'_{{}^0G}(y_{12}) \left[\mathrm{E}_G S_{{}^0G}(y_{12}) S'_{{}^0G}(y_{13})\right]^{-1/2}$$

и квадрантной корреляцией (для симметричных F, G) —

$$\rho_Q(H) = \left[\mathrm{E}_F S_F(x) S'_F(x)\right]^{-1/2} \mathrm{E}_H S_F(x) S'_G(y) \ \left[\mathrm{E}_G S_G(y) S'_G(y)\right]^{-1/2}.$$

Предполагается, что нормирующие знаковые и ранговые дисперсионные матрицы невырождены.

Данное определение корректно, так как для существования используемых знаковых/ранговых ковариационных матриц достаточно конечных первых моментов, и из невырожденности нормирующих матриц следует, что из них можно извлечь (притом единственным образом) корень степени $-1/2$. Симметричность распределения влечет единственность теоретической медианы Ойа и, значит, все используемые знаковые функции определены однозначно.

Заканчивается Раздел 2.3 доказательством асимптотической нормальности выборочных матричных корреляций (11). Матричная статистика $\mathbf{r}$ называется $\sqrt{n}$-асимптотически гауссовской оценкой для ρ, если при $n \to \infty$ величина $\sqrt{n}(\mathbf{r} - \rho)$ имеет предельное матричное нормальное распределение с нулевым средним.

Теорема 2.1. *Пусть распределение $(x', y')' \sim H$ имеет конечные вторые моменты и $\rho_S(H)$ определена. Тогда выборочная матричная корреляция $\mathbf{r}_S$, построенная по выборке объема n из распределения H, для ρ_S является $\sqrt{n}$-асимптотически гауссовской оценкой.*

Для изучения асимптотических свойств $\mathbf{r}_Q = \mathbf{r}_Q(\widehat{\mu}_{\mathbf{x}}, \widehat{\mu}_{\mathbf{y}})$ необходимо сначала обосновать возможность замены выборочных медиан Ойа их неизвестными теоретическими значениями. Запись $\mathbf{r}_Q(\mu_1, \mu_2)$ означает, что используемые для построения квадрантной матричной корреляции эмпирические знаки наблюденных x_i, y_i вычисляются относительно μ_1, μ_2 соответственно.

Лемма 2.2. *Пусть*

(Q-1) распределение $(x', y')' \sim H$ имеет конечные вторые моменты;

(Q-2) функции распределений x, y непрерывны;

(Q-3) выборочные медианы Ойа $\widehat{\mu}_{\mathbf{x}}$, $\widehat{\mu}_{\mathbf{y}}$, построенные по выборкам объемов n из распределений $x \sim F, y \sim G$, являются $\sqrt{n}$-состоятельными оценками нулевых центров распределений;

(Q-4) распределение H симметрично относительно нуля;

(Q-5) дифференцируемы в $\mu_1 = \mu_2 = 0$ компоненты матричных функций

$$\mathrm{E}_H S_F(x;\mu_1) S'_G(y;\mu_2), \quad \mathrm{E}_F S_F(x;\mu_1) S'_F(x;\mu_1), \quad \mathrm{E}_G S_G(y;\mu_2) S'_G(y;\mu_2).$$

Тогда при $n \to \infty$

$$\sqrt{n}\,[\mathbf{r}_Q(\widehat{\mu}_{\mathbf{x}}, \widehat{\mu}_{\mathbf{y}}) - \mathbf{r}_Q(0,0)] \xrightarrow{p} 0.$$

Достаточные условия для выполнения *(Q-3)* приведены Arcones и др[24].

Теорема 2.2. *Пусть выполнены условия Леммы 2.2 и $\rho_Q(H)$ определена. Тогда выборочная корреляция $\mathbf{r}_Q$, построенная по выборке объема n из распределения H, для ρ_Q является $\sqrt{n}$-асимптотически гауссовской оценкой.*

Теорема 2.3. *Пусть распределение $(x', y')' \sim H$ имеет конечные вторые моменты и $\rho_K(H)$ определена. Тогда выборочная матричная корреляция $\mathbf{r}_S$, построенная по выборке объема n из распределения H, для ρ_K является $\sqrt{n}$-асимптотически гауссовской оценкой.*

