От редакции

Редакционный совет журнала принял решение о публикации серии статей, посвященных актуальным для социологии математическим методам, мало доступным в настоящее время для широкого круга читателей (Социология: 4М. 2003. № 16. С. 125). Сложность прежде всего в том, что читателю-социологу неизвестно, где и что искать. В статье президента Российской ассоциации статистических методов А.И. Орлова дается обзор развития статистических методов за последние тридцать лет.

А.И. Орлов
(*Москва*)

СТАТИСТИЧЕСКИЕ МЕТОДЫ В РОССИЙСКОЙ СОЦИОЛОГИИ (ТРИДЦАТЬ ЛЕТ СПУСТЯ)

В статье дан обзор развития статистических методов в российской социологии за 1974–2004 гг. Обсуждаются основные научные события этих лет, прежде всего, формирование прикладной статистики и ее основы – статистики нечисловых данных (в социологии 70–90% переменных имеют нечисловой характер). Рассмотрены методы снижения размерности, в том числе методы состоятельной оценки размерности модели в многомерном шкалировании.

Ключевые слова: математика в социологии, статистические методы, вероятностно-статистические модели, прикладная статистика, статистика нечисловых данных, непараметрическая статистика, многомерное шкалирование, оценка размерности пространства.

Александр Иванович Орлов – доктор технических наук, профессор, директор Института высоких статистических технологий и эконометрики МГТУ им. Н.Э. Баумана. E-mail: orlov@professor.ru; http://orlovs.pp.ru.

«Болезни роста» современной российской социологии

В течение последних 15 лет российская социология бурно развивается по всем количественным параметрам. Если в 1989 г. в России было 6 социологических факультетов, отделений, кафедр, то в 2003 г. – уже 105. Число студентов-социологов выросло более чем в 100 раз. Во всех вузах преподают социологию – она вошла в перечень «Общих гуманитарных и социально-экономических дисциплин» государственных образовательных стандартов высшего профессионального образования. Издается более 20 социологических журналов. Каждый год ВАК утверждает около 50 докторских диссертаций по социологии [1, с. 2–3].

Очевидно, глубину исследованиям придает использование развитого научного аппарата – методологии и методов сбора и анализа данных, математических моделей. На наш взгляд, принципиальный прорыв был осуществлен в нашей стране в 1970-е гг. Именно тогда в арсенале отечественных социологов появились теория измерений и нечеткие множества, математические методы классификации и многомерное шкалирование, непараметрическая статистика и статистика нечисловых данных.

В дальнейшие десятилетия шло естественное развитие научного аппарата. К сожалению, нельзя сказать, что в последние годы темпы этого развития усилились. Исследователи 1970-х гг. выпустили учебники [2; 3; 4; 5], но поток научных результатов в области математических методов в социологии не расширился по сравнению с 1970-ми – периодом «бури и натиска». Из этого следует, в частности, что публикации тех лет [6; 7; 8] отнюдь не устарели, они представляют большой интерес для социологов XXI в.

Итак, социология бурно развивается вширь, но весьма медленно – вглубь. Это вполне естественно. Прочитав в 1970 г. популярную книгу В.Э. Шляпентоха [9], автор провел свое первое полевое исследование. Несмотря на простоту, оно позволило решить управленческие задачи, стоявшие перед автором как директором

Вечерней математической школы (ВМШ) при Московском математическом обществе. Итоги многолетней деятельности в ВМШ подведены в [10]. И сейчас наши ученики-маркетологи, готовя выпускные работы на степень магистра делового администрирования, обходятся полевыми исследованиями на столь же простом уровне (см. описание исследования «Потребители растворимого кофе» в [5, гл. 2]).

И лишь постепенно практики приходят к необходимости применять более сложные методы. Например, в крупном маркетинговом агентстве, опрашивающем за год до 0,5 млн. потребителей, в котором автор этих слов работал консультантом, был создан специализированный отдел обработки данных, сотрудники которого ежедневно применяли различные алгоритмы статистической обработки данных, включенные в известный пакет SPSS.

Как показывает анализ тезисов докладов и выступлений на II Всероссийском социологическом конгрессе «Российское общество и социология в XXI в.: социальные вызовы и альтернативы» [11], большинство участников конгресса не дозрело не только до применения математики, но и до проведения простейших полевых исследований. Математика необходима для продвинутых социологических исследований, когда простейших методов недостаточно. Очевидно, с укреплением социологических центров в них будут возникать подразделения анализа данных, которые сначала будут пользоваться стандартными статистическими пакетами, а затем востребуют и современные методы.

В социологии с успехом используются различные методы анализа данных и разнообразные математические модели [12]. Обсудим развитие методов обработки результатов выборочных исследований за последние тридцать лет.

