УДК 303.732.4 : 519.2

08.00.13 Математические и инструментальные методы экономики (экономические науки)

ОЦЕНИВАНИЕ РАЗМЕРНОСТИ ВЕРОЯТНОСТНО-СТАТИСТИЧЕСКОЙ МОДЕЛИ

Орлов Александр Иванович
д.э.н., д.т.н., к.ф.-м.н., профессор
РИНЦ SPIN-код: 4342-4994
Московский государственный технический университет им. Н.Э. Баумана, Россия, 105005, Москва, 2-я Бауманская ул., 5, prof-orlov@mail.ru

Вероятностно-статистические модели данных - основа методов прикладной статистики. При анализе статистически данных часто необходимо оценивать две составляющие вероятностно-статистических моделей - структуру моделей и их параметры. Методы расчета состоятельных оценок параметров хорошо известны (например, применяют методы одношаговых оценок, которые пришли на смену методам максимального правдоподобия). Структура модели обычно выбирается исследователем (можно сказать, что используются экспертные методы). Некоторые параметры структуры можно оценивать с помощью математико-статистических методов. Например, степень многочлена в регрессионной зависимости или число слагаемых в модели смеси распределений, используемой для классификации. Для подобных параметров модели используется общий термин - размерность вероятностно-статистической модели. Более общая составляющая модели - информативное подмножество признаков. В настоящей статье рассмотрено асимптотическое поведение оценок размерностей ряда моделей. Изучено асимптотическое поведение ряда оценок степени полинома при восстановлении зависимости. Получены состоятельные оценки размерности и структуры модели в регрессии. Рассмотрены подходы к оцениванию числа элементов смеси в задачах классификации. Обсуждаются оценки размерности модели в факторном анализе и многомерном шкалировании. С целью обоснования последовательного выполнения этапов статистического анализа данных анализируются проблемы "стыковки" алгоритмов классификации и регрессии. Полезными оказываются оптимизационные формулировки ряда задач прикладной статистики. Основные результаты касаются состоятельности оценок. Краткие формулировки ряда теорем содержатся в ранее вышедших публикациях. Проблема оценивания размерности вероятностно-статистической модели как самостоятельное направление прикладной статистики впервые

UDC 303.732.4 : 519.2

Mathematical and instrumental methods of Economics

ESTIMATION OF THE DIMENSION OF THE PROBABILITY-STATISTICAL MODEL

Orlov Alexander Ivanovich
Dr.Sci.Econ., Dr.Sci.Tech., Cand.Phys-Math.Sci., professor
Bauman Moscow State Technical University, Moscow, Russia

Probabilistic-statistical data models are the basis of applied statistics methods. When analyzing statistical data, it is often necessary to estimate two components of probabilistic-statistical models - the structure of the models and their parameters. Methods for calculating consistent parameter estimates are well known (for example, the one-step estimation methods are used, which replaced the maximum likelihood methods). The structure of the model is usually chosen by the researcher (we can say that expert methods are used). Some structural parameters can be estimated using mathematical-statistical methods. For example, the degree of a polynomial in the regression relationship or the number of terms in a mixture model used for classification. For such parameters of the model, a general term is used - the dimension of the probabilistic-statistical model. A more general component of the model is an informative sign subset. In this article, we consider the asymptotic behavior of the dimension estimates for a number of models. The asymptotic behavior of a number of estimates for the degree of a polynomial is studied when restoring the dependence. Consistent estimates of the dimension and structure of the model in regression are obtained. Approaches to estimating the number of elements in a mixture in classification problems are considered. Estimates of the model dimension in factor analysis and multivariate scaling are discussed. In order to substantiate the sequential execution of the stages of statistical data analysis, the problems of "docking" of classification and regression algorithms are analyzed. Optimization formulations of a number of problems in applied statistics turn out to be useful. The main results relate to the consistency of the estimates. Brief formulations of a number of theorems are contained in earlier publications. The problem of estimating the dimension of a probabilistic-statistical model as an independent direction of applied statistics was first considered here. For the first time, the proofs of the theorems included in this article are published. These theorems and detailed proofs are the main scientific results of article

рассмотрена здесь. Впервые публикуются доказательства включенных в настоящую статью теорем. Эти теоремы и подробные доказательства и являются основными научными результатами работы

Ключевые слова: ПРИКЛАДНАЯ СТАТИСТИКА, АНАЛИЗ ДАННЫХ, ОЦЕНИВАНИЕ, СОСТОЯТЕЛЬНОСТЬ, ОПТИМИЗАЦИЯ, РЕГРЕССИОННЫЙ АНАЛИЗ, МЕТОДЫ КЛАССИФИКАЦИИ, МНОГОМЕРНОЕ ШКАЛИРОВАНИЕ

Keywords: APPLIED STATISTICS, DATA ANALYSIS, ESTIMATION, CONSISTENCY, OPTIMIZATION, REGRESSION ANALYSIS, CLASSIFICATION METHODS, MULTI-DIMENSIONAL SCALING

DOI: http://dx.doi.org/10.21515/1990-4665-162-002

1. Введение

По статистическим данным необходимо оценивать две составляющие вероятностно-статистических моделей - структуру моделей и параметры. Методы расчета состоятельных оценок параметров хорошо известны (например, применяют метод одношаговых оценок, который пришел на смену методу максимального правдоподобия). Структура модели обычно выбирается исследователем (можно сказать, что используются экспертные методы). Некоторые параметры структуры можно оценивать с помощью математико-статистических методов. Например, степень многочлена в регрессионной зависимости или число слагаемых в модели смеси распределений, используемой для классификации. Для подобных параметров модели используется общий термин - размерность вероятностно-статистической модели. Более общая составляющая модели - информативное подмножество признаков. В настоящей статье рассмотрено асимптотическом поведении оценок размерностей ряда моделей. Изучено асимптотическое поведение ряда оценок степени полинома при восстановлении зависимости. Получены состоятельные оценки размерности и структуры модели в регрессии. Рассмотрены подходы к оцениванию числа элементов смеси в задачах классификации. Обсуждаются оценки размерности модели в факторном анализе и многомерном шкалировании. С целью обоснования

последовательного выполнения этапов статистического анализа данных анализируются проблемы "стыковки" алгоритмов классификации и регрессии. Полезными оказываются оптимизационные формулировки ряда задач прикладной статистики. Основные результаты касаются состоятельности оценок. Краткие формулировки ряда теорем содержатся в ранее вышедших публикациях. Проблема оценивания размерности вероятностно-статистической модели как самостоятельное направление прикладной статистики впервые рассмотрена здесь. Впервые публикуются доказательства включенных в настоящую статью теорем. Эти доказательства и являются основными научными результатами работы.

2. Асимптотическое поведение ряда оценок степени полинома в регрессии

Во многих прикладных задачах требуется установить зависимость переменной y от переменных x_1, x_2, ..., x_m. Простейшая вероятностно-статистическая модель имеет вид

$$y = a_1 x_1 + a_1 x_1 + ... + a_m x_m + \varepsilon , \quad (1)$$

где a_j - коэффициенты линейной регрессии, j = 1, 2, ..., m, а ε - остаточный член, рассматриваемый обычно как погрешность измерения или результат влияния неучтенных факторов.

Исходные данные для определения (т.е. оценивания) коэффициентов регрессии имеют вид

$$(y_i, x_{1i}, x_{2i}, ..., x_{mi}),\ i = 1, 2, ..., n, \quad (2)$$

Рассмотрим модель с детерминированными x_{ji}, j = 1, 2, ..., m, i = 1, 2, ..., n. В классической вероятностно-статистической модели предполагается, что

$$y_i = \sum_{1 \le j \le m} a_j x_{ji} + \varepsilon_i , \quad (3)$$

где $\varepsilon_i, i = 1,2,...,n$, - независимые нормальные случайные величины с нулевым математическим ожиданием и дисперсией σ^2. Модель (3) обычно записывают в матричной форме:

$$Y = aX + \mathrm{E}, \quad (4)$$

где $Y = (y_1, y_2, ..., y_n)^T$ - вектор значений зависимой переменной, $a = (a_1, a_2, ..., a_m)$ - вектор неизвестных коэффициентов, $X = ||x_{ji}||$ - матрица значений независимых переменных, называемая также матрицей плана (в терминах теории планирования экспериментов), $\mathrm{E} = (\varepsilon_1, \varepsilon_2, ..., \varepsilon_n)^T$ - вектор погрешностей, T - символ транспонирования.

Литература по регрессионному анализу практически необозрима (миллионы названий статей и книг). В частности, многообразие моделей регрессионного анализа обсуждается в [1, 2]. Классическая теория изложена в [3, 4]. Неклассический подход развит в [5]. Вычислительные вопросы рассмотрены в [6]. Оптимальный выбор матрицы плана - предмет теории планирования эксперимента [7]. Ряд ссылок будет дан ниже.

Для модели (3) - (4) теория хорошо развита. Параметры оценивают методом наименьших квадратов, проверяют различные гипотезы. Однако в практических исследованиях часто возникает необходимость выделения "информативного подмножества признаков (независимых переменных)". При этом вместо (3) предполагается справедливой модель

$$y_i = \sum_{j \in J} a_j x_{ji} + \varepsilon_i, \ i = 1,2,...,n, \quad (5)$$

где J - подмножество множества $\{1, 2, ..., m\}$. Например, если модель (3) используется для управления технологическим процессом или для иного массового применения, то сокращение числа независимых переменных приносит ощутимый экономический эффект от сокращения числа измерений. В научных исследованиях выделение "информативного подмножества признаков" позволяет установить основные факторы, влияющие на изучаемое явление или процесс, и т.д. В дискуссии по

прикладной статистике, проведенной во время IV международной вильнюсской конференции по теории вероятностей и математической статистике (Вильнюс, 1985 г.) именно проблема выделения ""информативного подмножества признаков" J была признана наиболее актуальной.

Возникает задача построения состоятельной оценки J_n множества J, т.е. оценки, удовлетворяющей соотношению

$$\lim_{n\to\infty} Card(J_n \Delta J) = 0, \qquad (6)$$

где $Card$ (A) - число элементов конечного множества A.

Разработано большое число методов выделения информативного подмножества признаков (см., например, [8, гл.12], [9, гл.6]). Однако они обычно излагаются как эвристические, свойства их не изучены, неизвестно даже, справедливо ли свойство состоятельности (6). А если оно не выполнено, то, вообще говоря, нельзя гарантировать, что линия регрессии оценивается состоятельно. В рамках статистики нечисловых данных может быть получено (6) на основе общих результатов для решений экстремальных статистических задач [10, 11].

