УДК 330.43 : 519.2

5.2.2. Математические, статистические и инструментальные методы экономики (экономические науки)

АСИМПТОТИКА РАСТУЩЕЙ РАЗМЕРНОСТИ

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

Рассмотрены статистические задачи, в которых число параметров статистических моделей не является пренебрежимо малым по сравнению с объем выборки. В предложенной А.Н. Колмогоровым асимптотике растущей размерности объем выборки и число параметров безгранично растут так, что их отношение стремится к некоторой положительной константе. В статье обращается внимание на недостаточность традиционного подхода, на основе которого даются не всегда удовлетворительные рекомендации по решению статистических задач, отличающихся большим числом оцениваемых параметров. Даны предложения по включению некоторых процедур, разработанных для многопараметрических задач, в программные продукты по статистическому анализу данных. Описаны эффекты, возникающие при оценивании большого числа параметров, и возможные подходы к решению многопараметрических задач прикладной статистики. Развитие многопараметрических методов статистического анализа опирается на свойства выборочных характеристик случайных матриц растущей размерности, прежде всего выборочных ковариационных матриц. Одним из серьезных достижений в развитии многопараметрических методов статистического анализа является разработка процедуры асимптотически экстремального дискриминантного анализа. Рассмотрен также случай зависимых компонент случайного вектора. Всё более важным становится класс многопараметрических статистических задач, в которых число показателей (компонент векторов наблюдений) настолько велико, что намного превосходит объем выборки. В качестве примера рассмотрена теория люсианов – конечных последовательностей независимых испытаний Бернулли с, вообще говоря, различными вероятностями успеха. Метод проверки гипотез по совокупности малых выборок (т.е. в асимптотике, когда объем выборки фиксирован, а число параметров безгранично растет) основан на

UDC 330.43: 519.2

5.2.2. Mathematical, statistical and instrumental methods of economics (economic sciences)

ASYMPTOTICS OF INCREASING DIMENSION

Orlov Alexander Ivanovich
Dr.Sci.Econ., Dr.Sci.Tech., Cand.Phys-Math.Sci., professor

Bauman Moscow State Technical University, Moscow, Russia

Statistical problems in which the number of parameters of statistical models is not negligible compared to the sample size are considered. In the asymptotics of increasing dimension proposed by A.N. Kolmogorov, the sample size and the number of parameters grow infinitely so that their ratio tends to some positive constant. The article draws attention to the insufficiency of the traditional approach, on the basis of which not always satisfactory recommendations are given for solving statistical problems with a large number of estimated parameters. Suggestions are given for the inclusion of some procedures developed for multiparametric tasks in software products for statistical data analysis. The effects arising from the estimation of a large number of parameters and possible approaches to solving multiparametric problems of applied statistics are described. The development of multiparametric methods of statistical analysis is based on the properties of sample characteristics of random matrices of increasing dimension, primarily sample covariance matrices. One of the major achievements in the development of multiparametric methods of statistical analysis is the development of a procedure for asymptotically extreme discriminant analysis. The case of dependent components of a random vector is also considered. A class of multiparametric statistical problems is becoming increasingly important, in which the number of indicators (components of observation vectors) is so large that it far exceeds the sample size. As an example, the theory of Lucians is considered – finite sequences of independent Bernoulli trials with, generally speaking, different success probabilities. The method of testing hypotheses for a set of small samples (i.e., in asymptotics, when the sample size is fixed and the number of parameters grows infinitely) is based on the use of unbiased estimates of zero, unbiased estimates of the variances of these estimates and distance (more precisely, pseudometrics) introduced using one or another axiom system. The necessity of further development

использовании несмещенных оценок нуля, несмещенных оценок дисперсий этих оценок и расстояния (точнее, псевдометрики), введенного с помощью той или иной системы аксиом. Обоснована необходимость дальнейшей разработки статистической теории в асимптотике Колмогорова, создания соответствующего методического и прикладного обеспечения, организации широкого внедрения уже полученных научных результатов

of the statistical theory in Kolmogorov's asymptotics, creation of appropriate methodological and applied support, organization of widespread implementation of the scientific results already obtained is substantiated

Ключевые слова: СТАТИСТИЧЕСКИЕ МЕТОДЫ, ПРИКЛАДНАЯ СТАТИСТИКА, АСИМПТОТИКА КОЛМОГОРОВА, МНОГОПАРАМЕТРИЧЕСКОЕ ОЦЕНИВАНИЕ, СЛУЧАЙНЫЕ МАТРИЦЫ, ДИСКРИМИНАНТНЫЙ АНАЛИЗ, ЛЮСИАН, РАССТОЯНИЕ

Keywords: STATISTICAL METHODS, APPLIED STATISTICS, KOLMOGOROV ASYMPTOTICS, MULTIVARIATE ESTIMATION, RANDOM MATRICES, DISCRIMINANT ANALYSIS, LUCIAN, DISTANCE

http://dx.doi.org/10.21515/1990-4665-205-022

Введение

Развитие математических, статистических и инструментальных методов экономики, прежде всего прикладной математической статистики, направлено на всё более полный учет особенностей реальных данных. Обсудим статистические задачи, в которых число p параметров статистических моделей не является пренебрежимо малым по сравнению с объем выборки n. Потребность в усложнении моделей диктуется неуклонным возрастанием значения научных методов в экономике и управлении и, одновременно, обеспечивается возрастанием возможностей информационно-коммуникационных технологий и искусственного интеллекта.