Основное событие – появление прикладной статистики

Математические методы выборочных исследований. Выборочные исследования – один из основных инструментов социологов. Для переноса выводов с выборки на всю интересующую исследователя совокупность необходимо использовать вероятностно-статистические методы и модели. Уже в 1970-х гг. в нашей стране активно разрабатывались продвинутые математические и статистические методы анализа данных социологических опросов (см., например, сборники [6; 7]). Отметим, что работы тех уже далеких лет, как правило, отнюдь не устарели и по-прежнему представляют интерес для специалистов по анализу социологических данных и математическому моделированию социальных процессов.

Одни и те же математические и статистические методы и модели могут с успехом применяться в самых разных областях науки и практики. Статистические методы и модели весьма эффективны в социологических, социально-экономических, управленческих, технических и технико-экономических исследованиях, медицине, истории – практически в любой прикладной отрасли и области знания[1].

[1] Очевидна связь между исследованиями, выполненными в рамках различных дисциплин. Например, на II Всероссийском социологическом конгрессе (2003 г.) активно обсуждалась такая традиционно экономическая тематика, как маркетинговые и инновационные исследования [11]. Однако для специалиста вполне естественным является желание «замкнуться» внутри своей предметной области. Например, довольно странным выглядело бы предложение о преподавании на социологическом факультете в соответствии с учебником по эконометрике [5]. Удивление значительно возросло бы при констатации того, что этот учебник составлен в основном из статей, опубликованных в журнале «Заводская лаборатория» (в прошлом – орган Министерства черной металлургии). Действительно, есть ли что-либо общее у инженера-металлурга, менеджера, экономиста и социолога? Необходимо известное интеллектуальное развитие, чтобы понять, что все эти

В рассматриваемой области основное событие последних тридцати лет – это становление научно-практической дисциплины «прикладная статистика», посвященной разработке и применению статистических методов и моделей.

Появление прикладной статистики. В нашей стране термин «прикладная статистика» вошел в широкое употребление в 1981 г. после выхода массовым тиражом (33940 экз.) сборника «Современные проблемы кибернетики (прикладная статистика)». В этом сборнике обосновывалась трехкомпонентная структура прикладной статистики [13]. Во-первых, в нее входят ориентированные на прикладную деятельность математико-статистические методы анализа данных (эту область можно назвать прикладной математической статистикой и включать также и в прикладную математику). Однако прикладную статистику нельзя целиком относить к математике. Она включает в себя две внематематические области. Во-первых, методологию организации статистического исследования: как планировать исследование, как собирать данные, как подготавливать данные к обработке, как представлять результаты. Во-вторых, организацию компьютерной обработки данных, в том числе разработку и использование баз данных и электронных таблиц, статистических программных продуктов, например, диалоговых систем анализа данных. В нашей стране термин «прикладная статистика» использовался и ранее 1981 г., но лишь внутри сравнительно небольших и замкнутых групп специалистов. Эти факты предыстории прикладной статистики также рассмотрены в сборнике [13].

Прикладная статистика и математическая статистика – это две разные научные дисциплины. Первая относится к статистике, вторая – к математике. Различие четко проявляется и при преподавании. Курс математической статистики состоит в основном

специалисты могут использовать одни и те же инструменты исследования – статистические методы и модели.

из доказательств теорем, как и соответствующие учебные пособия. В курсах прикладной статистики основное – методология анализа данных и алгоритмы расчетов, а теоремы приводятся для обоснования этих алгоритмов, доказательства же, как правило, опускаются (их можно найти в научной литературе).

Структура современной статистики. Внутренняя структура статистики как науки была выявлена и обоснована при создании в 1990 г. Всесоюзной статистической ассоциации (см. об этом, например, статью [14]).

Прикладная статистика – методическая дисциплина, являющаяся центром статистики. При применении методов прикладной статистики к конкретным областям знаний и отраслям народного хозяйства получаем научно-практические дисциплины типа «статистика в промышленности», «статистика в медицине» и др. С этой точки зрения эконометрика – это «статистические методы в экономике» [5]. Математическая статистика играет роль математического фундамента для прикладной статистики.

К настоящему времени любому специалисту очевидно четко выраженное размежевание математической статистики и прикладной статистики. Математическая статистика исходит из сформулированных в основном в 1930–1950 гг. постановок математических задач, происхождение которых связано с анализом статистических данных. Начиная с 70-х гг. XX в. исследования по математической статистике посвящены лишь обобщению и дальнейшему математическому изучению этих задач. Поток новых математических результатов (теорем) не ослабевает, но новые практические рекомендации по обработке статистических данных при этом не появляются. Можно сказать, что математическая статистика как научное направление замкнулась внутри себя.

