ЭКОНОМИКА
И МАТЕМАТИЧЕСКИЕ МЕТОДЫ
1990
Том 26, вып. 2

ПАРАДИГМЫ СОВРЕМЕННОЙ МАТЕМАТИЧЕСКОЙ СТАТИСТИКИ

Миркин Б. Г.

(Москва)

Для некоторых из основных математико-статистических понятий показано, что они допускают три способа интерпретации: теоретико-вероятностный, модельно-аппроксимационный и эвристико-алгоритмический, каждый со своей системой представлений и задач. Делаются определенные методологические и дидактические выводы.

Хорошо известно, что термин «статистика» допускает несколько взаимосвязанных, но различных толкований, из которых важнейшими являются следующие два. Статистика — это: а) информация о значениях показателей, характеризующих те или иные аспекты функционирования социально-экономической системы или ее элементов в определенный период; б) наука о способах обработки подобной информации с целью анализа, объяснения, прогноза или управления. Развитие статистики в обоих указанных аспектах немыслимо без применения математических моделей и методов, однако математические конструкции, разрабатываемые и используемые в них, могут сильно различаться из-за специфики возникающих задач.

Основные проблемы, которые нужно решать в связи с «информационным» толкованием статистики,— это разработка системы статистических показателей и процедур их измерения. По-видимому, такие разработки должны опираться на теоретические представления о функционировании того конкретного социально-экономического объекта (народного хозяйства, региона, отрасли, трудового коллектива и т. п.), для которого собирается информация. Статистические показатели призваны отражать соответствующие теоретические категории, а процедуры измерения — смысл этих категорий. Подлинная математизация данного направления поэтому включает математизацию теоретических представлений, т. е. разработку адекватной математической модели социально-экономического объекта и математических методов оценивания значений переменных этой модели. Указанный инструментарий, по-видимому, является необходимым и достаточным для решения задачи формирования и измерения системы статистических показателей (мы отвлекаемся здесь от проблем организационного обеспечения). Методология оценки параметров (коэффициентов) модели на базе реальной информации в значительной степени должна покрывать проблематику второго, «методологического» толкования термина «статистика».

К сожалению, состояние социально-экономических исследований далеко от описанного. Обычно мы не имеем ни сколько-нибудь полной системы категорий, ни однозначного понимания их смысла и взаимосвязей (достаточно вспомнить такие казалось бы четкие категории, как «национальный доход» или «прибыль»). В этих условиях использование и разработка математико-статистических моделей и методов приобретают ряд особенностей, которые до настоящего времени не систематизированы и не исследованы. В данной статье рассматриваются некоторые интенсивно развиваемые направления прикладной и теоретической статистики в социально-экономических исследованиях с точки зрения сосуществования в них различных интерпретаций одних и тех же понятий. Это позволяет уточнить соотношение между такими традиционными на-

правлениями математической статистики, как оценка среднего значения статистического показателя, и такими нетрадиционными, как кластерный анализ многомерных данных смешанной природы, а также сделать некоторые выводы.

1. ТРИ ПАРАДИГМЫ В ЗАДАЧАХ ОЦЕНКИ ПАРАМЕТРОВ

Значительное количество статистических задач связано с оценкой тех или иных «теоретических» параметров по данным наблюдений. Такими параметрами являются характеристики средних значений и разброса статистических показателей, функциональных зависимостей между ними, «внутренние» факторы и т. п. Для оценки этих параметров разрабатываются и используются методы, которые могут рассматриваться как реализации по крайней мере одного из нижеследующих подходов.

1. Эвристико-алгоритмический подход. Аналитическая модель связи имеющихся данных и оцениваемых параметров либо отсутствует вовсе, либо не принимается во внимание. В качестве модели берется такой алгоритм формирования параметров, каждый шаг которого достаточно ясно с интуитивной, геометрической или какой-либо иной точки зрения ориентирован на получение искомых характеристик. Результат рассматривается как искомая оценка.

Указанные алгоритмические вычисления могут использовать достаточно продвинутые математические понятия и конструкции, оптимизационные и иные построения, что не меняет их эвристического характера в силу того, что они не основаны на явной модели, связывающей оцениваемые теоретические величины с наблюденными данными.

2. Модельно-аппроксимационный подход. Имеется модель, увязывающая наблюденные данные X с оцениваемыми параметрами A зависимостью вида $X=F(A)$. Задача состоит в том, чтобы найти A по X. Количество неизвестных величин в массиве A должно быть не больше, чем число наблюденных значений в массиве X. Поэтому в общем случае никакие конкретные значения A не могут дать точного равенства $X=F(A)$, так что возникают «невязки» E, характеризующие отклонения от этого равенства, т. е. определяемые равенством $X=F(A)+E$.

Задача состоит в том, чтобы найти такой массив A, который минимизирует невязки E, причем является «допустимым», т. е. удовлетворяет априорным условиям на параметры типа $A\in D$, где D — множество допустимых A. Причины возникновения невязок (их три: неправильный выбор зависимости F или допустимого множества D, неточность или неполнота измерений X, стохастическая природа данных) при этом не анализируются и в модели не учитываются.