В классической схеме статистического исследования допускается возможность неограниченного выбора данных из генеральной совокупности при фиксированной модели. С этой схемой связана традиционная асимптотика математической статистики $p = const,\ n \to \infty$. Однако для современных применений статистического анализа эта схема часто оказывается непригодной. Например, при проведении испытаний промышленной продукции общее число изделий n приходится

ограничивать, но число p измеряемых параметров (показателей качества) целесообразно брать достаточно большим, поскольку затраты на увеличение числа измеряемых параметров зачастую существенно меньше затрат на увеличение количества изделий. В медицинских научных исследованиях число больных n обычно ограничено десятками и сотнями (например, из-за ограниченности коечного фонда клиники и временных рамок статистического исследования), в то время как число параметров p, описывающих больного (например, результатов различных анализов и обследований) может измеряться тысячами, как во многих видах автоматизированных историй болезни. При экономическом изучении предприятий определенной отрасли или региона общее число n рассматриваемых предприятий может быть значительно меньше числа p рассматриваемых показателей их финансово-хозяйственной деятельности (к настоящему времени число разработанных различными авторами подобных показателей измеряется тысячами). В социологических или маркетинговых исследованиях число вопросов в анкете также нельзя считать малым по сравнению с числом опрашиваемых (респондентов).

Отечественная научная школа в области теории вероятностей и математической статистики создана академиком АН СССР А.Н.Колмогоровым [1]. В течение полувека он интересовался статистическими постановками, в которых число неизвестных параметров растет вместе с объемом данных. К ним относится и весьма актуальная в настоящее время работа «К вопросу о пригодности найденных статистическим путем формул прогноза» (1933) (см. [2, с. 161-167]). А в 1970-х годах он стимулировал исследования по т.н. «асимптотике растущей размерности» (в современной терминологии – асимптотике Колмогорова)

$$n \to \infty,\ p \to \infty,\ \frac{p}{n} \to y > 0 \qquad (1)$$

при некотором положительном числе y, где p - число параметров, n – объем выборки. Эта асимптотика весьма актуальна как для многомерного статистического анализа, так и для статистики нечисловых данных [3], а также для задач статистического приемочного контроля [4] и анализа социологических данных (см., например, [5, гл. 13]).

В ответ на предложение А.Н. Колмогорова развернулись исследования по асимптотике растущей размерности. Важные результаты получили Л.В. Архаров, Д.А. Барсов, А.Д. Деев, В.И. Заруцкий, Л. Г.Малиновский, Л.Д. Мешалкин, В.И. Сердобольский и др. (обзор работ по этой тематике дан в [6]).

При практическом применении результатов, полученных в асимптотике растущей размерности, возникает вопрос – откуда взять значение y, предельное для p/n. Обычно в полученные формулы вместо y подставляют p/n. Поэтому в настоящее время в асимптотике Колмогорова от третьего предельного перехода в асимптотике (1) обычно отказываются, но взамен требуют отделенности дроби p/n от 0 и $+\infty$, т.е. требуют существования числа $\varepsilon > 0$ такого, что двойной предельный переход осуществляется в условиях

$$p \to \infty,\ n \to \infty,\ p/n > \varepsilon,\ p/n < 1/\varepsilon.$$

Статистические задачи, в которых величиной p/n пренебречь нельзя, образуют важный, но пока еще недостаточно изученный класс задач, относящихся к математическим методам исследования [7]. Задачи этого класса удобно называть существенно многопараметрическими.

В настоящей работе (продолжающей доклад [8]):

а) обращается внимание на недостаточность традиционного подхода, на основе которого даются не всегда удовлетворительные рекомендации по решению статистических задач, отличающихся большим числом оцениваемых параметров;

б) даны предложения по включению некоторых процедур, разработанных для существенно многопараметрических задач, в программные продукты по статистическому анализу данных;

в) описаны эффекты, возникающие при оценивании большого числа параметров, и возможные подходы к решению многопараметрических задач прикладной статистики.

Особенности задач многопараметрического оценивания

Современные статистические многомерные данные, подготовленные для обработки, представляют собой матрицы, в которых заданы значения показателей исследуемых объектов (наблюдений, статистических единиц). Число показателей часто бывает большим (десятки, сотни, тысячи) и сравнимым (или даже превосходящим на порядки) с объемом выборки. Обычная трудность, с которой сталкивается исследователь, применяющий методы многомерного статистического анализа, состоит в неустойчивости или даже невозможности операции обращения выборочной ковариационной матрицы (при $p > n$ её определитель равен 0). В тех случаях, когда результаты операции обращения всё же удается получить, они зачастую оказываются статистически незначимыми.

Практика пошла по пути искусственного снижения размерности статистической задачи, отказа от обработки части информации, отбора показателей или проекции тем или иным способом в пространство меньшей размерности. Включение в вероятностно-статистическую модель некоторых показателей может (при применении распространенных методов многомерного статистического анализа) даже ухудшить качество статистических решений. В подобной ситуации специалисту прикладной области обычно рекомендуется подбирать лучшие комбинации факторов тем или иным образом. Этой же цели служат, в частности, различные пошаговые процедуры.