Научное направление и сам термин «прикладная статистика» возникли как реакция на описанную выше тенденцию. Прикладная статистика нацелена на решение реальных задач. Поэтому в ней возникают новые постановки математических задач анализа

статистических данных, развиваются и обосновываются новые методы. Обоснование часто проводится математическими методами, т.е. путем доказательства теорем. Большую роль играет методологическая составляющая – как именно ставить задачи, какие предположения принять с целью дальнейшего математического изучения. Велика роль современных информационных технологий, в частности, компьютерного эксперимента.

Рассматриваемое соотношение математической и прикладной статистик отнюдь не является исключением в мире научных дисциплин. Как правило, математические дисциплины проходят в своем развитии ряд этапов. Вначале в какой-либо прикладной области возникает необходимость в применении математических методов, накапливаются соответствующие эмпирические приемы (для геометрии это – «измерение земли», т.е. землемерие, в Древнем Египте). Затем возникает математическая дисциплина со своей аксиоматикой (для геометрии это – время Евклида). Затем идет внутриматематическое развитие и преподавание (считается, что большинство результатов элементарной геометрии получено учителями гимназий в XIX в.). При этом на запросы исходной прикладной области перестают обращать внимание, и та порождает новые научные дисциплины (сейчас «измерением земли» занимается не геометрия, а геодезия и картография). Затем научный интерес к исходной дисциплине иссякает, но преподавание по традиции продолжается (элементарная геометрия до сих пор изучается в средней школе, хотя трудно понять, в каких практических задачах может понадобиться, например, теорема о том, что высоты треугольника пересекаются в одной точке). Следующий этап – окончательное вытеснение дисциплины из реальной жизни в историю науки (объем преподавания элементарной геометрии в настоящее время постепенно сокращается, в частности, ей все меньше уделяется внимания на вступительных экзаменах в вузах). К интеллектуальным дисциплинам, закончившим свой жизненный путь, относится средневековая схоластика. Как справедливо

отмечено, например, в [15], теория вероятностей и математическая статистика успешно двигаются по ее пути – вслед за элементарной геометрией.

Статистические данные собираются и анализируются с незапамятных времен (см., например, Книгу Чисел в Ветхом Завете). Однако современная математическая статистика была создана сравнительно недавно, а именно, в первой половине XX в. Именно тогда были разработаны ее основные идеи, получены результаты, излагаемые ныне в учебных курсах математической статистики. Затем математики занялись разработкой внутриматематических проблем, а для создания новых статистических технологий и теоретического обслуживания практики анализа статистических данных стала использоваться новая дисциплина – прикладная статистика.

Точки роста прикладной статистики

Внутри прикладной статистики наиболее значимым нам представляется создание и развитие статистики объектов нечисловой природы. Ее называют также статистикой нечисловых данных, или нечисловой статистикой. Большое значение имеет развитие непараметрической статистики и методов снижения размерности. Рассмотрим три перечисленные «точки роста» прикладной статистики.

Статистика объектов нечисловой природы как часть прикладной статистики. Согласно общепринятой в настоящее время классификации статистических методов прикладная статистика делится на четыре области: статистика (числовых) случайных величин; многомерный статистический анализ; статистика временных рядов и случайных процессов; статистика объектов нечисловой природы.

Первые три из этих областей являются классическими. Они были хорошо известны еще в первой половине XX в. Остановимся

на четвертой, сравнительно недавно вошедшей в массовое сознание специалистов. Анализ динамики развития прикладной статистики приводит к выводу, что в XXI в. статистика объектов нечисловой природы станет центральной областью прикладной статистики, поскольку содержит наиболее общие подходы и результаты.

Исходный объект в прикладной математической статистике – это выборка. В классической математической статистике элементы выборки – это числа. В многомерном статистическом анализе – вектора. А в нечисловой статистике элементы выборки – это объекты нечисловой природы, которые нельзя складывать и умножать на числа. Другими словами, объекты нечисловой природы лежат в пространствах, не имеющих векторной структуры. Примерами объектов нечисловой природы являются:

– значения качественных признаков, т.е. результаты кодировки, например, вариантов ответа на вопросы социологической анкеты;

– бинарные отношения – упорядочения (ранжировки), классификации (отношения эквивалентности), толерантности[1];

– результаты парных сравнений, т.е. последовательности из 0 и 1;

– множества (обычные или нечеткие);

– слова, предложения, тексты;

– вектора, координаты которых – совокупность значений разнотипных признаков, часть из них носит качественный характер, а часть – количественный.

В течение 1970-х гг. на основе запросов социологии [16], экономики, техники и медицины развивались конкретные направления статистики объектов нечисловой природы. Были установлены связи между конкретными видами таких объектов, разработаны для них вероятностные модели. Научные итоги этого периода подведены в монографии [8].