Раздел 2.4 посвящен многомерным тестам независимости. Мы приводим классические тестовые статистики и записываем их в виде функций выборочной матричной корреляции (3). Аналогичным образом строим новые критерии независимости с помощью ранговых и знаковых выборочных матричных корреляций (11). Пусть на основе n независимых реализаций $\{(x'_i, y'_i)'\}_{i=1}^n$ пары p- и q-мерных признаков $(x', y')'$ мы хотим проверить гипотезу

$$\mathcal{H}_0: \ x \text{ и } y \text{ независимы.} \tag{13}$$

Новыми тестовыми статистиками для проверки гипотезы $\mathcal{H}_0$ выбраны

$$r_S = \|\mathbf{r}_S\|, \qquad r_K = \|\mathbf{r}_K\|, \qquad r_Q = \|\mathbf{r}_Q\|, \tag{14}$$

где $\|\mathbf{r}\| \equiv \mathrm{tr}^{1/2}[\mathbf{r}'\mathbf{r}]$ — матричная норма Фробениуса. Лемма 2.1 немедленно влечет аффинную инвариантность предложенных статистик. Также отметим,

[24] M. Arcones, Z. Chen and E. Giné. "Estimators Related to U-process with Applications to Multivariate Medians: Asymptotic Normality". *Ann. Statist.*, Vol. 22, pp. 1460-1477, 1994.

что для применения наших тестов требуется существование *вторых* моментов распределений признаков (в то время как для классических — четвертые). Основным результатом Раздела 2.4 является установление распределений тестовых статистик (14) в условиях гипотезы независимости $\mathcal{H}_0$. Пусть χ^2_{pq} обозначает хи-квадрат распределение с pq степенями свободы.

Теорема 2.4. *Пусть распределения признаков имеют конечные вторые моменты и симметричны. Тогда при $\mathcal{H}_0$, $nr_S^2 \xrightarrow{d} \chi^2_{pq}$ для $n \to \infty$.*

Теорема 2.5. *Пусть распределения признаков удовлетворяют условиям (Q-1)-(Q-3) Леммы 2.2. Тогда при $\mathcal{H}_0$, $nr_Q^2 \xrightarrow{d} \chi^2_{pq}$ для $n \to \infty$.*

Теорема 2.6. *Пусть распределения признаков имеют конечные вторые моменты. Тогда при $\mathcal{H}_0$, $nr_K^2/4 \xrightarrow{d} \chi^2_{pq}$ для $n \to \infty$.*

Для конечных выборок многомерные версии знаковых и ранговых тестов (как предложенные нами, так и другие возможные обобщения), к сожалению, теряют важное свойство свободы от распределения. Это объясняется тем, что для многомерных данных не существует естественного упорядочения.

Остальные разделы второй главы посвящены изучению статистических свойств (эффективности и робастности) предложенных статистик (14) для эллиптической модели распределений признаков. Раздел 2.5 дает определение эллиптического распределения и вид векторных знаковых и ранговых функций в этом частном случае.

Определение. *Говорят, что случайный вектор x из $\mathbb{R}^p$ имеет эллиптическое распределение с параметрами $\mu \in \mathbb{R}^p$ и $\Lambda \in \mathbb{R}^p_p$, $\Lambda \succ 0$, если его плотность имеет вид*

$$f_x(x) = |\det \Lambda|^{-1/2} f_0(\|\Lambda^{-1/2}(x-\mu)\|^2), \tag{15}$$

где $\|\cdot\|$ — евклидовская норма, f_0 — заданная неотрицательная вещественная функция, не зависящая от μ, Λ, и $\int_0^\infty r^{p-1} f_0(r^2)dr = \Gamma(\frac{1}{2}p)/(2\pi^{p/2})$. Будем обозначать такие распределения $x \sim E_p(\mu, \Lambda)$. Если $x \sim E_p(\mu, \alpha I_p)$, $\alpha > 0$, тогда распределение x называют сферическим.

Говорят, что случайный вектор x из $\mathbb{R}^p$ имеет *эллиптическое p-мерное t-распределение Стьюдента* с ν степенями свободы и пишут $x \sim t_{\nu,p}(\mu, \Lambda)$, если его плотность имеет вид (15) с функцией $f_0(\|x\|^2)$, заданной в (6). Данное распределение представляет интерес для изучения различных свойств статистик, так как с помощью параметра ν можно варьировать «тяжесть хвостов» (при $\nu \to \infty$ получаем многомерный нормальный закон).

Если $x \sim E_p(\mu, \Lambda)$, то вектор $\Lambda^{-1/2}(x-\mu) \sim E_p(0, I_p)$ имеет сферическое распределение. Таким образом, эллиптическая модель является сдвигово-масштабным преобразованием сферических распределений.