Проблема минимизации невязок E по своей сути носит многокритериальный характер, поскольку для каждого элемента данных может возникать своя ненулевая невязка. Это роднит проблемы оценки параметров с проблемами принятия решений. Однако в статистике принято скаляризовать проблему минимизации невязок с помощью использования конкретной числовой функции от E — «нормы» $\|E\|$. Традиционно используется квадрат евклидовой нормы (метод наименьших квадратов), а с недавнего времени также сити-блок метрика (метод наименьших модулей) и равномерная (чебышевская) норма.

Качество модели $X=F(A)$ с точки зрения ее соответствия наблюденным данным характеризуется величиной $\|E\|$. Особенно удачна эта характеристика в случае, когда справедливо соотношение вида

$$\|X\|=\|F(A)\|+\|E\|,$$

которое позволяет говорить об «объясненной» доле «разброса» данных $\|E\|$, равной $|F(A)\|/\|X\|$, и «необъясненной» доле, равной $\|E\|/\|X\|$. Указанные доли являются безразмерными величинами, заключенными между 0 и 1, что удобно для оценки качества модели.

Дальнейшая декомпозиция $\|F(A)\|$ по отдельным элементам модели (характерная, например, для линейных зависимостей) дает возможность

сравнительной оценки «вклада» этих элементов в «разброс» данных $\|X\|$ и, на этой основе, их сравнительной значимости.

3. Вероятностный подход. Согласно ему, с наблюдениями ассоциируются те или иные распределения вероятностей, которые образуют теоретическую модель явления. Данные интерпретируются как выборочные реализации соответствующих случайных величин. Указанные распределения задаются так, чтобы искомые результаты были параметрами этих распределений. Задача математической теории — формирование и обоснование алгоритмов оценки таких параметров по выборочным данным. Как известно, на этом пути разработана достаточно разветвленная и интересная система понятий (точечные и интервальные оценки, состоятельность и т. п.).

В последнее время вероятностный подход особенно часто используется как развитие аппроксимационных построений. А именно, в модели вида $X=F(A)+E$ величины E (и, возможно, часть A) предполагаются случайными с заданным (хотя и не обязательно известным) вероятностным распределением. Конечно, подобный переход к вероятностной парадигме осмыслен далеко не всегда. Однако и в тех случаях, когда с данными не может быть связан какой-либо естественный случайный механизм, наличие вероятностной интерпретации аппроксимационного алгоритма может рассматриваться как его полезное свойство.

Описанные три парадигмы задают три разные системы интерпретации одних и тех же параметров, методов их оценки и возникающих в связи с ними проблем. Различение этих систем важно, во-первых, с точки зрения дальнейшего развития теории статистики, во-вторых, для обучения, в-третьих, для приложений, в частности, при построении диалоговых систем обработки социально-экономической информации, а также для выработки рекомендаций об адекватном использовании результатов вычислений. Ниже рассмотрим, как преломляются данные парадигмы в задачах: оценки среднего значения; оценки связи (корреляции) двух переменных; многомерного факторного анализа; кластерного анализа, и сделаем соответствующие выводы.

2. ПРОБЛЕМА ОЦЕНКИ СРЕДНЕГО ЗНАЧЕНИЯ

Возьмем n чисел $x_1, \ldots, x_n$ и поставим вопрос об оценке усредненной величины a, которая бы в некотором смысле являлась их «агрегированным» представителем. Эта проблема давно изучена и в аппроксимационной, и в вероятностной парадигмах, так что эвристико-алгоритмический подход здесь выглядит анахронизмом.

Тем не менее для показа общности возникающих конструкций и проблем рассмотрим, например, следующую эвристическую процедуру. Найдем крайние (минимальное и максимальное) значения x_i, равные $x_{\min}$ и $x_{\max}$, после чего заменим эту пару их полусуммой $x=(x_{\min}+x_{\max})/2$. В результате останется $n-1$ величин x_i. Продолжая процесс замен, через $n-1$ шагов придем к единственной полусумме x, которую и примем за искомый параметр a.

С эвристической точки зрения процедура представляется вполне приемлемой, поскольку каждый ее шаг осмыслен как шаг в направлении требуемой цели. Вместе с тем оставаясь в рамках эвристико-алгоритмической парадигмы, нельзя решить вопрос об адекватности данного алгоритма, но можно провести математическое исследование в этом направлении. В частности, устанавливать какие-либо свойства данной процедуры, «полезные» с точки зрения наших представлений о «среднем» значении (например, в данном случае таким свойством является то, что x не выходит за область значений x_i). Причем эти свойства могут носить характер достаточно тонких необходимых и (или) достаточных условий (например, того, чтобы a являлось медианой).