Хорошо известно, распределения реальных данных, как правило, не является нормальным [12, 13]. Однако математический аппарат в случае нормальности зачастую является более простым. Это связано с тем, что глубоко развита теория квадратичных форм в евклидовом пространстве (квадратичные формы стоят в степени экспоненты, описывающей плотность многомерного распределения). Поэтому для первоначального теоретического изучения считаем возможным использовать основанные на нормальности частные случаи регрессионных моделей.

В ряде случаев представляется естественным рассмотреть последовательность моделей вида (1) - (4).Например, изучается зависимость y от t. Естественно попытаться приблизить зависимость сначала константой, при недостаточной точности такого приближения

попробовать использовать линейную функцию, при неудаче - квадратическую, затем, если необходимо, - параболу третьего порядка, и т.д. [14]. Приближение y с помощью полинома порядка m - 1 описывается с помощью модели (1) - (4), если положить

$$x_1 = 1, x_2 = t, x_3 = t^2, ..., x_m = t^{m-1}. \quad (7)$$

В связи с (7) подчеркнем, что x_j в модели (1) - (4) не обязательно являются результатами прямых измерений. Более важным является случай, когда $x_j = f_j(x)$, j = 1, 2, ..., где x - исходные переменные, $f_j(x)$ - некоторые функции. (Модель (1) - (4) - частный случай такой формулировки, когда $x = (x_1, x_2, ..., x_m)$ и $f_j(x) = x_j$, $j = 1, 2, ..., m$. При этом x может иметь произвольную природу, в частности, быть объектом нечисловой природы.

В постановке (7) естественно считать, что модель (1) - (4) имеет место при некотором $m = m_0$, и искать это m_0, увеличивая m на 1, пока модель не будет адекватно описывать данные (подробнее см. ниже).Если априори задано достаточное (наверняка) число переменных M, то информативное подмножество признаков J естественно искать не среди всех подмножеств множества {1, 2, ..., M}, а среди подмножеств $J(m)$ = {1, 2, ..., m}, образующих расширяющуюся систему подмножеств $J(m) \subset J(m+1)$, m = 1, 2, ... Этим постановка (7) отличается от общей постановки (5). Другими словами, в случае (7) структуру модели задает не подмножество J, а натуральное число m_0, которое в соответствии с [15] называем *размерностью модели*.

Рассмотрим два метода, используемых прикладниками [9, 14, 16, 17] для оценки размерности модели m_0. Они основаны на применении "кажущейся ошибки", т.е. величины

$$\Delta_m = \frac{1}{n-m-1} \sum_{1 \le i \le n} (y_i - y_{im})^2 = \frac{1}{n-m-1} \Delta_m^0, \quad (8)$$

где y_{im} - сглаженные по методу наименьших квадратов значения зависимой переменной, полученные при принятии модели (7) с данным m.

Первый метод состоит в том, что в качестве оценки размерности модели, т.е. необходимого числа базисных функций, берут первый локальный минимум "кажущейся ошибки", т.е.

$$m_{1n} = \min\{m : \Delta_{m-1} > \Delta_m,\ \Delta_m \le \Delta_{m+1}\}. \quad (9)$$

Второй метод основан на проверке адекватности модели (3). При этом начинают с m =1 и увеличивают на 1 число параметров только в случае неадекватности, т.е. отклонения гипотезы о том, что данные (2) описываются моделью (3) при используемом m (в постановке (7) при этом увеличивается число используемых базисных функций f_j, т.е. степень полинома, но не число исходных независимых переменных). При известной дисперсии σ^2 для проверки указанной гипотезы можно воспользоваться тем, что Δ_m^0 имеет распределение $\sigma^2\chi^2_{n-m-1}$. Если σ^2 неизвестно, то применяют известный критерий Фишера: при $m_2 > m_1$ и справедливости (3) с $m = m_1$ статистика

$$f(m_1, m_2) = \frac{(n - m_2 - 1)(\Delta^0_{m_1} - \Delta^0_{m_2})}{(m_2 - m_1)\Delta^0_{m_2}} \quad (10)$$

имеет распределение Фишера с числом степеней свободы числителя m_2 - m_1 и знаменателя n - m_2 - 1, и гипотеза H_0: $m = m_1$ отвергается, если

$$f(m_1, m_2) \ge F_\alpha(m_2 - m_1; n - m_2 - 1), \quad (11)$$

где F_α есть $(1-\alpha)$-квантиль распределения Фишера с указанными степенями свободы, α - уровень значимости. Метод оценки размерности модели основан на том, что, рассматривая последовательно m_1 = 1, 2, ..., мы проверяем гипотезу H_0: $m = m_1$ с помощью (10) - (11) (выбор m_2 может быть проведен различными способами, например, $m_2 = m_1 + 1$ или $m_2 = 2m_1$) и останавливаемся на таком наименьшем m_1, что рассматриваемая гипотеза не отвергается. В постановке (7) наиболее естественно применять $m_2 = m_1 + 1$. При этом мы используем статистики

$$\xi_k = \frac{(g-k-2)(\Delta_k^0 - \Delta_{k+1}^0)}{\Delta_{k+1}^0}, \ k = 1,2,... \quad (12)$$

При $k \geq m_0$ статистика ξ_k имеет распределение Фишера с числом степеней свободы числителя 1 и знаменателя n - k - 2. В качестве оценки размерности модели используют согласно (11)

$$m_{2n} = \min\{k : \xi_k < F_\alpha(1, \ n-k-2). \quad (13)$$

Изучим поведение статистик m_{1n} и m_{2n} как оценок истинной размерности модели m_0. Заметим, что если модель (3) адекватна при $m = m_0$, то она адекватна и при любом $m' > m_0$ - достаточно положить $a_{m+1} = a_{m+2} = ... = a_{m'} = 0$. Поэтому истинная размерность m_0 - это минимальное m, при котором модель (3) адекватна.

Воспользуемся геометрической интерпретацией метода наименьших квадратов, рассмотренной А.Н. Колмогоровым [18] и изложенной, например, в [19, §§11,12]. Введем вектора

$$T_j = (x_{j1}, x_{j2}, ..., x_{jn})^T, j = 1, 2, ..., m. \quad (14)$$

В постановке (7) они имеют вид

$$T_1 = (1,1,...,1)^T, \ T_j = (t_1^{j-1}, t_2^{j-1}, ..., t_n^{j-1}), \ j = 2,..., m. \quad (15)$$

Тогда

$$Y = \sum_{1 \leq j \leq m} a_j T_j + \mathrm{E}, \quad (16)$$

где Y и E - те же, что и равенстве (4).

Введем в рассмотрение линейную оболочку

$$L_m = L_m (T_1, T_2, ..., T_m) \quad (17)$$

векторов T_1, T_2, ..., T_m. Ясно, что задача оценки параметров методом наименьших квадратов является частным случаем так называемой "общей линейной модели" [19, с.129]. Следовательно, наилучшей оценкой (в модели (3)) для вектора

$$Z = Y - \mathrm{E} = \sum_{1 \leq j \leq m} a_j T_j \quad (18)$$

является проекция Y как элемента евклидова пространства R^n на подпространство L_m. В случае линейной независимости векторов T_1, T_2, ..., T_m проекция однозначно определяет оценки $\hat{a}_j$ коэффициентов a_j, j = 1, 2, ..., m, а именно, оценкой a_j является коэффициент $\hat{a}_j$ в разложении проекции Y_{1m} по базису T_1, T_2, ..., T_m:

$$Y_{1m} = \mathrm{Pr}\,oj_{L_m} = \sum_{1 \le j \le m} \hat{a}_j T_j \,. \tag{19}$$

Имеем

$$Y = Y_{1m} + Y_{2m}, \tag{20}$$

где Y_{1m} - проекция Y на L_m, а Y_{2m} - проекция Y на ортогональное дополнение к L_m.. При этом

$$\Delta_m^0 = \| Y_{2m} \|^2 . \tag{21}$$

Пусть Q_{1n}, Q_{2n}, ..., Q_{mn} - ортонормированный базис в L_m (в предположении dim (L_m) = m). а $Q_{(m+1)n}$, $Q_{(m+2)n}$, ..., Q_{nn} - ортонормированный базис в $L_m^{\perp}$ - ортогональном дополнении к L_m . Тогда

$$Z = \sum_{1 \le j \le m} a_j T_j = \sum_{1 \le j \le n} b_{jn} Q_{jn} \,, \tag{22}$$

и

$$Y = \sum_{1 \le j \le n} \beta_{jn} Q_{jn} \,, \tag{23}$$

где

$$\beta_{jn} = b_{jn} + \sigma\delta_j \tag{24}$$

при j = 1, 2, ..., m_0 и

$$\beta_{jn} = \sigma\delta_j \tag{25}$$

при $j = m_0 + 1, ..., n$, где $\delta_1, \delta_2, ..., \delta_n$ - независимые нормальные случайные величины с нулевым математическим ожиданием и единичной дисперсией.

Что можно сказать о случайных величинах $|\beta_{jn}|$? Если исходный базис являлся ортогональным, то в пространстве L_m естественно использовать ортонормированный базис

$$Q_j = Q_{jn} = \frac{T_j}{\| T_j \|} \ . \tag{26}$$

Следовательно, в этом случае

$$\beta_{jn} = a_j \| T_j \| + \sigma\delta_j \tag{27}$$

при j = 1, 2, ..., m_0, а при $m > m_0$ величины β_{jn} задаются формулой (25). Поскольку

$$\| T_j \| = \sum_{1 \le i \le n} x_{ji}^2 \ , \tag{28}$$

в частности, в постановке (7) $\| T_1 \| = \sqrt{n}$, то для типичных прикладных задач

$$n^{-1} \| T_j \|^2 = O(1), \ \ n \| T_j \|^{-2} = O(1) \tag{29}$$

при $n \to \infty$.

Изучим асимптотическое поведение Δ_m при $n \to \infty$. При этом с изменением n вектора T_j размерности n, разумеется, меняются, и базис Q_{1n}, Q_{2n}, ..., Q_{mn} в L_m тоже, вообще говоря, меняется вместе с коэффициентами b_{jn}, j = 1, 2, ..., m, даже в случае, когда ортонормальный базис получаем ортогонализацией T_1, T_2, ..., T_m.

Для дальнейших рассуждений есть два пути. Один из них применили М.В. Гальченко и В.А. Гуревич [20]. Они ввели предположение, что матрица плана такова, что при каждом n вектора T_1, T_2, ..., T_m ортогональны. Примером является план [20, с.55] с

$$f_j(x) = \sqrt{2} \cos(j \arccos x), \ \ x \in [-1.1], \ \ x_{in} = \cos\left(\frac{2i-1}{2n} \pi \right). \tag{30}$$

Кроме того, они предполагают, что $a_j \ne 0$ при j = 1, 2, ..., m_0.