Пусть наблюдения $x_i \in \mathbb{R}^p$ имеют произвольное сферическое распределение F_0 с нулевым средним (т.е., ортогонально инвариантное) и конечными первыми моментами. Тогда функции (9) имеют вид

$$S_{F_0}(x) = c_{F_0}\,\{\|x\|^{-1}x\}, \qquad R_{F_0}(x) = \alpha_{F_0}(\|x\|)\,\{\|x\|^{-1}x\}, \tag{16}$$

где c_{F_0} — некоторая постоянная, и $\alpha_{F_0}(\cdot)$ — ограниченная вещественная функция. См. обзор Oja[25] и ссылки в нем. Также нам понадобится представление функции (12):

$$\mathrm{E}_{F_0}[S\text{o}_{F_0}(x-x_i)\mid x_i] = c\text{o}_{F_0}\,\mathrm{E}_{F_0}[\frac{x-x_i}{\|x-x_i\|}\mid x_i] = c\text{o}_{F_0}\,\beta_{F_0}(\|x\|)\frac{x}{\|x\|}, \tag{17}$$

где вещественная функция $\beta_{F_0}(\cdot)$ ограничена. См. Möttönen и др.[26]

В Разделе 2.6 новые ранговые и знаковые критерии сравниваются с другими известными в литературе с помощью асимптотической относительной эффективности (АОЭ) Питмена. Для этого вводится многомерная модель зависимости в качестве альтернативы к нулевой гипотезе независимости, впервые предложенная Gieser, Randles[27]:

$$\mathcal{H}_n: \quad \begin{pmatrix} x \\ y \end{pmatrix} = \begin{pmatrix} (1-\Delta)I_p & \Delta M_1 \\ \Delta M_2 & (1-\Delta)I_q \end{pmatrix} \begin{pmatrix} x^* \\ y^* \end{pmatrix}, \tag{18}$$

где вещественные параметр $\Delta = n^{-1/2}\delta$, $\delta \geq 0$; $x^* \in \mathbb{R}^p$, $y^* \in \mathbb{R}^q$ — независимые случайные векторы; $M_1, M_2' \in \mathbb{R}^p_q$ — фиксированные матрицы.

Мы предполагаем что x^*, y^* из (18) имеют эллиптические распределения с конечными вторыми моментами. Для нахождения распределения критериальных статистик (14) при близких альтернативах применена третья Лемма Ле Кама (см., например, Гаек, Шидак[28], van der Vaart[29]). При этом требуется установить локальную асимптотическую нормальность модели (18) и, как следствие, контигуальность последовательности альтернатив $\mathcal{H}_n$ гипотезе $\mathcal{H}_0$.

Лемма 2.3. *Пусть независимые x^*, y^* имеют сферические плотности $f_{x^*}(x^*) = f_0(\|x^*\|^2)$, $f_{y^*}(y^*) = g_0(\|y^*\|^2)$ такие, что f_0, g_0 — положительны, непрерывно дифференцируемы, и*

$$\mathrm{E}_{\mathcal{H}_0}\|x^*\|^2 < \infty, \quad \mathrm{E}_{\mathcal{H}_0}\|x^*\|^2\phi^2(\|x^*\|^2) < \infty, \quad \mathrm{E}_{\mathcal{H}_0}\|x^*\|^4\phi^2(\|x^*\|^2) < \infty,$$

где $\phi := f_0'/f_0$; аналогично для y^*. Тогда последовательность альтернатив $\mathcal{H}_n$ (18) контигуальна нулевой гипотезе $\mathcal{H}_0 : \Delta = 0$.

---

[25] H. Oja. “Affine Invariant Multivariate Sign and Rank Tests and Corresponding Estimates: a Review”. — *Scand. J. Statist.* (invited paper), Vol. 26, pp. 319-343, 1999.

[26] J. Möttönen, H. Oja, J. Tienari. “On the Efficiency of Multivariate Spatial Sign and Rank Tests”. *Ann. Statist.*, Vol. 25, pp. 542-552, 1997.

[27] P. W. Gieser, R. H. Randles. “A Nonparametric Test of Independence Between Two Vectors”. — *J. Amer. Statist. Assoc.*, Vol. 92, pp. 561-567, 1997.

[28] Я. Гаек, З. Шидак. *Теория ранговых критериев.* — М.: “Наука”, 1971.

[29] A.W. van der Vaart. *Asymptotic Statistics.* — Cambridge Univ. Press, 1998.

Установлено, что тестовые статистики (14) при альтернативах $\mathcal{H}_n$ (18) имеют предельные нецентральные хи-квадрат распределения с параметрами нецентральности общего вида:

$$\chi^2_{pq}(\delta^2(pq)^{-1}\|m_1M_1 + m_2M_2'\|^2), \tag{19}$$

где постоянные m_1, m_2 зависят только от распределений x^*, y^* из (18). Пусть $\mathcal{I}\{\cdot\}$ — индикатор события.