В аппроксимационной парадигме проблема может ставиться как задача оценки величины a, удовлетворяющей равенствам

$$x_i=a+\varepsilon_i, \quad i=1, \ldots, n, \qquad (1)$$

с как можно меньшими величинами невязок ε_i в смысле, например, одной из «норм» вида

$$\sum_i \varepsilon_i^2, \quad \sum_i |\varepsilon_i|, \quad \max_i |\varepsilon_i|. \tag{2}$$

Хорошо известно, что при использовании первого из этих критериев оптимальной оценкой a является среднее арифметическое, второго — медиана, третьего — «полуразмах» $(x_{\min}+x_{\max})/2$.

Наряду с (1) можно в качестве модельного соотношения рассматривать также равенства $x_i/a=1+\varepsilon_i$ или $a/x_i=1+\varepsilon_i$, имеющие несколько иное истолкование и совсем другие аппроксимационные оценки среднего значения.

Важный вопрос математической теории, связанной с аппроксимационным решением,— возможность его эвристико-алгоритмической интерпретации. Ясно, что третий из критериев (2) применительно к (1) приводит к решению, даваемому первым шагом вышеизложенной процедуры «полусредних». Решение по второму критерию (медиана) в этом смысле значительно сложнее. Наиболее близко к эвристико-алгоритмическому решению среднее арифметическое, которое во многих случаях совпадает с методом «полусредних». Можно исследовать необходимые и достаточные условия этого совпадения. Однако еще более продуктивным является исследование вопроса о такой модификации метода «полусредних», при которой он всегда приводит к среднему арифметическому. Хорошо известно, что для этого достаточно учитывать весовые характеристики получаемых чисел.

Таким образом, наличие аппроксимационной парадигмы позволяет, во-первых, ставить и решать аппроксимационную задачу, во-вторых, интерпретировать аппроксимационные построения в эвристико-алгоритмических терминах, в-третьих, подбирать и соответствующим образом модифицировать существующие алгоритмы, предложенные на чисто эвристических основаниях.

Обратимся теперь к вероятностной парадигме. Среди имеющихся вероятностных моделей, пожалуй, наиболее распространенная состоит в том, чтобы считать числа x_i случайными независимыми реализациями одной и той же случайной величины ξ с плотностью $f(x)$. При этом все рассмотренные выше конструкции, связанные с моделью (1), очевидно, дают несмещенную оценку математического ожидания ξ. В зависимости от предполагаемого вида $f(x)$ можно говорить о более тонких свойствах оценок. Например, для нормального распределения, как известно, состоятельную оценку дает только среднее арифметическое. Для гиперболического же распределения математическое ожидание может не существовать вовсе; в этой ситуации можно говорить о бессмысленности самой постановки задачи отыскания среднего — это новый момент, привносимый именно вероятностной парадигмой.

Другая распространенная модель, которую можно рассматривать как частный случай первой,— наделение вероятностной структурой аппроксимационной модели (1). В традиционной теории измерений, в частности, полагают, что ε_i — независимые реализации нормально распределенной случайной величины с нулевым математическим ожиданием, тогда как a — неизвестный, но неслучайный параметр.

Вероятностная парадигма позволяет дать, во-первых, модельные основания выбору критерия аппроксимации, во-вторых, вероятностные оценки качества получаемых величин, в-третьих, основу проверки статистических гипотез о событиях, связанных с анализируемыми параметрами. Конечно, все это — лишь при условии справедливости вероятностной модели: вероятностный подход предполагает хорошее понимание природы рассматриваемых данных.

3. ПРОБЛЕМА ОЦЕНКИ СВЯЗИ ДВУХ ПЕРЕМЕННЫХ

Пусть (x_i, y_i) — результаты наблюдения двух переменных x и y на объектах $i \in I$. Традиционно в качестве мер связи переменных x, y используются такие характеристики, как (линейный) коэффициент корреляции (если x, y — количественные) и коэффициент сопряженности Пирсона (если x, y — качественные). Рассмотрим возможные интерпретации этих коэффициентов в терминах описанных трех парадигм.

Для количественных признаков x, y эвристико-алгоритмическая интерпретация коэффициента корреляции основана на их представлении в виде векторов $x=(x_i)$ и $y=(y_i)$ их значений в $|I|$-мерном евклидовом пространстве. Близость переменных оценивается близостью векторов. В частности, удобно использовать косинус угла

$$\rho = (x, y)/[(x, x)(y, y)]^{1/2} = \Sigma x_i y_i / \sqrt{\Sigma x_i \, \Sigma y_i}, \qquad (3)$$

показывающий степень их сонаправленности (коллинеарности) со всеми соответствующими тригонометрическими истолкованиями.

В качественном случае связь переменных x, y выражается таблицей сопряженности $P=(p_{mt})$, где p_{mt} — доля наблюдений, имеющих m-ю градацию признаков x и t-ю градацию y. Согласно методологии, разрабатывавшейся Г. Юлом и К. Пирсоном, коэффициент связи качественных переменных должен отражать уровень отличия наблюденного распределения p_{mt} от совместного распределения признаков при их статистической независимости $p_{m+}p_{+t}$, где $p_{m+} = \sum_t p_{mt}$, $p_{+t} = \sum_m p_{mt}$ — соответствующие маргинальные величины. При этом используется не абсолютное квадратическое отклонение $\Sigma (p_{mt} - p_{m+}p_{+t})^2$, и не относительное квадратическое отклонение $\Sigma (p_{mt}/p_{m+}p_{+t} - 1)^2$, а промежуточная характеристика — коэффициент сопряженности

$$\Phi^2 = \Sigma (p_{mt} - p_{m+}p_{+t})^2 / p_{m+}p_{+t}. \qquad (4)$$

Коэффициент Φ^2 «победил» благодаря своей связи с вероятностной парадигмой: величина $|I|\Phi^2$ имеет распределение χ^2 при стандартной гипотезе о том, что наблюденные значения x, y — это случайная независимая выборка из двумерного распределения при статистически независимых переменных.