Специальный вид плана, на наш взгляд, излишнее ограничение. Дальнейшие рассуждения верны для плана "общего вида", нужны лишь некоторые условия регулярности, гарантирующие от "вырождения". Это и есть второй путь.

Имеем

$$\Delta_m = \frac{1}{n-m-1}\left(\beta^2_{(m+1)n} + ... + \beta^2_{nn}\right). \quad (31)$$

В силу (25) и справедливости (по теореме Чебышева) закона больших чисел для δ_j^2 имеем

$$\Delta_m \to \sigma^2 \quad (32)$$

по вероятности при $n \to \infty$, если $m \geq m_0$.

Пусть теперь $m < m_0$. Представим Δ_m в виде суммы двух слагаемых

$$\Delta_m = \frac{1}{n-m-1}\left(\beta^2_{(m+1)n} + ... + \beta^2_{m_0 n}\right) + \frac{1}{n-m-1}\left(\beta^2_{(m_0+1)n} + ... + \beta^2_{nn}\right). \quad (33)$$

Из (32) следует, что второе из них сходится по вероятности к σ^2 при $n \to \infty$. Если

$$\lim_{n\to\infty} \frac{b^2_{jn}}{n} = \gamma_j > 0, \quad j = 1,2,...,m_0, \quad (34)$$

то

$$\Delta_{m-1} - \Delta_m = \frac{1}{n}\beta^2_{mn}(1 + o(1)) \quad (35)$$

и для любого $\varepsilon > 0$

$$\lim_{n\to\infty} P(\Delta_{m-1} - \Delta_m \geq \gamma_m - \varepsilon) = 1. \quad (36)$$

Из (36) следует, что

$$\lim_{n\to\infty} P\{m_{1n} < m_0\} = 0. \quad (37)$$

Пусть теперь $m \geq m_0$. Имеем

$$\Delta_m - \Delta_{m+1} = \frac{1}{n-m-1}\left(\beta^2_{(m+1)n} - \frac{1}{n-m-2}\left(\beta^2_{(m+2)n} + ... + \beta^2_{nn}\right)\right). \quad (38)$$

В силу (32)

$$\lim_{n\to\infty} P\{\Delta_m - \Delta_{m+1} \leq 0\} = P\{\beta^2_{(m+1)n} \leq \sigma^2\} = \lambda, \quad (39)$$

где

$$\lambda = P\{\delta^2_{m+1} \leq 1\} = \frac{1}{\sqrt{2\pi}}\int_{-1}^{1} \exp\left\{-\frac{x^2}{2}\right\} dx = 0{,}68268... \quad (40)$$

Из (37) и (39) вытекает, что

$$\lim_{n\to\infty} P\{m_{1n} = m_0\} = \lim_{n\to\infty} P\{\Delta_{m_0} \le \Delta_{m_0+1}\} = \lambda. \quad (41)$$

В силу (38) и (32) величина

$$P\{m_{1n} = m_0 + k \mid m_{1n} \ge m_0\} = \lim_{n\to\infty} P\{\Delta_{m_0} > \Delta_{m_0+1}, \Delta_{m_0+1} > \Delta_{m_0+2}, \ldots, \Delta_{m_0+k-1} > \Delta_{m_0+k}, \Delta_{m_0+k} \le \Delta_{m_0+k+1}\} \quad (42)$$

сходится при $n \to \infty$ к

$$P\{\beta^2_{(m_0+1)n} > \sigma^2, \ldots, \beta^2_{(m_0+k)n} > \sigma^2, \beta^2_{(m_0+k+1)n} \le \sigma^2\}. \quad (43)$$

Из независимости $\beta_{(m_0+1)n}, \ldots, \beta_{(m_0+k+1)n}$ соотношений (25) и (39) вытекает, что

$$\lim_{n\to\infty} P\{m_{1n} = m_0 + k\} = \lambda(1-\lambda)^k, \ k = 0,1,2,\ldots, \quad (44)$$

где λ определено в (40). Итак, доказана следующая теорема, впервые полученная в [21].

Теорема 1. Пусть модель (1) - (4) верна при $m = m_0$. Пусть справедливы условия регулярности (34). Тогда имеют место предельные соотношения (37) и (44), т.е. распределение оценки m_{1n} в пределе является геометрическим.

Следствие. Оценка m_{1n} не является состоятельной (в смысле, принятом в математической статистике).

Замечание. Просматривается аналогия с последовательным анализом. В частности, соотношения типа (43) - (44) справедливы для декартовых последовательных критериев [22, с.485]. Специфика рассматриваемой задачи состоит в том, чтобы избавиться от зависимости последовательных проверок, что удается сделать в асимптотике с помощью соотношений типа (32). Представляется перспективным использование оптимальных правил остановки, разработанных в статистическом последовательном анализе [23]. Однако необходимо отметить, что типичные задачи последовательного анализа, в частности, задачи разладки и задачи последовательного различения простых гипотез с помощью критерия отношения вероятностей, существенно отличаются от рассматриваемых нами задач регрессионного анализа.

Условие (34) - это условие типа того, что мы находимся в ситуации "общего положения" (ср. [24]), т.е. отсутствует "вырождение". Если при всех n базис T_1, T_2, ..., T_m является ортогональным, как для плана (30), то согласно (23) и (26) $b_{jn} = a_j \|T_j\|$, а потому соотношение (34) эквивалентно тому, что $a_j \neq 0$ при $j = 1, 2, ..., m_0$ и

$$\lim_{n\to\infty}\left(\frac{1}{n}\|T_j\|^2\right) = \gamma_j',\ \ \gamma_j' = \frac{\gamma_j}{a_j},\ j = 1,2,...,m_0, \tag{45}$$

Соотношение (45) справедливо, например, для плана (30). Грубо говоря, условия (34) и (45) означают, что "вклады" вновь добавляемых переменных "не вырождаются", т.е. по порядку такие же, как вклад $T_1 = (1, 1, ..., 1)$ в постановке (7).

Рассмотрим теперь оценку m_{2n}. Согласно (10) и (21) имеем

$$f(m_1, m_2) = \frac{\frac{1}{m_2 - m_1}\left(\beta^2_{(m_1+1)n} + ... + \beta^2_{m_2 n}\right)}{\frac{1}{n - m_2 - 1}\left(\beta^2_{(m_2+1)n} + ... + \beta^2_{nn}\right)}. \tag{46}$$

Пусть выполнено условие (34). Тогда в силу (24) для любого $\varepsilon > 0$

$$\lim_{n\to\infty} P(\beta^2_{mn} > n\gamma_m(1-\varepsilon)) = 1,\ \ m = 1,2,...,m_0. \tag{47}$$

Если $m_1 < m_o$, то для числителя в (46) имеем:

$$\lim_{n\to\infty} P\left(\frac{1}{m_2 - m_1}\left(\beta^2_{(m_1+1)n} + ... + \beta^2_{m_2 n}\right) > \frac{n(1-\varepsilon)}{m_2 - m_1}\left(\gamma_{m_1+1} + ... + \gamma_{m_3}\right)\right) = 1 \tag{48}$$

для любого $\varepsilon > 0$, где $m_3 = \min(m_2, m_0)$.

Если $m_1 < m_0, m_2 \geq m_0$, то из (32) и (48) следует, что существует $C > 0$ такое, что

$$\lim_{n\to\infty} P\{f(m_1, m_2) > Cn\} = 1. \tag{49}$$

Пусть оценка m_{2n} размерности модели m_0 определяется с помощью последовательности троек

$$(m_1(k), m_2(k), F(k)),\ k = 1, 2, ...,\ \ m_1(k) < m_2(k), \tag{50}$$

где последовательности натуральных чисел $m_1(k)$, $m_2(k)$ возрастают. Гипотеза H_0: $m = m_1(k)$ против альтернативы H_1: $m = m_2(k)$ проверяется с помощью статистики $f(m_1(k), m_2(k))$, критическое значение выбирается согласно (11) с уровнем значимости $\alpha = F^{-1}_{m_2-m_1,n-m_2-1}(F(k))$. Это описание получения оценки m_{2n} - несколько более общее, чем данное ранее (формулы (10) - (13)), когда предполагалось, что $m_1(k) \equiv k$ и $F(k) \equiv F_\alpha$. Если гипотеза H_0 отвергается при k = 1, 2, ..., k_0 и впервые принимается при $k = k_0 + 1$, то полагаем $m_{2n} = m_1(k_0 + 1)$.

Теорема 2 [15]. Пусть выполнены условия (34), (52). Тогда

$$\lim_{n\to\infty} P(m_{2n} < m_0) = 0. \qquad (51)$$

Доказательство вытекает из соотношения (49), согласно которому при достаточно больших n гипотеза H_0 может быть принята только при $m_1(k) \ge m_0$. если известно, что $m_2(k) > m_0$. Остается рассмотреть случай, когда $m_2(k) \le m_0$. Для того, чтобы гипотеза H_0 отвергалась при любом $F(k)$ с вероятностью, стремящейся к 1 при $n \to \infty$, необходимо и достаточно, чтобы для любого $m < m_0$ было выполнено соотношение

$$\lim_{n\to\infty}\left(nb^2_{mn}\left(\sum_{m+1\le j\le m} b^2_{mn} \right)^{-1} \right) = \infty. \qquad (52)$$

Замечание. Как видно из проведенных рассуждений, для справедливости (51) нет необходимости требовать выполнения (34), достаточно справедливости (52) и условия

$$\lim_{n\to\infty} b^2_{mn} = \infty, \quad j = 1,2,...,m_0 \qquad (53)$$

Теорема 3 [15]. Пусть оценка размерности модели m_{2n} определяется с помощью последовательности проверок (50). Пусть выполнено (51). Тогда для любого целого $q \ge 0$ существует

$$p(q) = \lim_{n\to\infty} P\{m_{2n} = m_0 + q\}. \qquad (54)$$

Доказательство. С помощью (25) и (32) получаем из (46) и (11), что

$$\lim_{n\to\infty} P\{m_{2n} = m_0 + q\} = P\{\delta^2_{m_1(k)+1} + ... + \delta^2_{m_2(k)} \geq F(k)(m_2(k) - m_1(k)),$$
$$k_1 \leq k < k_2,\ \delta^2_{m_1(k)+1} + ... + \delta^2_{m_2(k)} < F(k)(m_2(k) - m_1(k)),\ k = k_2\}, \quad (55)$$

где $\{\delta_1, \delta_2, ..., \delta_m, ...\}$ - последовательность независимых нормальных случайных величин с нулевым математическим ожиданием и единичной дисперсией, $k_1 = \min\{k: m_1(k) \geq m_0\}$, число k_2 таково, что $m(k_2) = m_0 + q$. Если же $m_0 + q$ не принадлежит множеству $\{m_1(k), k = 1, 2, ...\}$, то очевидно, что $p(q) = 0$.