[1] Толерантность – это рефлексивное симметричное отношение. Отличается от классификации (отношения эквивалентности) возможным отсутствием транзитивности. Толерантностями естественно описывать отношения сходства или знакомства. Вероятностно-статистическая теория толерантностей содержится в монографии [8].

Следующий этап – выделение статистики объектов нечисловой природы в качестве самостоятельного направления в прикладной статистике, ядром которого являются методы статистического анализа данных произвольной природы. Программа развития этого нового научного направления впервые была сформулирована в статье [17]. Реализация этой программы была осуществлена в 1980-е гг. Для работ этого периода характерна сосредоточенность на внутренних проблемах нечисловой статистики. Предварительные итоги были подведены в сборнике научных статей [18], полностью посвященном нечисловой статистике.

К 1990-м гг. статистика объектов нечисловой природы с теоретической точки зрения была достаточно хорошо развита, основные идеи, подходы и методы были разработаны и изучены математически, в частности, доказано достаточно много теорем. Наступило время перейти к применению полученных результатов на практике. Одним из примеров такого применения являются работы по социологии науки [19].

Непараметрическая статистика. Из многих «точек роста» прикладной статистики, рассмотренных в [5], отметим непараметрическую статистику, или непараметрику. В первой трети XX в. в работах Спирмена и Кендалла появились первые непараметрические методы, основанные на коэффициентах ранговой корреляции. Но непараметрика, не делающая нереалистических предположений о том, что функции распределения результатов наблюдений принадлежат тем или иным параметрическим семействам распределений, стала заметной частью статистики лишь со второй трети XX в., после работ А.Н. Колмогорова и Н.В. Смирнова 1930-х гг. После второй мировой войны развитие непараметрической статистики пошло быстрыми темпами. К настоящему времени с помощью непараметрических методов можно решать практически тот же круг статистических задач, что и посредством параметрических. Все большую роль играют непараметрические оценки плотности, непараметрические методы регрессии

и распознавания образов (дискриминантного анализа). В нашей стране непараметрические методы получили достаточно большую известность после выхода в 1965 г. сборника статистических таблиц Л.Н. Большева и Н.В. Смирнова [20], содержащего подробные таблицы для основных непараметрических критериев.

Тем не менее, параметрические методы все еще популярнее непараметрических. Неоднократно публиковались (см., например, [5]) экспериментальные данные, показывающие, что распределения реально наблюдаемых случайных величин, в частности, ошибок измерения, в подавляющем большинстве случаев отличны от нормальных (гауссовских). Тем не менее, теоретики продолжают строить и изучать статистические модели, основанные на гауссовости, а практики – пытаться применять подобные методы и модели. С точки зрения прикладной статистики такие попытки напоминают поиск ключей под фонарем, где светло, а не там, где они потеряны.

Почему же неадекватные параметрические методы довольно часто позволяют получать практически полезные выводы? Прикладная статистика дает возможность изучать свойства конкретных алгоритмов анализа данных на основе вероятностно-статистических моделей. Например, рассмотрим простейший алгоритм анализа данных – расчет выборочного среднего арифметического. Если мы хотим перенести результаты с выборки на более широкую совокупность, то вынуждены использовать ту или иную модель порождения данных, например, рассматривать наблюденные значения как реализации независимых одинаково распределенных случайных величин (векторов или объектов иной природы). Эта модель позволяет обосновать использование выборочного среднего арифметического как точечной оценки теоретического среднего (математического ожидания), указать (доверительные) границы для теоретического среднего и решить иные задачи [5]. В частности, оказывается, что доверительные границы, рассчитанные при нереалистическом предположении

нормальности, при увеличении объема выборки сближаются с адекватными непараметрическими границами, построенными на основе центральной предельной теоремы теории вероятностей. В то же время методы отбраковки резко выделяющихся наблюдений, основанные на гипотезе нормальности, не являются адекватными [5].

Ряд непараметрических методов рассмотрен в обзоре [21]. Более подробное изложение можно найти в учебнике [5], в котором, в частности, продемонстрировано, что свой естественный вид многие непараметрические методы, предназначенные для оценивания среднего, плотности, регрессионной зависимости и решения других задач (в частности, в теории классификации [22]), приобретают в рамках статистики объектов нечисловой природы. Отметим также «широту» непараметрики – в нее входят все методы, не опирающиеся на ту или иную модель принадлежности функций распределения результатов наблюдений к некоторому параметрическому семейству распределений. Ранговые методы составляют лишь часть одномерной непараметрики, как и методы, предполагающие непрерывность функции распределения результатов наблюдений. Например, выборочное среднее арифметическое – это непараметрическая оценка среднего в модели, в которой результаты наблюдений имеют произвольную функцию распределения с конечной дисперсией [5].