В последнее время утвердилась (также в рамках эвристико-алгоритмической парадигмы) другая методология оценки связи качественных переменных — по тому, сколько информации дает знание градации одной переменной для прогнозирования градации второй. Эта методология позволяет сформулировать огромное количество различных показателей связи (см., например, серию статей Л. А. Гудмэна и У. Х. Крускэла [1]). Как показано в [2], коэффициент (4) также может быть получен с использованием подобных соображений. А именно, Φ^2 — среднее значение относительного прироста доли t при получении информаций о m, т. е.

$$\Phi(m, t) = (p_{mt}/p_{m+} - p_{+t})/p_{+t}.$$

Точнее,

$$\Phi^2 = \sum_{m,t} p_{mt} \Phi(m, t).$$

В аппроксимационной парадигме коэффициент корреляции характеризует качество линейного представления одной переменной через другую. Рассмотрим модель

$$y_i = ax_i + b + \varepsilon_i, \qquad (5)$$

где a и b — неизвестные параметры, оцениваемые из условия минимизации невязок ε_i. В частности, по методу наименьших квадратов (первый из критериев (2)), как известно, оптимальные оценки определяются формулами

$$a = \rho\sigma_y/\sigma_x, \quad b = \bar{y} - a\bar{x}, \qquad (6)$$

где $\bar{x}$, $\bar{y}$ — средние арифметические; σ_x, σ_y — средние квадратические отклонения от них, а ρ — коэффициент вида (3), рассчитанный для предварительно центрированных признаков x и y. Минимальное значение критерия наименьших квадратов (2) при этом, как известно, равно

$$\sum_i \varepsilon_i^2 = |I|(1-\rho^2)\sigma^2 y. \qquad (7)$$

Это определяет смысл коэффициента корреляции ρ в рамках аппроксимационной парадигмы. С одной стороны, знак ρ задает знак a, а с другой — ρ^2 показывает, какая доля разброса y объясняется его зависимостью (5) от x.

Любопытно, что подобным образом может быть интерпретирован и коэффициент Φ^2. Для этого следует только перейти к количественному представлению качественных признаков с помощью матриц связи между наблюдениями [2].

Вероятностная парадигма, блестяще разработанная К. Пирсоном, трактует коэффициент ρ как один из параметров двумерного нормального распределения, характеризующий степень «вытянутости» эллипсоидов рассеивания. Если наблюдения (x_i, y_i) рассматривать как независимые реализации двумерной нормальной случайной величины, то коэффициент (3) (при предварительно центрированных и нормированных x и y) является удачной оценкой этого параметра.

Часто берется также несколько другая, менее ограничительная схема, согласно которой значения y_i являются реализациями случайной величины $ax+b+\varepsilon$, причем x_i изменяются точно, a и b — неизвестны, а ε — случайная нормально распределенная ошибка.

Ясно, что обе эти вероятностные схемы пригодны только для количественных x и y. Для матричного представления признаков они не имеют смысла; выработать адекватные вероятностные представления здесь еще только предстоит.

Существует мнение, основанное на вероятностных представлениях, о том, что использование коэффициентов ρ и Φ^2 ограничено: Φ^2 может быть взят только для оценки значимости отклонения распределения качественных признаков от случая статистической независимости, а ρ осмыслен как коэффициент связи только для нормального распределения. Вышеизложенное показывает, что это не так. Эти коэффициенты имеют свой четкий смысл как коэффициенты связи и в других парадигмах. Более того, аппроксимационная парадигма позволяет дать обобщающую конструкцию, в которой обе эти величины являются специальными частными случаями матричного коэффициента (ковариации) [2].

4. ПРОБЛЕМА МНОГОМЕРНОГО ФАКТОРНОГО АНАЛИЗА

Модели и методы факторного анализа обсуждаются в [3—5]. Несмотря на 75-летнюю историю, состояние теории факторного анализа далеко от удовлетворительного.

Напомним, что проблема факторного анализа (в широком смысле этого понятия) состоит в построении некоторого числа неизмеримых «внутренних» факторов, объясняющих значительно большее число измеряемых, «косвенных» показателей, представляющих собой как бы внешнее выражение этих факторов.