Теорема 4 [15, 25]. Пусть $m_1(k) = k$, $m_2(k) = k + 1$, $F(k) = F$, $k = 1, 2, ...$ Пусть выполнены условия (52), (53). Тогда

$$p(k) = \lambda(1-\lambda)^q,\ q = 0,1,2,... \quad (56)$$

где

$$\lambda = P\{\delta_1^2 < F) = \Phi\left(\sqrt{F}\right) - \Phi\left(-\sqrt{F}\right). \quad (57)$$

Доказательство. При данном в теореме 4 виде последовательности (50) статистика $f(m_1, m_2)$ переходит в ξ_k из (12). Согласно теореме 2 справедливо (51). Согласно теореме 3

$$p(q) = P\{\delta_k^2 \geq F,\ m_0 \leq k < m_0 + q,\ \delta^2_{m_0+q} < F\} = [P\{\delta_1^2 \geq F\}]^q P\{\delta_1^2 < F\}, \quad (58)$$

откуда и следует требуемое. Сравним предельное распределение оценки m_{1n} (формулы (44), (40)) и предельное распределение оценки m_{2n} (формулы (56), (57)). Видим, что при $F = 1$, т.е. при $\lambda = 0{,}68268...$, предельные распределения этих оценок совпадают. Поэтому можно сказать, что оценка m_{2n} обобщает оценку m_{1n}.

Обсудим значение основных предпосылок, при которых получены теоремы 1 - 4, а именно, нормальности погрешностей ε_i в (3), "условия невырожденности" (34) и аналогичных ему условий (52) - (53).

Нормальность распределений случайных величин ε_i используется для получения следующих двух утверждений: после ортогонализации базиса, т.е. перехода от $\{T_j\}$ к $\{Q_{jn}\}$ (см. (22) - (23)) ошибки по-прежнему независимы и одинаково распределены; параметр λ в (44) и (56)

выражается через нормальное распределение по формулам (40) и (57) соответственно.

Сохранение независимости ошибок при переходе к другому базису - характеризационное свойство нормального распределения. Это - следствие известного цикла характеризационных теорем [26], начатого работой С.Н. Бернштейна 1941 г. [27] и продолженного в исследованиях Б.В. Гнеденко [28], В.П. Скитовичем, Г. Дармуа, Ю.В. Линником, А.А. Зингером и др.

Отказаться от нормальности можно в предположении, указанном в [25] и принятом за основу в [20], что план эксперимента имеет специальный вид, обеспечивающий ортогональность базиса $\{T_j\}$ (тогда переход к $\{Q_{jn}\}$ не нужен). Примером является план (30). Пусть в этом случае ошибки ε_i - независимые одинаково распределенные случайные величины с конечным начальным вторым моментом $M(\varepsilon_i^2) = \mu_2 < \infty$. Пусть выполнено (34). Тогда, как нетрудно убедиться, проследив проведенные выше выкладки, выполнены соотношения (37) и (44) с $\lambda = P\{\varepsilon_1^2 \le \mu_2\}$. Если выполнены условия (52) и (53), то справедливы соотношения (51) и (57) с $\lambda = P\{\varepsilon_1^2 \le F\}$. Очевидно, можно отказаться и от предположения одинаковой распределенности помех ε_i, как это сделано в [20], но это делать здесь не будем, поскольку принципиально новых результатов при этом не получено, а демонстрировать владение техникой предельных теорем нет необходимости.

Невыполнение одного из условий (34), (53) в силу (29) практически эквивалентно (в предположении, что $\{T_j\}$ - ортогональный базис при всех n) тому, что модель (3) верна при $m = m_0$, но при некотором $j < m_0$ имеем $a_j = 0$. Каковы свойства оценок m_{1n} и m_{2n} в этом случае?

Для упрощения описания поведения оценок предположим, что существуют

$$\lim_{n\to\infty} \frac{b_{tn}^2}{n} = \rho_t > 0, \quad t = 1,2,\ldots,m_0. \tag{59}$$

Тогда согласно [25]

$$\lim_{n\to\infty} P\{m_{1n}= j\} = P\left\{\delta_j^2 \le \frac{1}{\sigma^2}\left(\rho_{j+1}+\ldots+\rho_{m_0}\right)+1\right\} \qquad (60)$$

и

$$\lim_{n\to\infty} P\{m_{2n}= j\} = P\left\{\frac{\sigma^2\delta_j^2}{\rho_{j+1}+\ldots+\rho_{m_0}+\sigma^2} \le \left[\Phi^{-1}\left(1-\frac{\alpha}{2}\right)\right]^2\right\},\ m_1(k)=k,\ m_2(k)=k+1, \qquad (61)$$

т.е. с достаточно высокой вероятностью произойдет преждевременный останов. От предположений (59) можно избавиться, заменив предельные переходы на сближение левых и правых частей (60) и (61) и ρ_t на $\frac{b_{tn}^2}{n}$.

3. Состоятельные оценки размерности и структуры модели в регрессии

Рассмотренные в предыдущем разделе методы оценки истинной размерности модели (3) не являются состоятельными:

$$\lim_{n\to\infty} P(m_{in} = m_0) \ne 1,\ \ i=1,2\,. \qquad (62)$$

В настоящем разделе рассмотрим построение состоятельных оценок m_n^* параметра m_0, т.е. оценок, для которых

$$\lim_{n\to\infty} P(m_n^* = m_0) = 1\,. \qquad (63)$$

В [20] предложена состоятельная модификация m'_{1n} оценки m_{1n}. В отличие от (9) в качестве оценки взят не первый локальный минимум "кажущейся ошибки" Δ_m, а первый локальный минимум линейной функции от неё $A_{nm}\Delta_m + B_{nm}$, где A_{nm} и B_{nm} - некоторые константы. Предположение [20] о специальном виде плана излишне, от него можно избавиться методами предыдущего раздела. Другие подходы рассмотрены в [8, 9, 29, 30, 31].

Состоятельную модификацию оценки m_{2n} можно получить, заменив правую часть в (11) на величину, растущую с увеличением n так, что

правая часть в (55) стремится к 0 при $n \to \infty$, но при этом выполнено (51). В частности, рассмотрим оценку

$$m_{3n} = \min\{k : \xi_k < \omega(n)\}, \quad (64)$$

где ξ_k определено в (12), $\omega(n)$ - некоторая последовательность. Для справедливости (55) необходимо и достаточно, чтобы

$$\lim_{n\to\infty} \omega(n) = +\infty. \quad (65)$$

Для справедливости (51) согласно доказательству теоремы 2 достаточно выполнения соотношений (34), (52) и

$$\lim_{n\to\infty} \frac{\omega(n)}{n} = 0. \quad (66)$$

Из проведенных рассуждений вытекает следующая теорема [25].

Теорема 5. Пусть выполнены соотношения (34), (52), (65) и (66). Тогда оценка $m_n^* = m_{3n}$, заданная формулой (64), является состоятельной оценкой размерности модели, т.е. удовлетворяет соотношению (63).

Рассмотрим некоторые другие методы оценки размерности модели, а также выбора информативного подмножества признаков. При этом весьма полезной оказывается независимость в совокупности получаемых по (23) - (25) оценок β_{jn} параметров регрессии в ортонормальном базисе $\{Q_{jn}\}$.

Упорядочим оценки β_{jn} в порядке убывания их абсолютной величины:

$$|\beta_{j(1)n}| \geq |\beta_{j(2)n}| \geq \ldots \geq |\beta_{j(n)n}|. \quad (67)$$

Предположим сначала, что σ известно. Выберем ν_n из условия

$$2(1-\Phi(\nu_n)) = \frac{1}{n}. \quad (68)$$

Тогда, как известно [32, с.410],

$$\nu_n \cong 2\sqrt{\ln n} \quad (69)$$

и, кроме того,

$$P\left\{\max_{0\leq i\leq n-1} |\delta_i| > \nu_n\right\} \leq \frac{1}{n}. \quad (70)$$

Оценку m^* размерности модели m_0 найдем из условия

$$|\beta_{j(m^*)n}| \geq \sigma v_n, \ |\beta_{j(m^*+1)n}| < \sigma v_n . \quad (71)$$

Если (см. (28))

$$\lim_{n\to\infty} (\ln n)^{-1/2} \| T_j \| = +\infty, \ \ j = 1,2,..., \quad (72)$$

то условие (71) дает состоятельную оценку размерности модели m_0 = Card J, а множество

$$J_n = \{j(1), j(2), \ldots , j(m^*)\} \quad (73)$$

является состоятельной оценкой информативного подмножества признаков J (см.(5)) в смысле (6).

Пусть теперь σ неизвестно. Укажем семейство оценок σ. Пусть $0 < \theta < 1$. Рассмотрим $\beta_{j([\theta n])n},\ldots,\beta_{j(n)n}$. При $n \to \infty$ выборочная дисперсия $s^2(\theta)$ этих случайных величин сходится к дисперсии $\sigma\xi$, где ξ - урезанная на отрезок $\left[-\Phi\left(1-\theta/2\right), \Phi\left(1-\theta/2\right)\right]$ стандартная нормальная случайная величина, т.е. $s^2(\theta)$ сходится к $\sigma^2(1-\theta)$. Следовательно, оценкой параметра σ^2 является $(1-\theta)^2 s^2(\theta)$. Эту оценку можно использовать в (71). Состоятельность описанных выше оценок при этом сохраняется.