Причина неудач в применении ранее хорошо зарекомендовавших себя методов к новым задачам состоит в необоснованной экстраполяции области применимости традиционной асимптотики $p = const,\ n \to \infty$ и найденных в ней оптимальных решений. При разработке статистических процедур, как правило, сначала ставится и решается теоретич5еская задача, отыскивается наилучшее, но неизвестное исследователю решение. Затем в этом идеальном решении параметры заменяются их состоятельными оценками. Известны недостатки этого приема. Состоятельность, являющаяся, конечно, желательным свойством статистической оценки, никак не связана со свойствами оценки при фиксированном объеме выборки. Это особенно важно для многомерных и многопараметрических моделей. Обычные статистические процедуры оказываются, как правило, состоятельными неравномерно по p. Накопление погрешностей оценок приводит к смещениям порядка p/n [6], а иногда и p^2/n [9].

Некоторые специфические черты многопараметрических задач можно увидеть уже на следующем простом примере [8]. Пусть $X = (X_1, X_2, ..., X_p)$ - случайный вектор, описывающий наблюдения, имеющий многомерное нормальное распределение с математическим ожиданием μ и единичной ковариационной матрицей. Требуется оценить величину μ^2 (здесь и далее квадраты векторов означают квадраты их длин). Рассмотрим выборку из n независимых случайных векторов, распределенных одинаково с X. Пусть $\overline{X}$ - вектор, являющийся средним арифметическим элементов этой выборки. Можно показать, что математическое ожидание и дисперсия квадрата длины вектора $\overline{X}$ равны

$$M\left(\overline{X}^2\right) = \mu^2 + \frac{p}{n},\ D\left(\overline{X}^2\right) = \frac{4\mu^2}{n} + \frac{2p}{n^2}.$$

Из этих соотношений видно, что «естественная» оценка $\overline{X}^2$ величины μ^2

является асимптотически несмещенной и состоятельной при безграничном росте объема выборки, однако эти свойства неравномерны по p. По сравнению с ней оценка

$$\max\left(0,\ \overline{X}^2 - \frac{p}{n}\right)$$

обладает некоторыми преимуществами, а именно, при выполнении ограничений

$$\frac{p}{n} < c_1,\ \mu^2 < c_2$$

для некоторых c_1 и c_2 она является равномерно состоятельной и равномерно асимптотически несмещенной при $n \to \infty$. Эти свойства статистических оценок характерны для многопараметрического оценивания, при котором скалярные функции могут иметь значительные смещения при малой дисперсии.

При рассмотрении многопараметрических задач оценивания исследователь изучает суммарное влияние большого числа параметров, от каждого их которых рассматриваемые статистики зависят слабо. Влияние небольшого количества более важных параметров часто можно выделить и исследовать традиционными методами, в которых число параметров не зависит от объема выборки.

В общем случае в параметрическом пространстве выделяется точка, в окрестности которой локализуется вектор параметров. В статистической постановке задается семейство $\{F_p(x,\theta)\}$ функций распределения рассматриваемой случайной величины, зависящих от параметра $\theta \in R^p$, где размерность параметрического пространства $p \to \infty$, но длина параметра ограничена некоторым число c, т.е. $|\theta| < c$, где c от p не зависит [6, 10]. К такой постановке сводятся задачи дискриминантного и регрессионного анализа, когда вектор наблюдений задается растущим числом случайных компонент. Имеются в виду задачи, в которых требуется найти единичный

вектор, определяющий наилучшую разделяющую плоскость или вектор наилучших коэффициентов регрессии, длина которого априори ограничена.

Описание множества параметров и их оценок удобно задавать в виде функций распределения. Пусть $\theta = (\theta_1, ..., \theta_p)$ - вектор параметров, $\hat{\theta} = (\hat{\theta}_1, ..., \hat{\theta}_p)$ - вектор оценок θ по выборке n. Положим

$$F_{op}(t) = \frac{1}{p}\sum_{i=1}^{n} U(t - \sqrt{p}\theta_i), \quad F_p(t) = \frac{1}{p}\sum_{i=1}^{n} U(t - \sqrt{p}\hat{\theta}_i),$$

где $U(t) = 1$ при $t \geq 0$ и $U(t) = 0$ при $t < 0$, а множители $\sqrt{p}$ введены для удобства. Если компоненты (координаты) $\hat{\theta}$ попарно независимы, то

$$D(F_p(t)) \leq \frac{1}{2p}.$$

Для широкого класса оценок при $p \to \infty$ и $n \to \infty$ имеем

$$\sup_t D(F_p(t)) \to 0.$$

Такой эффект неслучайного поведения множества случайных величин в литературе назван «самоусреднением».

Пусть

$$|\theta| < c, \ n \to \infty, \ \frac{p}{n} \to y > 0, \ D(\theta_i) = \frac{1}{n}, \ i = 1, ..., p, \ F_{0p}(t) \to F_0(t)$$

равномерно по t, оценки $\hat{\theta}_i$ равномерно асимптотически нормальны и равномерно асимптотически попарно независимы. Тогда при $y > 0$

$$F_p(t) \to F(t) = \int \Phi\left(\frac{t - v}{\sqrt{y}}\right) dF_0(v)$$

(сходимость по вероятности), где $\Phi(z)$ - функция стандартного нормального распределения с нулевым математическим ожиданием и единичной дисперсией.