В рамках эвристико-алгоритмической парадигмы можно сформулировать ряд подходов к построению единственного фактора f, аппроксимирующего заданные признаки x_k, $k \in K$, например, в смысле максимизации одного из критериев

$$\sum_k \rho(f, x_k), \quad \sum_k |\rho(f, x_k)|, \quad \sum_k \rho^2(f, x_k),$$

где $\rho(f, x_k)$ — коэффициент корреляции искомого фактора f и переменной x_k. Хотя здесь оптимизируемые критерии и выписаны явно, их анализ не выходит за рамки эвристико-алгоритмической парадигмы, поскольку

эти критерии не сформулированы в терминах модели, выражающей x_k и f друг через друга явным образом.

Нетрудно видеть, что фактор f, максимизирующий первый из указанных критериев, получается как средний арифметический вектор $f = \sum_k x_k / |K|$ (для предварительно центрированных и нормированных признаков x_k), называемый обычно центроидом. Подобным образом получается и оптимальное решение по второму из критериев; следует только изменить направление шкалы некоторых из признаков x_k путем замены на $-x_k$. Третий критерий, как известно, является критерием компонентного анализа, который будет кратко охарактеризован ниже.

Для постановки и решения проблем оценки качества получаемого решения с точки зрения его соответствия данным, выбора и построения достаточного числа факторов, используется модельно-аппроксимационная парадигма. Обозначим через x_{ik} значение x_k, а через f_{im} — искомое значение m-го фактора f_m, $m \in M$, для i-го наблюдения, $i \in I$. Линейная модель факторного анализа имеет вид

$$x_{ik} = \sum_m a_{mk} f_{im} + \varepsilon_{ik}, \tag{8}$$

где a_{mk} — искомые коэффициенты, называемые факторными нагрузками, а ε_{ik} — невязки. Критерий оценки искомых a_{mk}, f_{im} по заданным x_{ik} можно выбирать, минимизируя: а) сами невязки в смысле их скалярных сверток, аналогичных (2)

$$\sum_{i,k} \varepsilon_{ik}^2, \quad \sum_{i,k} |\varepsilon_{ik}|, \quad \max_{i,k} |\varepsilon_{ik}|, \tag{9}$$

б) парные корреляции между «остаточными» переменными $\varepsilon_k = (\varepsilon_{ik})$. Первый из этих подходов может быть назван компонентным анализом, поскольку при использовании квадратичного критерия (9) широко известен как анализ главных компонент. Другие два критерия, применительно к модели (8), по-видимому, в литературе не рассматривались. Второй подход принято называть факторным анализом (в собственном смысле этого термина).

Следует особо подчеркнуть, что обе группы методов основаны на оптимизации критериев, зависящих только от $E = (\varepsilon_{ik})$. Дело в том, что решение по такому критерию существенно неоднозначно, оно определено только с точностью до подпространства, содержащего искомые факторы f_m [3].

Практика применения факторного анализа показала, что с точки зрения интерпретации наиболее ценными являются те решения, которые имеют «простую структуру» матрицы нагрузок A, когда значительная часть величин a_{mk} близка к 0, так что интерпретация определяется малым числом нагрузок a_{mk}, близких к ± 1. Это наблюдение вызвало поток работ по вторичной оптимизации модели (8) относительно a_{mk} при заданной матрице E с целью достижения как можно более простой структуры. При так называемом конфирматорном факторном анализе структура нулей матрицы A вообще задана исследователем. На наш взгляд, такое двухшаговое решение не является перспективным с точки зрения математизации данной проблемы, оно уводит из поля аппроксимационной парадигмы в пространство эвристико-алгоритмического подхода.

Концепция «простой структуры» матрицы A должна быть формализована в терминах исходной аппроксимационной модели: до, а не после выбора E. Увы, теории факторного анализа здесь похвалиться нечем. Нам известна только одна работа в указанном направлении — модель экстремальной группировки переменных советского исследователя Э. М. Бравермана. Согласно этой модели, структура нагрузок должна соответствовать разбиению множества исходных переменных на группы K_m таким образом, чтобы для каждого из факторов f_m ненулевыми были

только нагрузки a_{mk} на признаки x_k из той группы K_m, которая соответствует данному фактору [6].

Можно предложить другую реализацию идеи «простой структуры» — с помощью априорного требования, чтобы в качестве значений нагрузок a_{mk} могли служить только числа 0, 1, —1. Оказывается, при таком ограничении метод наименьших квадратов (9) для модели (8) приводит к модификации эвристико-алгоритмического метода формирования центроидов f_m, когда критерием является максимизация $\sum_{k \in K_m} |(f_m, x_k)|$, где K_m — конструируемое множество тех признаков x_k, для которых $a_{mk} \neq 0$.

Следует обсудить еще одну концепцию «простой» структуры, особенно сильно, хотя и в неафишируемой форме, эксплуатируемую в анализе главных компонент и других сводящихся к нему методах анализа данных (анализ соответствий [7, 8], дуальное шкалирование [9] и т. п.).