Оценки (71) и (73) рассмотрены согласно [25]. В ситуации, когда исходный базис не является ортонормальным, требуются некоторые пояснения типа тех, что были даны выше в связи с работой М.В. Гальченко и В.А. Гуревича [20] (см. (30)). От (5) следует перейти к аналогичной записи в ортонормальном базисе$\{Q_{jn}\}$, вообще говоря, зависящем от n. Примем, что базис $\{Q_{jn}\}$ получен ортогонализацией и нормированием исходного базиса. Тогда вместо (5) имеем

$$Y = \sum_{j\in J(n)} b_{jn} Q_{jn} + \mathrm{E}, \quad (74)$$

где

$$\max\{ j : j \in J(n)\} = ma[\{ j : j \in J\} = j_0 . \quad (75)$$

Если

$$\lim_{n\to\infty}\min_{1\le j\le j_0}(\ln n)^{-1/2}\,|\,b_{jn}\,|=+\infty \qquad (76)$$

то справедлив аналог состоятельности оценок

$$\lim_{n\to\infty} Card(J_n \Delta J(n)) = 0\,. \qquad (77)$$

Другой поход к нахождению информативного подмножества признаков - метод "всех регрессий" [8] - основан на статистике

$$\text{£}_{nk} = \underset{a_j, J\in A_k}{Arg\min}\sum_{i=1}^{n}\left(\sum_{j\in J} a_j x_{ji} - y_i\right)^2 = \underset{J\in A_k}{Arg\min}\, g\,, \qquad (78)$$

где

$$g = g(J, a_j, x_{ji}, y_i,\ 1\le j\le m,\ 1\le i\le n) = \sum_{i=1}^{n}\left(\sum_{j\in J} a_j x_{ji} - y_i\right)^2, \qquad (79)$$

а *Arg* min берется по всем J таким, что $J \in A_k$, т.е.

$$Card\,(J) \le k\,. \qquad (80)$$

Рассмотрим функцию

$$h_k(\omega) = \min_{J\in A_k}\ \min_{a_j, j\in J} g\,. \qquad (81)$$

Из результатов об асимптотике решений экстремальных статистических задач [10] следует, что по вероятности

$$\lim_{n\to\infty} h_k(\omega) = h_k\,, \qquad (82)$$

где (в общей ситуации) функция h_k сначала убывает при росте k от k = 1 до k = Card($J_{ист}$), затем остается постоянной (равной h), а

$$\lim_{n\to\infty} P\left(\text{£}_{nk} \in \left\{J : \min_{a_j} M\left(\sum_{j\in J} a_j x_{ji} - y_i\right)^2 = \min_J = h\right\}\right) = 1\,. \qquad (83)$$

Отсюда следует, что метод "всех регрессий", вообще говоря, не дает состоятельных оценок истинного множества информативных признаков $J_{ист}$, а даёт оценки "с завышением", что выражается формулой (83). Это означает, что разнообразные программно-алгоритмические методы нахождения "наилучшей" регрессии [8, гл.12; 9, гл.6], в которых не

обращается внимание на отличие (83) от желаемой состоятельности (6), нуждаются в более тщательном изучении.

4. Оценивание числа элементов смеси в задачах классификации

Среди задач классификации [33, 34] важное место занимают задачи расщепления смесей. В них принимают, что наблюдается выборка из распределения с плотностью

$$f(x) = \sum_{1 \le i \le m} \pi_i f_i(x), \qquad (84)$$

где плотности $f_i(x)$ описывают отдельные классы, а π_i - веса этих классов, $\pi_i > 0,\ \pi_1 + \pi_2 + ... + \pi_m = 1$. Часто считают, что $f_i(x) = f(x, \theta_i)$, т.е. плотности элементов смеси взяты из некоторого параметрического семейства, $\theta_i \in \Theta$. Запись (84) можно рассматривать также как приближение плотности $f(x)$ с помощью линейной комбинации плотностей $f_1(x)$, $f_1(x)$, ... в этом случае веса π_i не обязаны быть положительными, а вместо равенства (84) имеет быть предельный переход.

Смеси встречаются в различных прикладных задачах. Так, Э.С. Эренбург моделировал продолжительность безотказной работы изделий бытовой техники как смесь двух классов - изделий со скрытыми дефектами и изделий без скрытых дефектов [35].

Если число слагаемых в сумме (84) известно и все $\pi_i > 0$, то с теоретической точки зрения оценивание параметров π_i и θ_i не представляет трудностей - можно применять оценки максимального правдоподобия или одношаговые оценки [36]. Рассмотрим оценивание числа слагаемых. Вначале приведем один известный результат.

Пусть $\xi_1, \xi_2, ..., \xi_n$, - выборка из совокупности с плотностью $f(x, \theta)$, где параметр $\theta \in \Omega$ имеет размерность r. Пусть подпространство $\Omega_0 \subset \Omega$ имеет

размерность $r' < r$. Для проверки гипотезы $H_0 : \theta \in \Omega_0$ при альтернативе $H_1 : \theta \in \Omega \setminus \Omega_0$ применяют критерий отношения правдоподобия

$$\lambda(\Omega_0, \Omega) = \sup_{\theta \in \Omega_0} \prod_{1 \le i \le n} f(\xi_i, \theta) \Big(\sup_{\theta \in \Omega} \prod_{1 \le i \le n} f(\xi_i, \theta) \Big). \quad (85)$$

В [22, §13.8] при некоторых условиях регулярности показано, что при $\theta \in \Omega_0$ распределение случайной величины $(-2 \log \lambda(\Omega_0, \Omega)$ сходится при $n \to \infty$ к распределению хи-квадрат с r - r' степенями свободы. Это доказывается путем построения r - r' независимых стандартных нормальных случайных величин $\eta_1, \eta_2, ..., \eta_{r-r'}$, таких, что

$$(-2 \log \lambda(\Omega_0, \Omega) = \sum_{1 \le j \le r-r'} \eta_j^2 + o(1) \quad (86)$$

по вероятности при $n \to \infty$.

Рассмотрим последовательность описанных выше задач. Пусть $\Omega_0 \subset \Omega_1 \subset \Omega_2 \subset ...$ - последовательность пространств параметров,

$$\dim \Omega_i = r' + iq, \ i = 0,1,2,... \quad (87)$$

при некоторых r' и q. Пусть проводится проверка гипотез $H_i : \theta \in \Omega_i$ при альтернативах H_{i+1} последовательно при i = 0, 1, 2, ... Проверки проводятся с помощью статистики $\lambda(\Omega_i, \Omega_{i+1})$ (см. (85)), гипотеза H_i отвергается, если $(-2 \log \lambda(\Omega_i, \Omega_{i+1}) > \lambda_\gamma$, где λ_γ есть $100(1-\gamma)$-процентная точка распределения χ^2 с q степенями свободы. Пусть впервые при i = m* гипотеза H_i не отвергнута. Какова предельное распределение m* при $n \to \infty$?

Пусть $\theta \in \Omega_{m(0)}$ и $\theta \notin \Omega_{m(0)-1}$ Так же, как в разделе 2 настоящей статьи, можно показать, что при некоторых условиях регулярности [22]

$$\lim_{n \to \infty} P(m^* < m(0)) = 0, \ \lim_{n \to \infty} P(m^* = m(0) + a) = \gamma^a (1-\gamma), \ a = 0,1,2,... \quad (88)$$

При доказательстве используется независимость главных членов в разложениях типа (85) для $(-2 \log \lambda(\Omega_i, \Omega_{i+1}))$. Как и в разделе 3 настоящей статьи, состоятельную оценку $m(0)$ получаем, сделав γ зависящим от n.

С формальной точки зрения частным случаем рассматриваемой последовательности проверок является определение числа элементов смеси (параметра *m* в модели (84)). При этом $\Omega_s = \{(\pi_1,...,\pi_s,\theta_1,...,\theta_s)\}$. Тогда в (87) $r' = \dim\theta,\ q = \dim\theta + 1$.

Однако в силу специфики модели (84) соотношения (88) верны не всегда, в частности, они неверны, если рассматривается смесь нормальных распределений [37]. Поскольку необходимо $\Omega_i \subset \Omega_{i+1}$, а точка $\theta \in \Omega_0$ в (85) должна быть внутренней, то ограничения $\pi_i > 0$ или $\pi_i \geq 0$ противоречат условиям регулярности Уилкса. Поэтому не будем принимать эти ограничения. Далее, информационная матрица вырождается, если $f_i(x,\theta_i)$ и $f_{i+1}(x,\theta_{i+1})$ могут совпадать, как это имеет место для смеси нормальных распределений. Действительно, если $f_i(x,\theta_i) = f_{i+1}(x,\theta_{i+1})$, то

$$\pi_i f_i(x,\theta_i) = \pi'_i\ f_i(x,\theta_i) + (\pi_i - \pi'_i)\, f_{i+1}(x,\theta_{i+1}), \qquad (89)$$

т.е. разложение в (84) неоднозначно. Поэтому предложение использовать критерий Уилкса для нормальных смесей нельзя признать обоснованным.

Предельное распределение (88), полученное для смеси (84) в [38], имеет место при справедливости условий регулярности Уилкса, например, когда задана последовательность линейно независимых плотностей $f_1(x)$, $f_2(x)$, ... и $\pi_i \in R^1$. Интересные результаты получены А.М. Никифоровым [39].

5. Оценка размерности модели в факторном анализе и многомерном шкалировании

Идея многомерного шкалирования состоит в представлении каждого объекта точкой геометрического пространства небольшой размерности (обычно размерности 1, 2 или 3), координатами которой служат "скрытые значения факторов", в совокупности достаточно адекватно описывающих объект. Размерности 1 - 3 позволяют провести визуальный анализ (о нем на

примере клинической медицины см. [40]). В прикладном многомерном статистическом анализе имеется большое число методов снижения размерности - факторный анализ, метод главных компонент, многомерное шкалирование [41, 42]), целенаправленное проецирование [43, 44]) (этой группе методов посвятил свой доклад П. Хубер на Первом Всемирном Конгрессе Общества математической статистики и теории вероятностей им. Бернулли [45]). Цель всех этих методов - от большого числа признаков перейти к существенно меньшему, вообще говоря, вновь сконструированных признаков, которые тем не менее достаточно адекватно описывают рассматриваемые объекты. Многомерное шкалирование использует не сами объекты (как вектора в многомерном пространстве), а расстояния между ними ρ_{ij}, вычисленные по координатам векторов или заданные иными способами, например, с использованием экспертов. Требуется подобрать точки-представители в евклидовом пространстве небольшой размерности так, чтобы расстояния между ними r_{ij} мало отличались от расстояний между объектами ρ_{ij}. Согласно одной из формализаций (в т.н. метрическом шкалировании) должна достигать минимума величина

$$S = \sum_{i<j} | \rho_{ij} - r_{ij} | . \qquad (90)$$

В настоящем разделе мы не будем пытаться подробно рассматривать многообразие методов рассматриваемого типа (см. указанную выше литературу и наши публикации [46, 47]), а разберем модельную постановку оценки размерности итогового пространства.