Статистические методы и социология. Число актуальных для социологов публикаций по статистическим методам – не менее 100000 [5; 14]. Очевидна актуальность поиска необходимой исследователю информации. В среднесрочной перспективе можно ожидать помощи от Интернета (в http://orlovs.pp.ru). Однако в настоящее время основные результаты представлены на бумажных носителях. На наш взгляд, представленная в настоящей статье концепция развития статистических методов, разработанная Российской ассоциацией статистических методов, окажется полезной специалистам по анализу социологических данных.

Выше неоднократно отмечалась значимость для социологов работ, формально относящихся к экономике, управлению (менеджменту), техническим исследованиям. Обратим внимание на информационный поток, идущий из социологии в другие области. Отметим, например, работу по дискриминантному анализу [23], имеющую общестатистический интерес.

Самостоятельная проблема – внедрение в практику работы организации современных статистических методов. Обратим внимание на систему «Шесть сигм» организации подобного внедрения [24].

Мы рассмотрели развитие идей и научной области, а не персоналии. В краткой, но весьма содержательной сводке [25] описаны основные научные результаты большого числа отечественных исследователей в области статистических методов анализа социологических данных, названы основные исследовательские коллективы Москвы, Петербурга, Новосибирска и многих других городов, приведена обширная библиография (119 названий). Подробное же описание требует серии книг, а не статьи.

Методы снижения размерности

Как уже отмечалось, одной из «точек роста» прикладной статистики являются методы снижения размерности. Они все чаще используются при анализе социологических данных. Рассмотрим наиболее перспективные методы снижения размерности. В качестве примера применения вероятностно-статистического моделирования и результатов статистики нечисловых данных обоснуем состоятельность оценки размерности пространства, ранее предложенной Краскалом из эвристических соображений [26; 27].

В многомерном статистическом анализе каждый объект описывается вектором, размерность которого произвольна, но одна и та же для всех объектов. Однако человек может непосредственно воспринимать лишь числовые данные или точки на плоскости.

Анализировать скопления точек в трехмерном пространстве уже гораздо труднее. Непосредственное восприятие данных более высокой размерности невозможно. Поэтому вполне естественным является желание перейти от многомерной выборки к данным небольшой размерности, чтобы «на них можно было посмотреть».

Кроме стремления к наглядности, есть и другие мотивы для снижения размерности. Те факторы, от которых интересующая исследователя переменная не зависит, лишь мешают статистическому анализу: во-первых, на сбор информации о них расходуются финансовые, временные, кадровые ресурсы, во-вторых, как можно доказать, их включение в анализ ухудшает свойства статистических процедур, в частности, увеличивает дисперсию оценок параметров и характеристик распределений. Поэтому желательно избавиться от таких факторов.

При анализе многомерных данных обычно рассматривают не одну, а множество задач, в частности, по-разному выбирая независимые и зависимые переменные. Поэтому рассмотрим задачу снижения размерности в следующей формулировке. Дана многомерная выборка. Требуется перейти от нее к совокупности векторов меньшей размерности, максимально сохранив структуру исходных данных, по возможности не теряя информации, содержащихся в данных. Задача конкретизируется в рамках каждого метода снижения размерности.

Метод главных компонент является одним из наиболее часто используемых методов снижения размерности. Основная его идея состоит в последовательном выявлении направлений, в которых данные имеют наибольший разброс. Пусть выборка состоит из векторов, одинаково распределенных с вектором $X = (x(1), x(2), \ldots, x(n))$. Рассмотрим линейные комбинации

$$Y(\lambda(1), \lambda(2), \ldots, \lambda(n)) = \lambda(1)x(1) + \lambda(2)x(2) + \ldots + \lambda(n)x(n),$$

где

$$\lambda^2(1) + \lambda^2(2) + \ldots + \lambda^2(n) = 1.$$

Здесь вектор $\lambda = (\lambda(1), \lambda(2), \ldots, \lambda(n))$ лежит на единичной сфере в n-мерном пространстве.

В методе главных компонент прежде всего находят направление максимального разброса, т.е. такое λ, при котором достигает максимума дисперсия случайной величины $Y(\lambda) = Y(\lambda(1), \lambda(2), \ldots, \lambda(n))$. Тогда вектор λ задает первую главную компоненту, а величина $Y(\lambda)$ является проекцией случайного вектора X на ось первой главной компоненты.

Затем, выражаясь терминами линейной алгебры, рассматривают гиперплоскость в n-мерном пространстве, перпендикулярную первой главной компоненте, и проектируют на эту гиперплоскость все элементы выборки. Размерность гиперплоскости на 1 меньше, чем размерность исходного пространства.