Аналогичные интегральные связи в асимптотике (1) возникают и в других задачах оценивания. Значение этих связей для статистического

анализа состоит в том, что они дают информацию о множественных характеристиках оцениваемых параметров (например, в виде функции $F_{0p}(t)$). Эта информация возникает как следствие априорного предположения $|\theta|<c$. Предельное интегральное уравнение относительно $F_0(t)$ разрешается однозначно. Но при подстановке оценок возникают трудности в виде некорректных задач.

Случайные матрицы растущей размерности

Развитие многопараметрических методов статистического анализа опирается на свойства выборочных характеристик случайных матриц растущей размерности. Теория таких матриц разработана В.Л. Гирко [11] и другими авторами. Выборочные ковариационные матрицы близки по свойствам к матрицам, изученным в этих работах, и на них можно распространить выводы, полученные в [11].

Пусть $X \in R^p$ - вектор наблюдений, $\Sigma = \mathrm{cov}(X, X)$ – его ковариационная матрица порядка $p \times p$, а C – выборочная ковариационная матрица. Положим

$$h_p(z) = p^{-1} tr(I - zC)^{-1},\ F_{0p}(t) = p^{-1}\sum_{i=1}^{p} U(t - \lambda_i(\Sigma)),\ F_{0p}(t) = p^{-1}\sum_{i=1}^{p} U(t - \lambda_i(C)),$$

где $\lambda_i(\Sigma)$, $\lambda_i(C)$ - соответствующие собственные числа матриц. В [11] установлено, что если при каждом p существует система координат (пусть неизвестная наблюдателю), в которой компоненты X независимы, выполняется некоторое условие типа условия Линдеберга в центральной предельной теореме теории вероятностей, справедливы соотношения (1) и

$$F_{0p}(t) \to F_0(t),$$

при каждом $t \geq 0$, то

$$F_p(t) \to F(t)$$

почти наверное и при мнимых z

$$h_p(z) \to h(z)$$

почти наверное.

Справедливо простое нелинейное уравнение

$$h(z) = \int (1 - zs(z)u)^{-1} dF_0(u), \;\; s(z) = 1 + y(h(z) - 1), \qquad (2)$$

которое однозначно связывает предельные спектры матриц Σ и C. Этот факт означает, что в асимптотике (1) влияние деталей распределения X усредняется и инвариантные относительно вращений функции C сходятся к функциям только от Σ, причем того же вида, что и для нормально распределенных X. Процедуры, инвариантные относительно вращений, оказываются асимптотически свободными от предположений о распределениях. Для $\Sigma = I$ плотность $F'(x)$ пропорциональна

$$u^{-1}\sqrt{(u_2 - u)(u - u_1)}, \;\; u_{1,2} = (1 \pm \sqrt{y})^2.$$

Обратим внимание на существенность учета эффектов оценивания многих параметров даже при небольшой размерности. Например, при p = 5 и n = 500 спектр матриц C «размыт» на отрезке значительной длины 0,4.

Некоторые применения

При решении многопараметрических задач получен ряд результатов, которые могут быть рекомендованы для использования в приложениях.

Рассмотрим задачу уменьшения квадратичного риска оценивания векторов математических ожиданий векторов $X \in R^p$ с независимыми компонентами. Отметим, что для известной [12] оценки Стейна $\hat{\mu}^S$ при p > 2 для любых n и $\mu = M(X)$ справедливо неравенство

$$M(\mu - \hat{\mu}^S)^2 < \inf_{\hat{\mu} \in K} M(\mu - \hat{\mu})^2 + \frac{4}{n},$$

где K – класс оценок вида $\hat{\mu} = \eta \bar{X}$ с неслучайным η. Таким образом, оценки Стейна асимптотически доминируют K. В [6] найдено семейство оценок векторов $\mu = M(X)$ ограниченной длины, асимптотически доминирующее

класс оценок, построенных путем взвешивания компонент $\bar{X}$, а также развит асимптотически экстремальный подход к дискриминантному анализу наблюдений из сближающихся совокупностей с независимыми компонентами. Дискриминантная функция имеет вид

$$g(x) = \sum_{i=1}^{p} \rho_i \ln \frac{f_i(x_i, \hat{\theta}_{1i})}{f_i(x_i, \hat{\theta}_{2i})},$$

где $f_i(x_i, \hat{\theta}_i)$ - плотности из некоторого регулярного семейства, $\hat{\theta}_{1i}$ и $\hat{\theta}_{2i}$ – оценки компонент $i = 1,\ldots p$ векторов θ_1 и θ_2, задающих совокупности 1 и 2. Обозначим J_i и $\hat{J}_i$ вклад и оценку вклада переменной X_i в расстояние Кульбака между совокупностями. Пусть

$$\max_i J_i = O(p^{-1})$$

(условие равномерного сближения) и рассматриваются коэффициенты взвешивания $\rho_i = \eta(n\hat{J}_i / 2)$, где $\eta(t)$ - функции из некоторого широкого класса. В [6] в асимптотике растущей размерности (1) найдены пределы для вероятностей ошибок в виде функционалов от $\eta(t)$, решены экстремальные задачи и построены оценки асимптотически экстремальной функции $\eta(t) = \eta^0(t)$.