Пусть объекты описываются точками d_1, d_2, ..., d_n, в k-мерном евклидовом пространстве. Пусть L_m - пространство размерности m. Пусть $\rho(d_i, L_m)$ - расстояние между точкой d_i и линейным пространством L_m, и

$$f_n(m) = \min_{\{L_m\}} \sum_{1 \le i \le n} \rho(d_i, L_m) \ - \qquad (91)$$

- сумма расстояний точек d_1, d_2, ..., d_n до их наилучшего приближения гиперплоскостью размерности m. Пусть в рассматриваемой вероятностной модели

$$d_i = d_i^0 + \varepsilon_i, \quad (92)$$

где ε_i - независимые нормальные случайные вектора с математическим ожиданием 0 и ковариационной матрицей $\sigma^2 I$, где I - единичная матрица, точки d_i^0 лежат в гиперплоскости размерности m_0 и не лежат (одновременно все вместе) ни в какой гиперплоскости меньшей размерности. Тогда методами раздела 2 настоящей статьи установлено [46, с.68-70], что при $n \to \infty$ и соответствующих условиях регулярности (типа данных выше в разделе 2)

$$f_n(m) \to f(m) = f_1(m) + \sigma^2(k-m), \ m = 1,2,...,k, \quad (93)$$

по вероятности, где $f_1(m)$ - функция, зависящая от расположения точек $d_1^0, d_2^0, ..., d_n^0$. Примем для первичного анализа ситуации, что эти точки имеют круговое нормальное распределение в том подпространстве размерности m_0, в котором они лежат, т.е.

$$d_i^0 = \xi_i(1)e(1) + \xi_i(2)e(2) + ... + \xi_i(m_0)e(m_0), \quad (94)$$

где $e(1)$, $e(2)$, ..., $e(m_0)$ - ортонормальный базис в этом пространстве, а $\xi_i(1), \xi_i(2), ..., i = 1,2,...,n,$ - независимые нормальные случайные величины с математическими ожиданиями 0 и одинаковыми дисперсиями σ_0^2. Тогда в силу (93) имеем

$$f_1(m) = \begin{cases} \sigma_0^2(m_0 - m), & m < m_0, \\ 0, & m \geq m_0. \end{cases} \quad (95)$$

Таким образом, функция $f(m)$ из (93) линейна на отрезках [1, m_0] и [m_0, k], причем на первом отрезке она убывает быстрее, чем на втором. Отсюда следует, что статистика

$$m^* = \underset{m}{Arg\max}(f_n(m+1) - 2f_n(m) + f_n(m-1) \quad (96)$$

является состоятельной оценкой истинной размерности m_0 модели многомерного шкалирования.

Примечание. Если справедлива модель (94), упомянутые выше условия регулярности (типа рассматриваемых в разделе 2 настоящей статьи) выполнены.

Итак, из вероятностно-статистической теории вытекает рекомендация - определять размерность факторного пространства по правилу (96). Отметим: подобная рекомендация была сформулирована как эвристическая одним из основателей многомерного шкалирования Краскалом на основе опыта практического использования этого метода и вычислительных экспериментов (см., например, [42]). Вероятностная теория позволила обосновать эту эвристическую рекомендацию. Точнее, выше показано, что в достаточно естественной модели она приводит к состоятельной оценке.

К тематике настоящего раздела относятся также работы [48, 49].

6. Регрессия после классификации

Известно, что регрессионный анализ дает доступные интерпретации результаты лишь применительно к достаточно однородным совокупностям (см. обсуждение понятия "однородность" в [50]. Поэтому исходные данные рекомендуют разбить на однородные группы и лишь затем применять регрессионный анализ к каждой из них по отдельности.

Программный продукт по прикладной статистике обычно включает в себя ряд методов классификации и регрессии. Поскольку статистическое исследование включает в себя, как правило, последовательное применение не одного, а многих алгоритмов, работа предыдущего алгоритма может, вообще говоря, нарушать условия применимости последующих. Поэтому раздел 6 Рекомендаций ВНИИС [51] посвящен вопросам "стыковки" последовательно выполняемых алгоритмов: "При последовательном

применении нескольких методов обработки данных необходимо обеспечить проверку условий применения каждого последующего метода" [51, с.9].

Рассмотрим "стыковку" алгоритмов классификации и регрессии [52]. Пусть в результате работы некоторого алгоритма классификации выделена группа "однородных" наблюдений. Можно ли применять тот или иной метод регрессионного анализа [1] к элементам этой группы? Во-первых, эти элементы, вообще говоря, не являются независимыми, т.к. границы группы определяются по исходной выборке, а не задаются априорно. Во-вторых, наблюдения не могут иметь нормальное распределение, поскольку элементы группы ограничены по крайней мере с некоторых сторон (например, несколькими гиперплоскостями). Следовательно, обычные предпосылки регрессионного анализа не выполнены, а потому влияние отклонений от этих предпосылок на свойства алгоритмов требуют специального изучения (прежде всего, в рамках общей схемы устойчивости [53]).

В качестве примера рассмотрим "стыковку" алгоритмов классификации и регрессии, когда классификация сводится к расщеплению смеси (см. раздел 4 выше). Пусть для простоты $m = 2$ в смеси (84). Находят состоятельные оценки параметров смеси и строят с их помощью дискриминантную поверхность

$$g(x, \alpha_n) = \beta_n \qquad (97)$$

где x - элемент того пространства, в котором лежат наблюдения, функция g задает вид дискриминантной поверхности (в простейшем случае g - линейная функция), α_n и β_n - оценки параметров дискриминантной (разделяющей) поверхности. Если $g(\xi_j, \alpha_n) > \beta_n$, то наблюдение ξ_j относят к первому классу (совокупности)ЯЯ, в противном случае - ко второму. Зависимость наблюдений, попавших в один класс, имеет своей причиной

то, что параметры α_n и β_n определяются по всей исходной выборке, в том числе и по тем наблюдениям, что попали в рассматриваемый класс. Однако обычно существуют предельные значения α и β такие, что $\alpha_n \to \alpha$ и $\beta_n \to \infty$ по вероятности при $n \to \infty$. Тогда, как легко видеть, совместное распределение фиксированного конечного числа элементов одного класса стремится к совместному распределению независимых случайных элементов, распределение которых получено из рассмотрения соответствующего слагаемого в исходной смеси (84) усечением на область $\{x : g(x,\alpha) > \beta\}$ (для первого класса) или на область $\{x : g(x,\alpha) \le \beta\}$ (для второго класса).

Хотя в каждом из двух классов (кластеров) наблюдения и являются асимптотически независимыми, их распределения отличаются от $f_1(x)$ и $f_2(x)$ соответственно, т.е. от распределений, описывающих исходные классы. В частности, математические ожидания и ковариационные матрицы отличаются от исходных, поэтому с помощью выборочны характеристик, рассчитанных по кластерам, нельзя непосредственно оценивать характеристики исходных классов. Аналогичные выводы справедливы и для иных способов кластеризации [38].

Укажем два практически важных способа корректной "стыковки" алгоритмов классификации и регрессии. Один из них основан на объединении двух задач в одну. Так, принимая модель смеси (84), параметры регрессии определяют при помощи оценок параметров π_j и θ_j в (84). Действительно, при расщеплении смеси нормальных распределений оценивают математические ожидания и ковариационные матрицы каждого из исходных классов (описываемых плотностями $f(x,\theta_j)$), а этого достаточно для нахождения регрессии. Недостатками этого способа "стыковки" являются: "привязка" к определенной параметрической модели

(84), ограничение свободы выбора алгоритма классификации, большой объем вычислений.

Второй способ основан на использовании методов устойчивой регрессии, не опирающихся на предположение нормальности. При этом метод предварительной классификации может быть любым, но результаты расчетов относятся не к исходным классам в модели типа (84), а именно к тем таксонам (кластерам), что выделены алгоритмом классификации.

Мы видим, что двухэтапность обработки данных, при которой на первом этапе выделяются объекты нечисловой природы - кластеры, влечет необходимость выполнения определенных требований на втором этапе, а также предъявляется определенные требования к интерпретации результатов расчетов. Здесь методология статистики нечисловой природы вторгается в классическую область многомерного статистического анализа.

7. Использование оптимизационной формулировки ряда задач прикладной статистики

Основные задачи прикладной статистики допускают оптимизационную формулировку [11, 12], а потому предельная теория решений экстремальных статистических задач [10] позволяет получать полезные следствия для них. Так, результаты, относящиеся к экстремумам аддитивных статистик, непосредственно приложимы к статистикам минимального контраста. Частными случаями оценок минимального контраста являются оценки максимального правдоподобия, устойчивые оценки Тьюки-Хубера, оценки параметров в задаче аппроксимации (параметрической регрессии). Состоятельность оценок минимального контраста означает состоятельность всех перечисленных оценок, а также справедливость законов больших чисел в пространствах произвольной природы. (Отметим, что результаты [10 - 12] обобщают результаты [54].)

Поэтому каждая общая теорема типа полученных в [10 - 12] влечет за собой соответствующие следствия, касающиеся перечисленных и других конкретных областей. Так, например, в задаче конструирования факторов [55] результаты [10 - 12] описывают поведение отношения, аппроксимирующего систему матриц.

В качестве примера рассмотрим подробнее метод главных компонент. Пусть $\xi_1, \xi_2, ..., \xi_n$ - независимые одинаково распределенные случайные вектора размерности p. Кратко опишем экстремальную задачу, решаемую в методе главных компонент. Введем в рассмотрение координаты векторов: $\xi_j = (\xi_j(1), \xi_j(2), ..., \xi_j(p)), j = 1,2,..., n$. Рассмотрим p' линейных комбинаций

$$z_\infty(i) = \sum_{1 \le k \ge p} c_{ik} (\xi_1(k) - M\xi_1(k)), \ i = 1,2,..., p'. \quad (98)$$

В методе главных компонент используется функционал

$$I(C) = \frac{D(z_\infty(1)) + D(z_\infty(2)) + ... + D(z_\infty(p'))}{D(\xi_1(1)) + D(\xi_1(2)) + ... + D(\xi_1(p))}, \quad (99)$$

где $C=||c_{ik}||$. Формула (99) относится к вероятностной модели. При анализе статистических данных аналогом $I(C)$ является функционал $I_n(C)$, в котором теоретические дисперсии заменены выборочными. Легко видеть, что при $n \to \infty$ для любой матрицы C

$$I_n(C) \to I(C) \quad (100)$$

(сходимость по вероятности). Рассмотрим решения экстремальных задач

$$C_n = Arg \min_C(-I_n(C)), \ C_\infty = Arg \min_C(-I(C)) \quad (101)$$

Легко видеть, что условия асимптотической равномерной разбиваемости [10 - 12] выполнено, а потому

$$\lim_{n \to \infty} C_n = C_\infty \quad (102)$$

по вероятности, с учетом единственности решений задач (101).

В литературе по методу главных компонент (см., например, обзор [56]), теорему о справедливости соотношения (102) обнаружить не

удалось. Основное внимание уделяется нереалистическому случаю многомерной нормальности.