В рассматриваемой гиперплоскости процедура повторяется. В ней находят направление наибольшего разброса, т.е. вторую главную компоненту. Затем выделяют гиперплоскость, перпендикулярную первым двум главным компонентам. Ее размерность на 2 меньше, чем размерность исходного пространства. Далее – следующая итерация.

С точки зрения линейной алгебры речь идет о построении нового базиса в n-мерном пространстве, ортами которого служат главные компоненты.

Дисперсия, соответствующая каждой новой главной компоненте, меньше, чем для предыдущей. Обычно останавливаются, когда она меньше заданного порога. Если отобрано k главных компонент, то это означает, что от n-мерного пространства удалось перейти к k-мерному, т.е. сократить размерность, практически не исказив структуру исходных данных.

Для визуального анализа данных часто используют проекции исходных векторов на плоскость первых двух главных компонент. Обычно хорошо видна структура данных, выделяются компактные кластеры объектов и отдельно выделяющиеся вектора.

Метод главных компонент является одним из методов *факторного анализа* [28]. Различные алгоритмы факторного анализа объединены тем, что во всех них происходит переход к новому базису в исходном *n*-мерном пространстве. Важным является понятие «нагрузка фактора», применяемое для описания роли исходного фактора (переменной) в формировании определенного вектора из нового базиса.

Новая идея по сравнению с методом главных компонент состоит в том, что на основе нагрузок происходит разбиение факторов на группы. В одну группу объединяются факторы, имеющие сходное влияние на элементы нового базиса. Затем из каждой группы рекомендуется оставить одного представителя. Иногда вместо выбора представителя расчетным путем формируется новый фактор, являющийся центральным для рассматриваемой группы. Снижение размерности происходит при переходе к системе факторов, являющихся представителями групп. Остальные факторы отбрасываются.

Описанная процедура может быть осуществлена не только с помощью факторного анализа. Речь идет о кластер-анализе признаков (факторов, переменных). Для разбиения признаков на группы можно применять различные алгоритмы кластер-анализа [22]. Достаточно ввести расстояние (меру близости, показатель различия) между признаками. Пусть X и Y – два признака. Различие $d(X,Y)$ между ними можно измерять с помощью выборочных коэффициентов корреляции:

$$d_1(X,Y) = 1 - |r_n(X,Y)|, \quad d_2(X,Y) = 1 - |\rho_n(X,Y)|,$$

где $r_n(X,Y)$ – выборочный линейный коэффициент корреляции Пирсона, $\rho_n(X,Y)$ – выборочный коэффициент ранговой корреляции Спирмена.

Многомерное шкалирование. На использовании расстояний (мер близости, показателей различия) $d(X,Y)$ между признаками X и Y основан обширный класс методов многомерного шкалирования [29; 30]. Основная идея этого класса методов состоит в

представлении каждого объекта точкой геометрического пространства (обычно размерности 1, 2 или 3), координатами которой служат значения скрытых (латентных) факторов, в совокупности достаточно адекватно описывающих объект. При этом отношения между объектами заменяются отношениями между точками – их представителями. Так, данные о сходстве объектов – расстояниями между точками, данные о превосходстве – взаимным расположением точек [31].

В практике анализа социологических данных используется ряд различных моделей многомерного шкалирования. Во всех них встает проблема оценки истинной размерности факторного пространства. Рассмотрим эту проблему на примере обработки данных о сходстве объектов с помощью метрического шкалирования.

Пусть имеется n объектов $O(1)$, $O(2)$, …, $O(n)$, для каждой пары объектов $O(i)$, $O(j)$ задана мера их сходства $s(i,j)$. Считаем, что всегда $s(i,j) = s(j,i)$. Происхождение чисел $s(i,j)$ не имеет значения для описания работы алгоритма. Они могли быть получены либо непосредственным измерением, либо с использованием экспертов, либо путем вычисления по совокупности описательных характеристик, либо как-то иначе.

В евклидовом пространстве рассматриваемые n объектов должны быть представлены конфигурацией n точек, причем в качестве меры близости точек-представителей выступает евклидово расстояние $d(i,j)$ между соответствующими точками. Степень соответствия между совокупностью объектов и совокупностью представляющих их точек определяется путем сопоставления матриц сходства $||s(i,j)||$ и расстояний $||d(i,j)||$. Метрический функционал сходства имеет вид

$$S = \sum_{i<j} |\, s(i,\, j) - d(i,\, j)\,|^2.$$

Геометрическую конфигурацию надо выбирать так, чтобы функционал S достигал своего наименьшего значения [29; 30].