Теорема 2.7. *Пусть $x^* \sim F_0, y^* \sim G_0$ удовлетворяют условиям Леммы 2.3. Тогда при $\mathcal{H}_n$, nr_S^2 сходится по распределению при $n \to \infty$ к (19), где*

$$m_1 = (a_{F_0}a_{G_0})^{-1}[\mathrm{E}\alpha'_{F_0}(\|x^*\|) + (p-1)\mathrm{E}\alpha_{F_0}(\|x^*\|)\|x^*\|^{-1}]\mathrm{E}\alpha_{G_0}(\|y^*\|)\|y^*\|,$$
$$m_2 = (a_{F_0}a_{G_0})^{-1}[\mathrm{E}\alpha'_{G_0}(\|y^*\|) + (q-1)\mathrm{E}\alpha_{G_0}(\|y^*\|)\|y^*\|^{-1}]\mathrm{E}\alpha_{F_0}(\|x^*\|)\|x^*\|.$$

Здесь постоянная величина $a^2_{F_0} = \mathrm{E}_{F_0}\alpha^2_{F_0}(\|x^*\|)$ для функции α_{F_0} из (16), и α'_{F_0} обозначает производную (существование которой предполагается). Аналогичные обозначения использованы для G_0.

Теорема 2.8. *Пусть $x^* \sim F_0, y^* \sim G_0$ удовлетворяют условиям Леммы 2.3 и Теоремы 2.5. Тогда при $\mathcal{H}_n$, nr_Q^2 сходится по распределению при $n \to \infty$ к (19), где постоянные*

$$m_1 = [2f_0(0)\mathcal{I}\{p=1\} + (p-1)\,\mathrm{E}\|x^*\|^{-1}]\,\mathrm{E}\|y^*\|,$$
$$m_2 = [2g_0(0)\mathcal{I}\{q=1\} + (q-1)\,\mathrm{E}\|y^*\|^{-1}]\,\mathrm{E}\|x^*\|.$$

Теорема 2.9. *Пусть $x^* \sim F_0, y^* \sim G_0$ удовлетворяют условиям Леммы 2.3. Тогда при $\mathcal{H}_n$, $nr_K^2/4$ сходится по распределению при $n \to \infty$ к (19), где*

$$m_1 = (2b_{F_0}b_{G_0})^{-1}[2\mathrm{E}f_0(\|x^*\|^2)\mathcal{I}\{p=1\} + (p-1)\mathrm{E}\|x^*_{12}\|^{-1}]\mathrm{E}\|y^*_{12}\|,$$
$$m_2 = (2b_{F_0}b_{G_0})^{-1}[2\mathrm{E}g_0(\|y^*\|^2)\mathcal{I}\{q=1\} + (q-1)\mathrm{E}\|y^*_{12}\|^{-1}]\mathrm{E}\|x^*_{12}\|.$$

Здесь постоянная величина $b^2_{F_0} = \mathrm{E}_{F_0}\beta^2_{F_0}(\|x^*\|)$ с функцией β_{F_0} из (17). Аналогичные обозначения использованы для G_0.

Следствие 2.1. *Асимптотическая эффективность Питмена многомерных критериев Спирмена, Кендэлла и квадрантного теста nr_S^2, $nr_K^2/4$, nr_Q^2 относительно критерия отношения правдоподобий для альтернатив $\mathcal{H}_n$* (18) *в случае $M_1 = M_2'$ равна*

$$(4pq)^{-1}(m_1 + m_2)^2, \tag{20}$$

где постоянные m_1, m_2 даются в Теоремах 2.7-2.9.

Численные результаты для АОЭ (20) в случаях многомерного нормального и t распределений x^*, y^* были получены Möttönen в совместной работе [6]. Эти выкладки приведены в Разделе 2.8. Для многомерных нормальных распределений $F_0, G_0\,(\nu = \infty)$, асимптотическая эффективность предложенных критериев хуже по сравнению с критерием отношения правдоподобия, и улучшается с ростом размерностей (стремится к 1 при $p, q \to \infty$). Для распределений с более «тяжелыми хвостами» (малые ν), по эффективности r_K, r_S превосходят классический тест, и r_Q его превосходит для больших размерностей p, q. Среди предложенных непараметрических знаковых и ранговых критериев (14), асимптотически лучше в рассмотренных случаях многомерная версия критерия Кендэлла.