Процедура асимптотически экстремального дискриминантного анализа

Одним из серьезных достижений в развитии многопараметрических методов статистического анализа можно считать разработку процедуры асимптотически экстремального дискриминантного анализа [6]. Диссертация В.С. Степанова [13] посвящена ее изучению. Программная реализация этой процедуры оформлена в виде алгоритма «ЭЛДА».

Согласно процедуре рассчитывают регуляризованную оценку $\Gamma(C)$ обратной ковариационной матрицы Σ^{-1}, которая обеспечивает минимум предельной вероятности ошибки классификации в асимптотике (1) с $y < 1$.

Дискриминантная функция имеет вид

$$g(x) = (\bar{x}_1 - \bar{x}_2)^T \Gamma(C)(x - (\bar{x}_1 + \bar{x}_2)/2),$$

где $\bar{x}_1$ и $\bar{x}_2$ - вектора выборочных средних для двух совокупностей. Матрица $\Gamma(C)$ диагонализируется вместе с C, при этом величины $\Gamma(\lambda_i)$ стоят на главной диагонали, где λ_i - собственные числа C. Процедура сводится к асимптотически экстремальному взвешиванию с весами $\Gamma(\lambda_i)$ компонент вектора наблюдений в системе координат, в которой C диагональна. Наилучшее предельное значение порога классификации также зависит от матрицы $\Gamma(C)$.

Описанный алгоритм испытывался методом Монте-Карло в сравнении с другими алгоритмами статистической классификации в системе сравнения алгоритмов COPPA [9]. Статистические испытания показали преимущество процедуры ЭЛДА, особенно при больших p, сравнимых с n. Достигнутый в испытаниях выигрыш в уменьшении вероятности ошибки объясняется, в первую очередь, эффектом регуляризации оценки Σ^{-1}, и во вторую – подавлением вкладов тех компонент X, которые вносят в дискриминацию (т.е. в разность ожидаемых значений дискриминантной функции $g(x)$ при x из одной и из другой совокупности) незначимые вклады, но заметно увеличивают дисперсию дискриминантной функции $g(x)$.

Экстремальные свойства процедуры ЭЛДА могут быть установлены в условиях применимости спектрально теории выборочных ковариационных матриц, сформулированных в [6]. Поэтому они не ограничены предположениями о нормальности, сделанными в [13]. От распределений результатов наблюдений (элементов выборок) требуется выполнение некоторых условий регулярности. Необходимо, чтобы существовала система координат, в которой компоненты векторов наблюдений были бы независимы, дисперсии компонент ограничены

снизу, а их центральные моменты – сверху. Еще одно условие ограничивает расстояние Махаланобиса между совокупностями не очень большими значениями. Расстояние между совокупностями должно быть таким, при котором задача статистической классификации имеет содержательный смысл, т.е. не вырождается. Асимптотическая вероятность ошибки классификации и эффективность процедуры ЭЛДА зависят только от математических ожиданий $M(X_v)$ и ковариационных матриц $M[(X - M(X_v))(X - M(X_v))^T]$ для X из совокупностей $v = 1, 2$, и для распределений из широких классов имеют тот же вид, что и для нормальных распределений.

Случай зависимых компонент

В [6] решена также задача построения оценок для совокупностей с зависимыми компонентами X в предположениях применимости спектральной теории выборочных ковариационных матриц. Построено семейство оценок, асимптотически доминирующее в асимптотике (1) класс оценок вектора $\mu = M(X), |\mu| < const$ вида $\hat{\mu} = \Gamma(C)\bar{x}$, где $\Gamma(C)$ - матрицы из некоторого класса, диагонализирующиеся вместе с C. В этих оценках компоненты векторов выборочных средних в системе координат, в которой C диагональна, взвешиваются с асимптотически экстремальными весами, зависящими от собственных значений C. Выигрыш достигается за счет подавления вкладов компонент $\bar{x}$, незначимо отличающихся от 0.

В асимптотике (1) решена задача минимизации предельного значения величины

$$p^{-1}M(tr(\Sigma - \hat{\Sigma}(C))^2)$$

где $\hat{\Sigma}(C)$ - матрица, диагонализирующаяся вместе с C и представляющая собой оценку Σ из класса оценок, зависящих от произвольной функции ограниченного изменения. Решена задача построения (в классе

регуляризированных оценок матрицы Σ^{-1}) оценок $\Gamma^0(C)$, асимптотически доминирующих некоторый широкий класс оценок.

Место асимптотики Колмогорова в многообразии математических методов теории классификации обсуждается в [14].

Теория люсианов

Всё более важным становится класс многопараметрических статистических задач, в которых число показателей (компонент векторов наблюдений) настолько велико, что намного превосходит объем выборки. Такие «сверхпараметрические» задачи характерны для моделей технических и организационных систем, биологических, медицинских, социологических, экономических постановок, экспертных оценок и задач управления качеством продукции, моделирования процессов принятия решений (см., например, [15, 16]). Здесь статистические процедуры обычно могут быть разработаны лишь в условиях достаточно слабой зависимости или зависимости специального вида между компонентами векторов наблюдений.