Замечание. В неметрическом шкалировании вместо близости самих мер близости и расстояний рассматривается близость упорядочений на множестве мер близости и множестве соответствующих расстояний. Вместо функционала S используются аналоги ранговых коэффициентов корреляции Спирмена и Кендалла. Другими словами, неметрическое шкалирование исходит из предположения, что меры близости измерены в порядковой шкале.

Пусть евклидово пространство имеет размерность m. Рассмотрим минимум среднего квадрата ошибки

$$\alpha_m = \frac{2}{n(n-1)} \min S,$$

где минимум берется по всем возможным конфигурациям n точек в m-мерном евклидовом пространстве. Можно показать, что рассматриваемый минимум достигается на некоторой конфигурации. Ясно, что при росте m величина α_m монотонно убывает (точнее, не возрастает). Можно показать, что при $m \geq n - 1$ она равна 0 (если $s(i,j)$ – метрика). Для увеличения возможностей содержательной интерпретации желательно действовать в пространстве возможно меньшей размерности. При этом, однако, размерность необходимо выбрать так, чтобы точки представляли объекты без больших искажений. Возникает вопрос: *как рационально выбирать размерность простpранства*, т.е. натуральное число m?

В рамках детерминированного анализа данных обоснованного ответа на этот вопрос, видимо, нет. Следовательно, необходимо изучить поведение α_m в тех или иных вероятностных моделях. Если меры близости $s(i,j)$ являются случайными величинами, распределение которых зависит от «истинной размерности» m_0 (и, возможно, от каких-либо еще параметров), то можно в классическом математико-статистическом стиле ставить задачу оценки m_0, искать состоятельные оценки и т.д.

Начнем строить вероятностные модели. Пусть объекты представляют собой точки в евклидовом пространстве размерности k, где k достаточно велико. То, что «истинная размерность» равна

m_0, означает, что все эти точки лежат на гиперплоскости размерности m_0. Примем для определенности, что совокупность рассматриваемых точек представляет собой выборку из кругового нормального распределения с дисперсией $\sigma^2(0)$. Это означает, что объекты $O(1)$, $O(2)$, …, $O(n)$ являются независимыми в совокупности со случайными векторами, каждый из которых строится как $\zeta(1)e(1) + \zeta(2)e(2) + \ldots + \zeta(m_0)e(m_0)$, где $e(1)$, $e(2)$, … , $e(m_0)$ – ортонормальный базис в подпространстве размерности m_0, в котором лежат рассматриваемые точки, а $\zeta(1)$, $\zeta(2)$, … , $\zeta(m_0)$ – независимые в совокупности одномерные нормальные случайные величины с математическим ожиданием 0 и дисперсией $\sigma^2(0)$.

Рассмотрим две модели получения мер близости $s(i,j)$. В первой из них $s(i,j)$ отличаются от евклидова расстояния между соответствующими точками из-за того, что точки известны с искажениями. Пусть $c(1)$, $c(2)$, …, $c(n)$ – рассматриваемые точки. Тогда

$$s(i,j) = d(c(i) + \varepsilon(i), c(j) + \varepsilon(j)),\ i,j = 1, 2, \ldots, n,$$

где d – евклидово расстояние между точками в k-мерном пространстве, вектора $\varepsilon(1)$, $\varepsilon(2)$, …, $\varepsilon(n)$ представляют собой выборку из кругового нормального распределения в k-мерном пространстве с нулевым математическим ожиданием и ковариационной матрицей $\sigma^2(1)I$, где I – единичная матрица. Другими словами, $e(i) = \eta(1)e(1) + \eta(2)e(2) + \ldots + \eta(k)e(k)$, где $e(1)$, $e(2)$, …, $e(k)$ – ортонормальный базис в k-мерном пространстве, а $\{\eta(i,t), i = 1, 2, \ldots, n, t = 1, 2, \ldots, k\}$ – совокупность независимых в совокупности одномерных случайных величин с нулевым математическим ожиданием и дисперсией $\sigma^2(1)$.

Во второй модели искажения наложены непосредственно на сами расстояния:

$$s(i,j) = d(c(i), c(j)) + \varepsilon(i,j),\ \ i,j = 1, 2, \ldots, n,\ i \neq j,$$

где $\{\varepsilon(i,j),\ i,j = 1, 2, \ldots, n\}$ – независимые в совокупности нормальные случайные величины с математическим ожиданием и дисперсией $\sigma^2(1)$.

В работе [32] показано, что для обеих сформулированных моделей минимум среднего квадрата ошибки α_m при $n \to \infty$ сходится по вероятности к

$$f(m) = f_1(m) + \sigma^2(1)(k - m),\ \ m = 1, 2, \ldots, k,$$

где

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

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

$$m^* = Arg \min_m \{\alpha_{m+1} - 2\alpha_m + \alpha_{m-1}\}$$

является состоятельной оценкой истинной размерности m_0.