В Разделе 2.7 изучен вопрос устойчивости к засорениям матричных корреляций (11) (и основанных на них статистик (14)). В литературе сложилась традиция описывать воздействие засорения на оценку посредством ее функции влияния (см. Хьюбер[30], Хампель и др.[31]).

Определение. *Пусть $R(\,\cdot)$ — некоторый функционал, определенный на множестве функций распределений из $\mathbb{R}^l$. Функцией влияния (засорения $z \in \mathbb{R}^l$ на функционал R в точке H) называют предел, если этот предел существует,*

$$IF(z; R, H) = \lim_{\varepsilon \downarrow 0} \frac{R(H_\varepsilon) - R(H)}{\varepsilon}, \tag{21}$$

где $H_\varepsilon = (1-\varepsilon)H + \varepsilon\Delta_z$ и Δ_z — это мера Дирака, сосредоточенная в точке z.

Робастные оценки должны иметь ограниченную функцию влияния. Вольно говоря, это влечет, что любые единичные засорения не имеют произвольно большого влияния на значение оценки.

Теорема 2.10. *Функции влияния засорения $(x', y')'$ на ранговые/знаковые матричные корреляции ρ_S, ρ_K, ρ_Q в точке H_0 такой, что ее маргинальные распределения F_0, G_0 — стандартные сферические, имеют линейный порядок роста по $\|x\|, \|y\|$ (для больших значений x, y) и в одномерном случае $p = q = 1$ ограничены. Предполагается, что H_0 удовлетворяет условиям Теорем 2.1-2.3 и везде, где это необходимо, порядок дифференцирования и взятия математического ожидания можно менять.*

В ходе доказательства этой теоремы получены функции влияния одномерных коэффициентов корреляций Спирмена, Кендэлла и квадратной корреляции. Несмотря на то, что их вывести не сложно, эти результаты, по-видимому, отсутствуют в изданиях по робастной статистике и опубликованы совсем недавно в работе Croux, Dehon[32], и только для квадрантной корреляции — у Пасмана, Шевлякова[33].

---

[30] П. Хьюбер. *Робастность в статистике.* — М.: "Мир", 1984.

[31] Ф. Хампель, Э. Рончетти, П. Рауссеу, В. Штаэль. *Робастность в статистике: подход на основе функции влияния.* — М.: "Мир", 1989.

[32] C. Croux, C. Dehon. "Robustness versus Efficiency for Nonparametric Correlation Measures". — *ECORE discussion paper*, 2008

[33] В. Р. Пасман, Г. Л. Шевляков. "Робастные методы оценивания коэффициента корреляции". — *Автоматика и Телемеханика*, т. 27, сс. 70-80, 1987.

Итак, предложенные нами статистики более устойчивы к засорениям, чем классические тесты, но, тем не менее, не робастны, поскольку их функции влияния не ограничены (за исключением одномерного случая $p = q = 1$). В этом смысле, многомерные знаковые и ранговые корреляции напоминают весьма эффективные и более устойчивые по сравнению с МНК оценки наименьших модулей.

3 Благодарности

Автор выражает глубокую благодарность доктору физико-математических наук, профессору Юрию Николаевичу Тюрину, под руководством которого проходила работа над диссертацией, за постановку задачи и постоянное внимание. Автор благодарит Ханну Ойа за многочисленные обсуждения и Юрки Моттонена за помощь в получении численных результатов.

4 Список публикаций автора по теме диссертации

1. Е. М. Суханова. “Многомерные знаковые и ранговые тесты независимости”. — *Успехи Математических Наук*, т. 63, вып. 5, сс. 199-200, 2008.

2. E. M. Sukhanova. “A Test for Independence of Two Multivariate Samples”. — *Mathematical Methods of Statistics*, Vol. 17, No. 1., pp. 74-86, 2008.

3. Е. М. Суханова. “Медиана Ойа: свойство согласованности с центром симметрии”. — *Сб. Статистические методы оценивания и проверки гипотез*, Пермь: Пермский университет, сс. 62-68, 2008.

4. Е. М. Суханова. “Матричная корреляция”. — *Труды VI Колмогоровских Чтений*, сс. 176-181, 2008.

5. E. M. Sukhanova. “Matrix Correlation”. — *Abstracts of the International Conference on Robust Statistics*, p. 96, 2008.

6. E. M. Sukhanova, J. Möttönen, H. Oja. “Multivariate Test of Independence Based on Matrix Rank Correlation”. — *Abstracts of the International Conference on Robust Statistics*, p. 70, 2008.

   (Сухановой Е. М. принадлежат теоретические результаты, Моттонен Ю., Ойа Х. получили численные результаты для асимптотических эффективностей критерия в некоторых случаях).