В ряде других задач прикладной статистики решения находятся путем минимизации функционала, также не являющегося аддитивным. Таковы различные варианты задач классификации, решаемые путем минимизации функционала качества, факторный анализ, метод экстремальной группировки признаков, отбор наиболее информативных признаков в моделях дискриминантного анализа, построение множества наиболее информативных переменных в моделях восстановления зависимостей (некоторые постановки разобраны выше в разделе 3), скалярная редукция многокритериальной оптимизационной схемы, т.е. экспертно-статистический метод построения обобщенного показателя "качества" в случае, когда экспертная информация - ранжировки, разбиения или результаты парных сравнений [57]. Во всех перечисленных задачах результаты [10-12] позволяют изучить асимптотическое поведение получаемых решений. Мы не будем подробно расписывать соответствующие результаты, поскольку это означало бы дать обзор основных задач прикладной статистики (см., в частности, [12, 58, 59]), обширный по объему, но не содержащий принципиально новых идей по сравнению со сказанным выше в настоящей статье и предыдущих публикациях.

Литература

1. Орлов А.И. Вероятностно-статистические модели корреляции и регрессии / Научный журнал КубГАУ. 2020. №160. С. 130–162.
2. Орлов А.И. Многообразие моделей регрессионного анализа (обобщающая статья) / Заводская лаборатория. Диагностика материалов. 2018. Т.84. №5. С. 63-73.
3. Кендалл М.Дж., Стьюарт А. Статистические выводы и связи. - М: Наука, 1973. - 900 с.
4. Демиденко Е.З. Линейная и нелинейная регрессия. - М.: Финансы и статистика, 1982. - 126 с.
5. Алгоритмы и программы восстановления зависимостей / Под ред. В.Н. Вапника. - М.: Наука, 1984. - 816 с.

6. Петрович М.Л. Регрессионный анализ и его математическое обеспечение на ЕС ЭВМ: Практическое руководство. - М.: Финансы и статистика, 1982. - 193 с.
7. Математическая теория планирования эксперимента / Справочная математическая библиотека. - М.: Наука, 1983. - 392 с.
8. Себер Дж. Линейный регрессионный анализ. - М.: Мир, 1980. - 456 с.
9. Дрейпер Н., Смит Г. Прикладной регрессионный анализ: Книга 2. - М.: Финансы и статистика, 1987. - 351 с.
10. Орлов А.И. Предельная теория решений экстремальных статистических задач / Научный журнал КубГАУ. 2017. №133. С. 579–600.
11. Орлов А.И. Организационно-экономическое моделирование: : учебник : в 3 ч. Ч.1: Нечисловая статистика. — М.: Изд-во МГТУ им. Н. Э. Баумана, 2009. — 542 с.
12. Орлов А.И. Прикладная статистика. - М.: Экзамен, 2006 - 671 с.
13. Орлов А.И. Распределения реальных статистических данных не являются нормальными / Научный журнал КубГАУ. 2016. №117. С. 71–90.
14. Налимов В.В. Теория эксперимента. - М.: Наука, 1971. - 208 с.
15. Орлов А.И. Оценка размерности модели в регрессии / Алгоритмическое и программное обеспечение прикладного статистического анализа. - М.: Наука, 1980. - С. 92-99.
16. Митропольский А.К. Техника статистических вычислений. - М.: Наука, 1971. - 570 с.
17. Пустыльник Е.И. Статистические методы анализа и обработки наблюдений. - М.: Наука, 1968. - 288 с.
18. Колмогоров А.Н. К обоснованию метода наименьших квадратов / Успехи математических наук. 1946. Т.1. Вып. 1. С.57-70.
19. Тутубалин В.Н. Теория вероятностей. - М.: МГУ, 1972. - 232 с.
20. Гальченко М.В., Гуревич А.В. Почти параметрическая оценка регрессии / Статистические методы оценивания и проверки гипотез: Межвузовский сборник научных трудов. - Пермь: Пермский ун-т, 1984. - С. 52-59.
21. Орлов А.И. Предельное распределение одной оценки числа базисных функций в регрессии / Прикладной многомерный статистический анализ. - М.: Наука, 1978. -= С. 380-381.
22. Уилкс С. Математическая статистика. - М.: Наука, 1967. - 632 с.
23. Ширяев А.Н. Статистический последовательный анализ: Оптимальные правила остановки. 2-е изд., перераб. - М.: Физматлит, 1976. - 272 с.
24. Арнольд В.И. О локальных задачах анализа / Вестник МГУ. Сер. матем. и мех. 1970. №2. С. 52-56.
25. Орлов А.И. Асимптотика некоторых оценок размерности модели в регрессии / Прикладная статистика. - М.: Наука, 1983. - С. 260-265.
26. Каган А.М., Линник Ю.В., Рао С.Р. Характеризационные задачи математической статистики. - М.: Наука, 1972. - 656 с.
27. Бернштейн С.Н. Об одном свойстве, характеризующем закон Гаусса / Труды Ленинградского политехн. ин-та. 1941. №3. С. 21-22. - Перепеч. в кн.: Бернштейн С.Н. Собрание сочинений: Т.IV: Теория вероятностей и математическая статистика. - М.: Наука, 1964. - С. 394-395, 569.
28. Гнеденко Б.В. Об одной теореме С.Н. Бернштейна / Известия АН СССР, Сер. матем. 1948. Т.12. №1. С. 97-100.
29. Боганик Г.Н. Об установлении порядка уравнения параболической регрессии / Теория вероятностей и её применения. 1967. Т.XII. №4. С. 750-763.
30. Киричук В.С. Выбор степени полинома, сглаживающего результаты измерений / Автометрия. 1970. №3. С. 26-71. 31.

31. Ковалерчук Б.Я., Лавков В.В. Поиск максимального верхнего нуля для минимизации числа признаков в регрессионном анализе / Журнал вычислительной математики и математической физики. 1984. Т.24. №3. С. 1241-1249.
32. Крамер Г. Математические методы статистики. - М.: Мир, 1975. - 648 с.
33. Орлов А.И. Математические методы теории классификации / Научный журнал КубГАУ. 2014. №95. С. 423 – 459.
34. Орлов А.И. Базовые результаты математической теории классификации / Научный журнал КубГАУ. 2015. №110. С. 219 – 239.
35. Эренбург Э.С. Смеси распределений в надежности. - М.: Знание, 1983. - 48 с.
36. Орлов А.И. Оценивание параметров: одношаговые оценки предпочтительнее оценок максимального правдоподобия / Научный журнал КубГАУ. 2015. №109. С. 208–237.
37. White H. Maximum likelihood estimation of misspecified models / Econometrics. 1982. V.50. N 1. P.1-25.
38. Орлов А.И. Некоторые вероятностные вопросы теории классификации / Прикладная статистика. - М.: Наука, 1983. - С.166-179.
39. Никифоров А.М. Исследование некоторых вопросов статистической теории распознавания образов с самообучением и анализа данных с пропусками применительно к задаче обработки клинических данных / Дисс. ... канд. физ.-мат. наук. - М.: МФТИ, 1987. - 144 с.
40. Волынский Ю.Д., Курочкина А.И. Многомерный анализ клинических данных / Вестник АМН СССР. 1987. № 1. С. 84-93.
41. Перекрест В.Т. Нелинейный типологический анализ социально-экономической информации: Математические и вычислительные методы. - Л.: Наука, 1983. - 176 с.
42. Терехина А.Г. Анализ данных методами многомерного шкалирования. - М.: Наука, 1986. - 168 с.
43. Енюков И.С. Методы, алгоритмы, программы многомерного статистического анализа: Пакет ППСА. - М.: Финансы и статистика, 1986. - 232 с.
44. Huber P.J. Projection Pursuit / Ann. Statist. 1985. V/13. N 3. P. 435-476.
45. Орлов А.И. Первый Всемирный конгресс Общества математической статистики и теории вероятностей им. Бернулли // Надежность и контроль качества. 1987. №6. С. 54-59.
46. Орлов А.И. Общий взгляд на статистику объектов нечисловой природы / Анализ нечисловой информации в социологических исследованиях. - М.: Наука, 1985. - С. 58-92.
47. Тюрин Ю.Н., Литвак Б.Г., Орлов А.И., Сатаров Г.А., Шмерлинг Д.С. Анализ нечисловой информации / Препринт. - М.: Научный Совет АН СССР по комплексной проблеме "Кибернетика", 1981. - 80 с.
48. Орлов А.И. Методы снижения размерности / Приложение 2 к книге: Толстова Ю.Н. Основы многомерного шкалирования. - М.: Издательство КДУ, 2006. С. 113-120.
49. Орлов А.И., Луценко Е.В. Методы снижения размерности пространства статистических данных / Научный журнал КубГАУ. 2016. №119. С. 92–107.
50. Смоляк С.А., Титаренко Б.П. Устойчивые методы оценивания: Статистическая обработка неоднородных совокупностей. - М.: Статистика, 1980. - 208 с.

51. Рекомендации: Прикладная статистика. Методы обработки данных. Основные требования и характеристики / Орлов А.И., Миронова Н.Г., Фомин В.Н., Черчинцев А.Н. - М.: ВАНИИСБ 1987. - 64 с.
52. Орлов А.И. Основные требования к методам анализа данных (на примере задач классификации) / Научный журнал КубГАУ. 2020. №159. С. 239–267.
53. Орлов А.И. Новый подход к изучению устойчивости выводов в математических моделях / Научный журнал КубГАУ. 2014. № 100. С. 146-176.
54. Reise R.D. Consistency of minimum contrast estimators in nonstandart case / Metriks. 1978. V.25. N 3. P. 129-142.
55. Миркин Б.Г. Анализ качественных признаков и структур. - М.: Статистика, 1980. - 319 с.
56. Андрукович П.Ф. Некоторые свойства метода главных компонент / Многомерный статистический анализ в социально-экономических исследованиях. - М.: Наука, 1974. - С. 189-228.
57. Орлов А.И. Организационно-экономическое моделирование : учебник : в 3 ч. Ч.2. Экспертные оценки. — М.: Изд-во МГТУ им. Н. Э. Баумана, 2011. — 486 с.
58. Орлов А.И., Луценко Е.В. Системная нечеткая интервальная математика. Монография (научное издание). – Краснодар, КубГАУ. 2014. – 600 с.
59. Лойко В.И., Луценко Е.В., Орлов А.И. Высокие статистические технологии и системно-когнитивное моделирование в экологии : монография. – Краснодар : КубГАУ, 2019. – 258 с.