В качестве примера рассмотрим теорию люсианов – один из разделов статистики нечисловых данных (статистики объектов нечисловой природы, нечисловой статистики). Люсиан (X_1, X_2, …, X_p) – это конечная последовательность независимых испытаний Бернулли с, вообще говоря, различными вероятностями успеха (т.е. совпадение вероятностей успеха для различных элементов этой конечной последовательности не предполагается). Распределение люсиана описывается вектором вероятностей успехов в испытаниях Бернулли ($P(X_1 = 1)$, $P(X_2 = 1)$, …, $P(X_p = 1)$), т.е. p параметрами. Люсианы возникают в различных экономических, технических, социологических, медицинских, психологических и других приложениях, обзор которых дан, например, в [16, 17].

Одна из типичных задач в теории люсианов – проверка гипотезы

согласованности n люсианов (X_{1i}, X_{2i}, …, X_{pi},), i = 1, 2, …, n, между собой, т.е. гипотезы

$$H_0 : P(X_{j1} = 1) = P(X_{j2} = 1) = \ldots = P(X_{jn} = 1), j = 1, 2, \ldots, p. \quad (3)$$

При справедливости H_0 вероятностная модель описывается p параметрами.

Метод проверки гипотез по совокупности малых выборок, т.е. в асимптотике

$$n = \text{const}, \ p \to \infty,$$

предложенный в [18, разд.4.5], основан на использовании несмещенных оценок нуля, несмещенных оценок дисперсий этих оценок и введенного в [18, разд.4.3] с помощью некоторой системы аксиом расстояния (точнее, псевдометрики) между множествами

$$d(A, B) = \mu(A \Delta B), \quad (4)$$

где $A \Delta B$ - симметрическая разность множеств A и B, т.е.

$$A \Delta B = (A \setminus B) \bigcup (B \setminus A),$$

а μ - некоторая мера. (При использовании конечных множеств - примером меры служит число элементов подмножества.) С помощью этого метода при $n \geq 4$ построены критерии проверки гипотезы (3) согласованности люсианов, а также гипотез однородности и независимости люсианов. Получены предельные распределения соответствующих статистик при нулевых гипотезах и при альтернативах, изучена мощность критериев [16, 17, 19].

Скорость сходимости распределений статистик к предельным нормальным распределениям изучалась методом Монте-Карло [16]. Оказалось, что отклонение от предельного распределения меньше ошибки метода Монте-Карло (при 10 000 испытаниях) уже при p = 15 – 30. Получены также оценки скорости сходимости типа неравенств Берри-Эссеена [20].

В теории люсианов рассмотрены также задачи классификации, для

которых получено статистически обоснованное ограничение на размер кластера, а также задачи оценивания среднего в пространстве бинарных отношений, интерпретируемые как задачи агрегирования мнений экспертов [15, 21].

В [22] на основе некоторой системы аксиом, отличной от приводящей к соотношению (4), получено расстояние (псевдометрика)

$$D(A,B)=\frac{\mu(A\Delta B)}{\mu(A\cap B)} \qquad (5)$$

между множествами A и B. Ведется построение предельной теории на основе расстояния (5), «параллельной» (т.е. решающей те же задачи) теории, ранее развитой на основе псевдометрики (4). В частности, доказана асимптотическая нормальность расстояния (5) между независимыми люсианами.

В пакеты программ прикладного статистического анализа целесообразно включать следующие программы, реализованные на ЭВМ и испытанные нами [16, 17]:

- проверка согласованности $n \geq 4$ люсианов;
- проверка однородности двух групп люсианов;
- проверка независимости люсианов.

Методы теории люсианов успешно применялись в целом ряде приложений: для проведения и управления научными медицинскими исследованиями, в частности, при анализе кинетокардиограмм [5, гл. 11], при разработке и принятии управленческих решений на основе экспертных оценок [21], для управления качеством продукции, в частности, в задачах статистического приемочного контроля [4], для анализа данных психологических тестов типа MMPI и т.д. [16].

Заключение

К настоящему времени решения, учитывающие влияние

существенно большого числа параметров, найдены лишь для некоторого числа из наиболее распространенных задач прикладной статистики. С точки зрения многопараметрического подхода недостаточно анализировались особенности решения задач регрессионного, дисперсионного и факторного анализов, метода главных компонент, многомерного шкалирования и оптимального проецирования, кластерного анализа. Недостаточно исследовано влияние большого числа параметров на методы проверки гипотез [23].

Не используются сколько-нибудь широко уже найденные асимптотические и асимптотически экстремальные решения. Не разработаны экспертные системы (системы искусственного интеллекта), которые могли бы подсказать пользователю выбор подходящего метода решения трудных статистических задач анализа многомерных данных и построения многопараметрических моделей.

Внедрение и обмен уже созданными алгоритмами затрудняется из-за отсутствия нужной информации и из-за общей неудовлетворительной организации накопления и распространения алгоритмов. Здесь следует отметить недостаточное обеспечение авторского права на алгоритмы, отсутствие должного приемочного контроля (сертификации), а также возможностей и средств распространения программной продукции.