В анализе главных компонент фактически используется следующий принцип, который может быть назван «принципом последовательного исчерпания»: факторы должны формироваться последовательно, так чтобы каждый следующий из них f_m учитывал максимально возможную долю разброса остаточных данных (т. е. величин $x_{ik}^{(m)} = x_{ik} - \sum_{t<m} a_{tk} f_{it}$), «не объясненных» предыдущими факторами f_t, $t<m$. Если следовать этому принципу, то нет надобности вращать полученные факторы, решение является окончательным. Понятно, что этот принцип применим лишь для определенного круга явлений, когда имеется упорядоченность факторов по степени их проявления в данных.

Таким образом, можно считать, что в настоящее время теория факторного анализа как бы «застыла перед рывком» к таким концепциям, которые будут непосредственно включать структуру факторных нагрузок в аппроксимационную модель. Вероятно, для этого потребуется более глубоко систематизировать и уточнить представления о «хорошо» интерпретируемом решении, «простой» структуре факторного пространства.

Вероятностная парадигма ничего нового в эту проблематику не вносит [4].

5. ПРОБЛЕМА КЛАСТЕР-АНАЛИЗА

Кластер-анализ — направление методологии анализа данных, ориентированное на разделение совокупности наблюдений на однородные в том или ином смысле группы. В этой сфере в настоящее время преобладает эвристико-алгоритмический подход, причем алгоритмы кластер-анализа разрабатываются не только в чисто геометрических терминах, но и в теоретико-графовых, алгебраических, статистических и др.

Не вдаваясь в обсуждение многочисленных идей, подходов, методов и алгоритмов, обратим внимание на три типа возникающих математических задач (впрочем, общих для решения всех проблем в рамках эвристико-алгоритмической парадигмы): а) исследование свойств решений, обеспечиваемых конкретным алгоритмом; б) анализ условий, при которых различные алгоритмы дают одинаковые или сходные результаты; в) построение алгоритмов решения оптимизационных, комбинаторных и иных задач.

Проиллюстрируем сказанное на одном классе алгоритмов, включающих один из наиболее распространенных методов кластер-анализа — метод k-средних [10], являющийся в свою очередь частным случаем метода динамических сгущений [7]. Эти алгоритмы оптимизируют класс критериев, обобщающий один из наиболее интересных как для теории, так и для практики критериев — минимум суммарной дисперсии переменных внутри каждого кластера (взвешенной численностями кластеров) [8].

Обозначим матрицу данных $X = (x_{ik})$, $i \in I$, $k \in K$. Каждый кластер m будем характеризовать ядром (эталоном) — $|K|$-мерным вектором a_m и функцией $r_m(i)$ принадлежности объектов $i \in I$ к кластеру m. Величина

$r_m(i)$ равна 1 или 0 в зависимости от того, принадлежит ли объект (наблюдение) i кластеру m (случай «четких» кластеров). Для так называемых нечетких кластеров величина $r_m(i)$ лежит между 0 и 1 так, что $\sum_m r_m(i) = 1$.

Критерием качества кластерной структуры $R = \{r_1, \ldots, r_M\}$ с функциями принадлежности r_m и ядрами a_m является функционал вида

$$F(R, a) = \sum_{i \in I} \sum_{m=1}^{M} \varphi(r_m(i))\, d(i, a_m), \tag{10}$$

где $d(i, a_m)$ — мера близости между ядром a_m и отдельным объектом i, представленным i-й строкой матрицы данных X, а φ — некоторая заданная монотонная функция (ясно, что для случая четких кластеров знак φ можно в (10) опустить).

Функционал (10) определяет класс алгоритмов «покоординатной» минимизации. Начиная с некоторых заданных a_m, находятся величины $r_m(i)$, минимизирующие (10). Затем при заданных $r_m(i)$ отыскиваются ядра a_m, минимизирующие (10), и так до тех пор, пока процесс не стабилизируется. Конечно, начинать эту процедуру можно не с ядер, а с некоторого заданного разбиения R.

Хорошо известны некоторые свойства данной процедуры. Например, для случая четких кластеров, когда d — евклидова или сити-блок метрика и расчет ведется, начиная с заданных M ядер, кластеры таковы, что каждое наблюдение относится к тому ядру a_m, к которому оно ближе, а ядра a_m являются центрами тяжести (медианами) своих кластеров. Известная теорема Мак Кина [10] устанавливает некоторые вероятностные свойства данной процедуры. В [8] показано, что данный класс алгоритмов весьма широк и включает не только метод k-средних, но и ряд других алгоритмов, разработанных из чисто эвристических соображений для формирования четких или нечетких классификаций.

Это не противоречит тому факту, что и критерий (10), и основанные на нем методы относятся к эвристико-алгоритмической парадигме, поскольку они не связаны явно ни с какой моделью представления данных с помощью кластерной структуры.