References

1. Orlov A.I. Veroyatnostno-statisticheskie modeli korrelyacii i regressii / Nauchnyj zhurnal KubGAU. 2020. №160. S. 130–162.
2. Orlov A.I. Mnogoobrazie modelej regressionnogo analiza (obobshchayushchaya stat'ya) / Zavodskaya laboratoriya. Diagnostika materialov. 2018. T.84. №5. S. 63-73.
3. Kendall M.Dzh., St'yuart A. Statisticheskie vyvody i svyazi. - M: Nauka, 1973. - 900 s.
4. Demidenko E.Z. Linejnaya i nelinejnaya regressiya. - M.: Finansy i statistika, 1982. - 126 s.
5. Algoritmy i programmy vosstanovleniya zavisimostej / Pod red. V.N. Vapnika. - M.: Nauka, 1984. - 816 s.
6. Petrovich M.L. Regressionnyj analiz i ego matematicheskoe obespechenie na ES EVM: Prakticheskoe rukovodstvo. - M.: Finansy i statistika, 1982. - 193 s.
7. Matematicheskaya teoriya planirovaniya eksperimenta / Spravochnaya matematicheskaya biblioteka. - M.: Nauka, 1983. - 392 s.
8. Seber Dzh. Linejnyj regressionnyj analiz. - M.: Mir, 1980. - 456 s.
9. Drejper N., Smit G. Prikladnoj regressionnyj analiz: Kniga 2. - M.: Finansy i statistika, 1987. - 351 s.
10. Orlov A.I. Predel'naya teoriya reshenij ekstremal'nyh statisticheskih zadach / Nauchnyj zhurnal KubGAU. 2017. №133. S. 579–600.
11. Orlov A.I. Organizacionno-ekonomicheskoe modelirovanie: : uchebnik : v 3 ch. CH.1: Nechislovaya statistika. — M.: Izd-vo MGTU im. N. E. Baumana, 2009. — 542 s.
12. Orlov A.I. Prikladnaya statistika. - M.: Ekzamen, 2006 - 671 s.
13. Orlov A.I. Raspredeleniya real'nyh statisticheskih dannyh ne yavlyayutsya normal'nymi / Nauchnyj zhurnal KubGAU. 2016. №117. S. 71–90.
14. Nalimov V.V. Teoriya eksperimenta. - M.: Nauka, 1971. - 208 s.

15. Orlov A.I. Ocenka razmernosti modeli v regressii / Algoritmicheskoe i programmnoe obespechenie prikladnogo statisticheskogo analiza. - M.: Nauka, 1980. - S. 92-99.
16. Mitropol'skij A.K. Tekhnika statisticheskih vychislenij. - M.: Nauka, 1971. - 570 s.
17. Pustyl'nik E.I. Statisticheskie metody analiza i obrabotki nablyudenij. - M.: Nauka, 1968. - 288 s.
18. Kolmogorov A.N. K obosnovaniyu metoda naimen'shih kvadratov / Uspekhi matematicheskih nauk. 1946. T.1. Vyp. 1. S.57-70.
19. Tutubalin V.N. Teoriya veroyatnostej. - M.: MGU, 1972. - 232 s.
20. Gal'chenko M.V., Gurevich A.V. Pochti parametricheskaya ocenka regressii / Statisticheskie metody ocenivaniya i proverki gipotez: Mezhvuzovskij sbornik nauchnyh trudov. - Perm': Permskij un-t, 1984. - S. 52-59.
21. Orlov A.I. Predel'noe raspredelenie odnoj ocenki chisla bazisnyh funkcij v regressii / Prikladnoj mnogomernyj statisticheskij analiz. - M.: Nauka, 1978. -= S. 380-381.
22. Uilks S. Matematicheskaya statistika. - M.: Nauka, 1967. - 632 s.
23. SHiryaev A.N. Statisticheskij posledovatel'nyj analiz: Optimal'nye pravila ostanovki. 2-e izd., pererab. - M.: Fizmatlit, 1976. - 272 s.
24. Arnol'd V.I. O lokal'nyh zadachah analiza / Vestnik MGU. Ser. matem. i mekh. 1970. №2. S. 52-56.
25. Orlov A.I. Asimptotika nekotoryh ocenok razmernosti modeli v regressii / Prikladnaya statistika. - M.: Nauka, 1983. - S. 260-265.
26. Kagan A.M., Linnik YU.V., Rao S.R. Harakterizacionnye zadachi matematicheskoj statistiki. - M.: Nauka, 1972. - 656 s.
27. Bernshtejn S.N. Ob odnom svojstve, harakterizuyushchem zakon Gaussa / Trudy Leningradskogo politekhn. in-ta. 1941. №3. S. 21-22. - Perepech. v kn.: Bernshtejn S.N. Sobranie sochinenij: T.IV: Teoriya veroyatnostej i matematicheskaya statistika. - M.: Nauka, 1964. - S. 394-395, 569.
28. Gnedenko B.V. Ob odnoj teoreme S.N. Bernshtejna / Izvestiya AN SSSR, Ser. matem. 1948. T.12. №1. S. 97-100.
29. Boganik G.N. Ob ustanovlenii poryadka uravneniya parabolicheskoj regressii / Teoriya veroyatnostej i eyo primeneniya. 1967. T.XII. №4. S. 750-763.
30. Kirichuk V.S. Vybor stepeni polinoma, sglazhivayushchego rezul'taty izmerenij / Avtometriya. 1970. №3. S. 26-71. 31.
31. Kovalerchuk B.YA., Lavkov V.V. Poisk maksimal'nogo verhnego nulya dlya minimizacii chisla priznakov v regressionnom analize / ZHurnal vychislitel'noj matematiki i matematicheskoj fiziki. 1984. T.24. №3. S. 1241-1249.
32. Kramer G. Matematicheskie metody statistiki. - M.: Mir, 1975. - 648 s.
33. Orlov A.I. Matematicheskie metody teorii klassifikacii / Nauchnyj zhurnal KubGAU. 2014. №95. S. 423 – 459.
34. Orlov A.I. Bazovye rezul'taty matematicheskoj teorii klassifikacii / Nauchnyj zhurnal KubGAU. 2015. №110. S. 219 – 239.
35. Erenburg E.S. Smesi raspredelenij v nadezhnosti. - M.: Znanie, 1983. - 48 s.
36. Orlov A.I. Ocenivanie parametrov: odnoshagovye ocenki predpochtitel'nee ocenok maksimal'nogo pravdopodobiya / Nauchnyj zhurnal KubGAU. 2015. №109. S. 208–237.
37. White H. Maximum likelihood estimation of misspecified models / Econometrics. 1982. V.50. N 1. P.1-25.
38. Orlov A.I. Nekotorye veroyatnostnye voprosy teorii klassifikacii / Prikladnaya statistika. - M.: Nauka, 1983. - S.166-179.

39. Nikiforov A.M. Issledovanie nekotoryh voprosov statisticheskoj teorii raspoznavaniya obrazov s samoobucheniem i analiza dannyh s propuskami primenitel'no k zadache obrabotki klinicheskih dannyh / Diss. ... kand. fiz.-mat. nauk. - M.: MFTI, 1987. - 144 s.

40. Volynskij YU.D., Kurochkina A.I. Mnogomernyj analiz klinicheskih dannyh / Vestnik AMN SSSR. 1987. № 1. S. 84-93.

41. Perekrest V.T. Nelinejnyj tipologicheskij analiz social'no-ekonomicheskoj informacii: Matematicheskie i vychislitel'nye metody. - L.: Nauka, 1983. - 176 s.

42. Terekhina A.G. Analiz dannyh metodami mnogomernogo shkalirovaniya. - M.: Nauka, 1986. - 168 s.

43. Enyukov I.S. Metody, algoritmy, programmy mnogomernogo statisticheskogo analiza: Paket PPSA. - M.: Finansy i statistika, 1986. - 232 s.

44. Huber P.J. Projection Pursuit / Ann. Statist. 1985. V/13. N 3. P. 435-476.

45. Orlov A.I. Pervyj Vsemirnyj kongress Obshchestva matematicheskoj statistiki i teorii veroyatnostej im. Bernulli // Nadezhnost' i kontrol' kachestva. 1987. №6. S. 54-59.

46. Orlov A.I. Obshchij vzglyad na statistiku ob"ektov nechislovoj prirody / Analiz nechislovoj informacii v sociologicheskih issledovaniyah. - M.: Nauka, 1985. - S. 58-92.

47. Tyurin YU.N., Litvak B.G., Orlov A.I., Satarov G.A., SHmerling D.S. Analiz nechislovoj informacii / Preprint. - M.: Nauchnyj Sovet AN SSSR po kompleksnoj probleme "Kibernetika", 1981. - 80 s.

48. Orlov A.I. Metody snizheniya razmernosti / Prilozhenie 2 k knige: Tolstova YU.N. Osnovy mnogomernogo shkalirovaniya. - M.: Izdatel'stvo KDU, 2006. S. 113-120.

49. Orlov A.I., Lucenko E.V. Metody snizheniya razmernosti prostranstva statisticheskih dannyh / Nauchnyj zhurnal KubGAU. 2016. №119. S. 92–107.

50. Smolyak S.A., Titarenko B.P. Ustojchivye metody ocenivaniya: Statisticheskaya obrabotka neodnorodnyh sovokupnostej. - M.: Statistika, 1980. - 208 s.

51. Rekomendacii: Prikladnaya statistika. Metody obrabotki dannyh. Osnovnye trebovaniya i harakteristiki / Orlov A.I., Mironova N.G., Fomin V.N., CHerchincev A.N. - M.: VANIISB 1987. - 64 s.

52. Orlov A.I. Osnovnye trebovaniya k metodam analiza dannyh (na primere zadach klassifikacii) / Nauchnyj zhurnal KubGAU. 2020. №159. S. 239–267.

53. Orlov A.I. Novyj podhod k izucheniyu ustojchivosti vyvodov v matematicheskih modelyah / Nauchnyj zhurnal KubGAU. 2014. № 100. S. 146-176.

54. Reise R.D. Consistency of minimum contrast estimators in nonstandart case / Metriks. 1978. V.25. N 3. P. 129-142.

55. Mirkin B.G. Analiz kachestvennyh priznakov i struktur. - M.: Statistika, 1980. - 319 s.

56. Andrukovich P.F. Nekotorye svojstva metoda glavnyh komponent / Mnogomernyj statisticheskij analiz v social'no-ekonomicheskih issledovaniyah. - M.: Nauka, 1974. - S. 189-228.

57. Orlov A.I. Organizacionno-ekonomicheskoe modelirovanie : uchebnik : v 3 ch. CH.2. Ekspertnye ocenki. — M.: Izd-vo MGTU im. N. E. Baumana, 2011. — 486 s.

58. Orlov A.I., Lucenko E.V. Sistemnaya nechetkaya interval'naya matematika. Monografiya (nauchnoe izdanie). – Krasnodar, KubGAU. 2014. – 600 s.

59. Lojko V.I., Lucenko E.V., Orlov A.I. Vysokie statisticheskie tekhnologii i sistemno-kognitivnoe modelirovanie v ekologii : monografiya. – Krasnodar : KubGAU, 2019. – 258 s.