Из сказанного ясно, что необходима дальнейшая разработка статистической теории в асимптотике Колмогорова, создание соответствующего методического и прикладного обеспечения, широкое внедрение уже полученных научных результатов на основе адекватного решения организационных вопросов управления научно-техническим прогрессом в рассматриваемой весьма перспективной области прикладной статистики.

Литература

1. Орлов А.И. Вероятностно-статистические методы в работах А.Н. Колмогорова // Научный журнал КубГАУ. 2014. №98. С. 158–180.

2. Колмогоров А.Н. Теория вероятностей и математическая статистика. – М.: Наука, 1986. - 535 с.

3. Орлов А.И. Искусственный интеллект: нечисловая статистика. — М.: Ай Пи Ар Медиа, 2022. — 446 с.

4. Орлов А.И. Метод проверки гипотез по совокупности малых выборок и его применение в теории статистического контроля // Научный журнал КубГАУ. 2014. №104. С. 38–52.

5. Орлов А.И. Искусственный интеллект: статистические методы анализа данных. — М.: Ай Пи Ар Медиа, 2022. — 843 с.

6. Сердобольский В.И. Асимптотическая теория статистического анализа наблюдений высокой размерности : диссертация ... доктора физико-математических наук / Моск. гос. ун-т им. М. В. Ломоносова. Фак. вычислит. математики и кибернетики — М.: 2000. — 245 с.

7. Загоруйко Н.Г., Орлов А.И. Некоторые нерешенные математические задачи прикладной статистики // Современные проблемы кибернетики (прикладная статистика). - М.: Знание, 1981. - С. 53-63.

8. Орлов А.И., Сердобольский В.И. Статистический анализ при большом числе параметров // Тезисы докладов III Всесоюзной школы-семинара «Программно-алгоритмическое обеспечение прикладного многомерного статистического анализа». - М.: ЦЭМИ АН СССР, 1987.- С. 151-160.

9. Раудис Ш.Ю. Статистическая классификация при существенно ограниченных выборках: автореферат дисс. … доктора технических наук / АН ЛатвССР. Ин-т электрон. и вычислит. техники. – Рига, 1978. – 39 с.

10. Сердобольский В.И. Теория существенно многомерного статистического анализа // Успехи математических наук. 1999. Т.54. Вып.2. С. 85–112.

11. Гирко В.Л. Спектральная теория случайных матриц. – М.: Наука, 1988. - 375 с.

12. Стейн Ч. Лекции по теории оценивания многих параметров // Исследования по статистической теории оценивания. Том I. Зап. научн. сем. ЛОМИ. 1977. № 74. С. 4–65.

13. Степанов В.С. Некоторые задачи дискриминантного анализа наблюдений большой размерности : автореферат дис. ... кандидата физико-математических наук / МГУ им. М. В. Ломоносова. фак. вычислит. математики и кибернетики. – М.: 1987.- 16 с.

14. Орлов А.И. Математические методы теории классификации // Научный журнал КубГАУ. 2014. №95. С. 423–459.

15. Орлов А.И. Парные сравнения в асимптотике Колмогорова // Экспертные оценки в задачах управления. - М.: Изд-во Института проблем управления АН СССР, 1982 - С. 58-66.

16. Рыданова Г.В. Некоторые вопросы статистического анализа случайных бинарных векторов. Дисс. ... канд. физ.-мат. наук. - М.: МГУ, ф-т вычислительной математики и кибернетики, 1987. - 139 с.

17. Орлов А.И. Теория люсианов // Научный журнал КубГАУ. 2014. №101. С. 275–304.

18. Орлов А.И. Устойчивость в социально-экономических моделях. - М.: Наука,1979. - 296 с.

19. Орлов А.И. Случайные множества с независимыми элементами (люсианы) и их применения // Алгоритмическое и программное обеспечение прикладного

статистического анализа. Ученые записки по статистике. Т.36. - М.: Наука, 1980. С. 287-308.

20. Дылько Т.Н. Выбор коэффициентов статистики для проверки гипотезы одинаковой распределенности в задачах анализа дихотомической информации // Вестник Белорусского государственного университета / Сер. 1: Физика, математика и механика. 1988. № 2. С. 36–40.

21. Орлов А.И. Искусственный интеллект: экспертные оценки. — М.: Ай Пи Ар Медиа, 2022. — 436 с.

22. Орлов А.И. Расстояния в пространствах статистических данных // Научный журнал КубГАУ. 2014. №101. С. 227–252.

23. Сердобольский В.И. Многопараметрические методы - новое направление в статистике // Статистические методы оценивания и проверки гипотез: межвузовский сборник научных трудов. Вып.19. – Пермь: Пермский государственный университет. 2006. - С.188-203.

24. Орлов А.И. Прикладной статистический анализ. — М.: Ай Пи Ар Медиа, 2022. — 812 с.

References

1. Orlov A.I. Veroyatnostno-statisticheskie metody v rabotah A.N. Kolmogorova // Nauchnyj zhurnal KubGAU. 2014. №98. S. 158–180.