В качестве модели можно рассматривать основное соотношение факторного анализа (8), если ограничить область значений факторов f_{im} величинами вида $r_m(i)$, т. е. 0 и 1 для четких кластеров или интервалом [0, 1] — для нечетких (конечно, при дополнительном ограничении $\sum_m f_{im} = 1$). При этих ограничениях соотношение (8) приобретает следующий смысл. Векторы $a_m = (a_{mk})$ представляют собой ядра кластеров m, так что каждый объект i (точнее, представляющая его строка матрицы X) с точностью до невязок равен либо ядру a_m кластера m, содержащего i, (случай четких неперекрывающихся кластеров), либо же выпуклой комбинации ядер всех кластеров a_m, взвешенных весами принадлежности $r_m(i)$ (случай размытых кластеров). Критерии (9) адекватности факторной структуры в данном случае могут быть записаны в виде функционала, аналогичного (10)

$$G(R, a) = \sum_{i \in I} d\left(i, \sum_{m=1}^{M} r_m(i)\, a_m\right), \tag{11}$$

где d — квадрат евклидова расстояния, сити-блок или равномерная метрика в зависимости от того, какой из критериев (9) используется.

Нетрудно видеть, что при четких кластерах, когда $r_m(i) = 0$ или 1, критерии (10) и (11) эквивалентны. Это придает критерию (10) и основанным на нем алгоритмам статус аппроксимационных со всеми вытекающими отсюда последствиями. В частности, появляется возможность оценивать степень разброса данных, «объясняемую» кластерной струк-

турой и отдельными ее элементами, что дает удобный инструмент интерпретации результатов.

Напротив, для случая размытых кластеров критерии (10) и (11) не эквивалентны. Так, ядра a_m размытых кластеров, оптимизирующие (10), представляют собой усредненные характеристики данного кластера, тогда как для (11) ядрами являются крайние точки «облака» объектов в $|K|$-мерном пространстве, порождающие точки, соответствующие объектам, как выпуклые комбинации. Это отражает различия между концепциями «реальных» и «идеальных» типов в логике. Анализ некоторых алгоритмов, получающихся на основе принципа последовательного исчерпания для кластерной интерпретации соотношений (8), содержится в [11, 12].

Та же аппроксимационная модель может быть использована и при неколичественной исходной информации, представленной (1, 0)-столбцами матрицы X, соответствующими отдельным градациям признаков. При этом выявляются некоторые связи с традиционными характеристиками корреляции качественных признаков. Например, при подходящей стандартизации данных в процессе формирования четкого кластера m величина вклада градации k в их квадратичный разброс в точности равна отдельному (m, k)-му слагаемому коэффициента сопряженности Φ^2 (4).

Таким образом, развитие аппроксимационной парадигмы в кластер-анализе представляется и своевременным, и полезным.

Вероятностные модели кластер-анализа в настоящее время либо не связаны с основным потоком работ и предлагаемых методов, как это имеет место для задачи о разделении смесей распределений, либо же играют вспомогательную роль, характеризуя те или иные свойства эвристических алгоритмов. Разработка аппроксимационных вероятностных моделей ждет своего часа.

6. ЗАКЛЮЧЕНИЕ

Перечислим некоторые общие выводы, вытекающие из предыдущего обсуждения. Хотя и опущен был ряд важных тем многомерного статистического анализа, таких как сравнительный анализ средних значений в подмножествах наблюдений (дисперсионный анализ), многомерное шкалирование, моделирование дискретных распределений (логлинейный и другие «анализы»), нижеследующее представляется справедливым и для этих направлений прикладной математической статистики.

1. В математической статистике сегодня сосуществуют три парадигмы: эвристико-алгоритмическая, модельно-аппроксимационная и вероятностная. Основные характеристики структуры данных (среднее, дисперсия (разброс), корреляция, кластер, фактор и т. п.) могут рассматриваться в каждой из них, получая свой смысл в зависимости от парадигмы.

2. Эвристико-алгоритмический подход принято считать ненаучным, поскольку он не вписывается в традиционную схему математизации науки: формирование модели, ее проверка, усовершенствование и т. п. Более того, известный критерий научности — возможность опровержения конкретными данными — не всегда может быть применен в указанной парадигме. Редко удается определить, правильны или нет результаты использования эвристических алгоритмов; значительно точнее говорить о том, хуже или лучше они вписываются в теоретические представления о том явлении, к которому относятся данные. Однако этот подход не менее научен, чем остальные. Просто он применяется на самых ранних этапах исследования, когда система понятий и закономерностей не сформирована. В таких условиях каждый отдельный расчет может казаться субъективным, но будучи повторенными на различных данных, в чем-то совпадая, в чем-то отличаясь, получаемые результаты приводят к накоплению знаний.

3. К магистральным могут быть отнесены только те математико-статистические построения, которые допускают интерпретацию во всех трех

парадигмах. При этом аппроксимационная модель должна являться ведущей, а методы ее идентификации — допускать как эвристико-алгоритмическую, так и вероятностную интерпретацию.

В задачах многомерной статистики аппроксимационный подход, по существу, делает лишь первые шаги. Хотелось бы отметить проблему многокритериальности (для минимизации отдельных невязок). Из вышесказанного вытекают также многообещающие направления развития аппроксимационных моделей в задачах факторного и кластерного анализов: обработка переменных, измеренных в разных шкалах; переход к нелинейным моделям и неквадратическим критериям; разработка алгоритмов аппроксимации; средств интерпретации решений и т. п.

4. В зависимости от используемой математико-статистической парадигмы необходимо применять компьютерные системы анализа данных, ориентированные на различные стратегии исследования. Когда модельно вероятностная парадигма была господствующей, системы статистической обработки были ориентированы на пользователя — специалиста, хорошо знающего модель и основанные на ней методы, поскольку незнание модели приводит к ее неверному использованию и неверным выводам. Неквалифицированный (хотя и вполне профессиональный в своей предметной сфере) экономист или социолог был презираем и изгонялся, платя тем же по отношению к математической статистике.

Сейчас на основе практики разработки и применения средств обработки данных эти представления существенно пошатнулись (если не опровергнуты). Можно рассматривать три типа систем статистической обработки данных.

В рамках алгоритмической парадигмы главное — не сам метод, а адекватность решения, получаемого с его помощью. Пользователь должен быть специалистом в своей области, а система обработки данных — дружественной к нему: обращаться к пользователю на языке, приближенном к языку пользователя, с вопросами о конкретных значениях параметров и начальных условий алгоритмов и предоставлять ему максимально полную информацию о получаемых решениях в терминах используемых данных. Система должна быть удобной для введения экспертной информации.

Ясно, что, оставаясь в рамках алгоритмического подхода, мы получаем решения, многие параметры которого определены пользователем. На первых этапах исследования выбор параметров становится во многом субъективным, что преодолевается обычно с помощью многократной обработки одних и тех же данных с изменяемыми значениями параметров. Переход к аппроксимационным методам существенно уменьшает объем вычислений и труд по интерпретации решений, поскольку за счет включения более агрегированных представлений о структуре данных позволяет автоматизировать выбор параметров и расширить спектр средств интерпретации решений (включая сравнительные оценки вкладов, даваемых отдельными элементами решения).

В системе, основанной на вероятностной парадигме, уровень автоматизируемости и интерпретируемости еще выше. Однако здесь важную составляющую должны представлять средства проверки моделей на реальной информации, а также средства «чистки» данных для достижения их соответствия модельным предположениям.

5. В настоящее время сосуществует много наименований для различных, подчас не очень четко очерченных, аспектов математической статистики: многомерный статистический анализ, анализ данных, вычислительная статистика, прикладная статистика и т. п. Безусловно, все эти термины имеют право на существование, поскольку каждый из них отражает важные реальные процессы развития статистической науки и практики. Материал статьи позволяет еще более расширить этот список и предложить три термина — эвристико-алгоритмическая статистика, аппроксимационная статистика и вероятностная статистика — для отражения феноменов, связанных с описанными парадигмами. На наш взгляд, такое различие может сыграть существенную роль в обучении статистике: с

точки зрения как облегчения в усвоении традиционных понятий, так и некоторого обновления содержания вузовских курсов статистики. Как известно, вузовский курс статистики для студентов экономических специальностей в настоящее время разбит на три части: общая теория статистики, социально-экономическая статистика, математическая статистика. При этом учебники математической статистики включают только вероятностные построения, а общей теории статистики — наряду с методологическими рассмотрениями и совершенно конкретные сведения о способах организации сбора статистической информации, которые явно относятся к первому толкованию термина статистика и соответственно к курсу социально-экономической статистики. По-видимому, более обоснованным было бы объединение методологических глав указанных дисциплин в единый курс «Прикладная статистика».

ЛИТЕРАТУРА

1. *Goodman L. A., Kruskal W. H.* Measures of associations for cross-classifications//J. Amer. Stat. Assoc. 1954. V. 49, 1959. V. 54, 1963. V. 58, 1972. V. 67.
2. *Миркин Б. Г.* Группировки в социально-экономических исследованиях. М.: Финансы и статистика, 1975.
3. *Харман Г.* Современный факторный анализ. М.: Статистика, 1972.
4. *Лоули Д., Максвелл А.* Факторный анализ как статистический метод. М.: Мир, 1967.
5. Статистические методы для электронно-вычислительных машин. М.: Мир, 1986.
6. *Браверман Э. М., Мучник И. Б.* Структурные методы обработки эмпирической информации. М.: Наука, 1983.
7. Методы анализа данных. М.: Финансы и статистика, 1985.
8. *Айвазян С. А., Бухштабер В. М., Енюков И. С., Мешалкин Л. Д.* Прикладная статистика: многомерная классификация и сокращение размерности. М.: Финансы и статистика, 1989.
9. *Nishisato S.* Dual Scaling. Toronto, 1980.
10. *MacQueen J.* Some Methods for Classification and Analysis of Multivariate Observations//Proc. 5th Berceley Symp. Math. Stat. Probab. 1967. V. 1.
11. *Миркин Б. Г.* Метод главных кластеров//Автоматика и телемеханика. 1987. № 10.
12. *Миркин Б. Г.* Линейные модели агрегирования данных и метод последовательного исчерпания//III Всесоюз. школа — семинар «Программно-алгоритмическое обеспечение прикладного многомерного статистического анализа». Ч. I. М.: ЦЭМИ АН СССР, 1987.

Поступила в редакцию
9 VI 1989