2. Kolmogorov A.N. Teoriya veroyatnostej i matematicheskaya statistika. – M.: Nauka, 1986. - 535 s.

3. Orlov A.I. Iskusstvennyj intellekt: nechislovaya statistika. — M.: Aj Pi Ar Media, 2022. — 446 c.

4. Orlov A.I. Metod proverki gipotez po sovokupnosti malyh vyborok i ego primenenie v teorii statisticheskogo kontrolya // Nauchnyj zhurnal KubGAU. 2014. №104. S. 38–52.

5. Orlov A.I. Iskusstvennyj intellekt: statisticheskie metody analiza dannyh. — M.: Aj Pi Ar Media, 2022. — 843 c.

6. Serdobol'skij V.I. Asimptoticheskaya teoriya statisticheskogo analiza nablyudenij vysokoj razmernosti : dissertaciya ... doktora fiziko-matematicheskih nauk / Mosk. gos. un-t im. M. V. Lomonosova. Fak. vychislit. matematiki i kibernetiki — M.: 2000. — 245 s.

7. Zagorujko N.G., Orlov A.I. Nekotorye nereshennye matematicheskie zadachi prikladnoj statistiki // Sovremennye problemy kibernetiki (prikladnaya statistika). - M.: Znanie, 1981. - S. 53-63.

8. Orlov A.I., Serdobol'skij V.I. Statisticheskij analiz pri bol'shom chisle parametrov // Tezisy dokladov III Vsesoyuznoj shkoly-seminara «Programmno-algoritmicheskoe obespechenie prikladnogo mnogomernogo statisticheskogo analiza». - M.: CEMI AN SSSR, 1987.- S. 151-160.

9. Raudis SH.YU. Statisticheskaya klassifikaciya pri sushchestvenno ogranichennyh vyborkah: avtoreferat diss. … doktora tekhnicheskih nauk / AN LatvSSR. In-t elektron. i vychislit. tekhniki. – Riga, 1978. – 39 s.

10. Serdobol'skij V.I. Teoriya sushchestvenno mnogomernogo statisticheskogo analiza // Uspekhi matematicheskih nauk. 1999. T.54. Vyp.2. C. 85–112.

11. Girko V.L. Spektral'naya teoriya sluchajnyh matric. – M.: Nauka, 1988. - 375 s.

12. Stejn CH. Lekcii po teorii ocenivaniya mnogih parametrov // Issledovaniya po statisticheskoj teorii ocenivaniya. Tom I. Zap. nauchn. sem. LOMI. 1977. № 74. S. 4–65.

13. Stepanov V.S. Nekotorye zadachi diskriminantnogo analiza nablyudenij bol'shoj razmernosti : avtoreferat dis. ... kandidata fiziko-matematicheskih nauk / MGU im. M. V. Lomonosova. fak. vychislit. matematiki i kibernetiki. – M.: 1987.- 16 s.

14. Orlov A.I. Matematicheskie metody teorii klassifikacii // Nauchnyj zhurnal KubGAU. 2014. №95. S. 423–459.

15. Orlov A.I. Parnye sravneniya v asimptotike Kolmogorova // Ekspertnye ocenki v zadachah upravleniya. - M.: Izd-vo Instituta problem upravleniya AN SSSR, 1982 - S. 58-66.

16. Rydanova G.V. Nekotorye voprosy statisticheskogo analiza sluchajnyh binarnyh vektorov. Diss. ... kand. fiz.-mat. nauk. - M.: MGU, f-t vychislitel'noj matematiki i kibernetiki, 1987. - 139 s.

17. Orlov A.I. Teoriya lyusianov // Nauchnyj zhurnal KubGAU. 2014. №101. S. 275–304.

18. Orlov A.I. Ustojchivost' v social'no-ekonomicheskih modelyah. - M.: Nauka,1979. - 296 s.

19. Orlov A.I. Sluchajnye mnozhestva s nezavisimymi elementami (lyusiany) i ih primeneniya // Algoritmicheskoe i programmnoe obespechenie prikladnogo statisticheskogo analiza. Uchenye zapiski po statistike. T.36. - M.: Nauka, 1980. S. 287-308.

20. Dyl'ko T.N. Vybor koefficientov statistiki dlya proverki gipotezy odinakovoj raspredelennosti v zadachah analiza dihotomicheskoj informacii // Vestnik Belorusskogo gosudarstvennogo universiteta / Ser. 1: Fizika, matematika i mekhanika. 1988. № 2. S. 36–40.

21. Orlov A.I. Iskusstvennyj intellekt: ekspertnye ocenki. — M.: Aj Pi Ar Media, 2022. — 436 c.

22. Orlov A.I. Rasstoyaniya v prostranstvah statisticheskih dannyh // Nauchnyj zhurnal KubGAU. 2014. №101. S. 227–252.

23. Serdobol'skij V.I. Mnogoparametricheskie metody - novoe napravlenie v statistike // Statisticheskie metody ocenivaniya i proverki gipotez: mezhvuzovskij sbornik nauchnyh trudov. Vyp.19. – Perm': Permskij gosudarstvennyj universitet. 2006. - S.188-203.

24. Orlov A.I. Prikladnoj statisticheskij analiz. — M.: Aj Pi Ar Media, 2022. — 812 c.