Итак, из вероятностной теории вытекает рекомендация – в качестве оценки размерности факторного пространства использовать m^*. Отметим, что подобная рекомендация была сформулирована как эвристическая одним из основателей многомерного шкалирования Дж. Краскалом [26; 27; 29]. Он исходил из опыта практического использования многомерного шкалирования и вычислительных экспериментов. Вероятностная теория позволила обосновать эту эвристическую рекомендацию.

ЛИТЕРАТУРА

1. *Осипов Г.В.* Российская социология в XXI веке // Материалы II Всероссийского социологического конгресса. М., 2003.

2. *Толстова Ю.Н.* Анализ социологических данных: методология, дескриптивная статистика, изучение связей между номинальными признаками. М.: Научный мир, 2000.

3. *Толстова Ю.Н.* Измерение в социологии. М.: Инфра-М, 1998.

4. *Татарова Г.Г.* Методология анализа данных в социологии (введение): Учебник для вузов. М.: NOTA BENE, 1999.

5. *Орлов А.И.* Эконометрика. М.: Экзамен, 2002 (1-е изд.), 2003 (2-е изд.), 2004 (3-е изд.).

6. Методы современной математики и логики в социологических исследованиях / Под ред. Э.П. Андреева. М.: Институт социологических исследований АН СССР, 1977.

7. Математические методы и модели в социологии / Под ред. В.Н. Варыгина. М.: Институт социологических исследований АН СССР, 1977.

8. *Орлов А.И.* Устойчивость в социально-экономических моделях. М.: Наука, 1979.

9. *Шляпентох В.Э.* Социология для всех: некоторые проблемы, результаты, методы. М.: Советская Россия, 1970.

10. *Орлов А.И.* О теоретических основах внеклассной работы по математике и опыте Вечерней математической школы при Московском математическом обществе // Бюллетень № 2 Всесоюзного центра статистических методов и информатики. М.: ВЦСМИ, 1991.

11. Тезисы докладов и выступлений на II Всероссийском социологическом конгрессе «Российское общество и социология в XXI веке: социальные вызовы и альтернативы»: В 3 т. М.: Альфа-М, 2003.

12. Актуальные проблемы социологической науки и социальной практики: Научная конференция «Сорокинские чтения – 2002»: Москва, МГУ им. М.В. Ломоносова, 17-18 декабря 2002 г.: Сб. науч. докл. в 3 т.: Том 3: Математическое моделирование социальных процессов. Вып. 5 / Под общ. ред. А.А. Самарского, В.И. Добренькова, А.П. Михайлова. М.: МАКС Пресс, 2003.

13. Современные проблемы кибернетики (прикладная статистика) / Под ред. А.И. Орлова. М.: Знание, 1981.

14. *Орлов А.И.* О перестройке статистической науки и ее применений // Вестник статистики. 1990. № 1. С. 65–71.

15. *Тутубалин В.Н.* Границы применимости (вероятностно-статистические методы и их возможности). М.: Знание, 1977.

16. *Орлов А.И.* Статистика объектов нечисловой природы и обработка социологических данных // Математические методы в социологическом исследовании. М.: Наука, 1981. С. 67–75.

17. *Орлов А.И.* Статистика объектов нечисловой природы и экспертные оценки // Экспертные оценки: Вопросы кибернетики. Вып. 58. М.: Научный совет АН СССР по комплексной проблеме «Кибернетика», 1979. С.17–33.

18. Анализ нечисловой информации в социологических исследованиях / Под ред. В.Г. Андреенкова, А.И. Орлова, Ю.Н. Толстовой. М.: Наука, 1985.

19. *Орлов А.И., Нечаева Е.Г., Соколов А.В.* Статистика объектов нечисловой природы и анализ данных о научном потенциале // Социология: методология, методы, математические модели. 1995. № 5–6. С. 118–136.

20. *Большев Л.Н., Смирнов Н.В.* Таблицы математической статистики. М.: Наука, 1965 (1-е изд.), 1968 (2-е изд.), 1983 (3-е изд.).

21. *Тюрин Ю.Н., Шмерлинг Д.С.* Непараметрические методы статистики // Социология: методология, методы, математические модели. 2004. № 18. С. 154–166.

22. *Орлов А.И.* Заметки по теории классификации // Социология: методология, методы, математические модели. 1991. № 2. С. 28–50.

23. *Бессокирная Г.П.* Дискриминантный анализ для отбора информативных переменных // Социология: методология, методы, математические модели. 2003. № 16. С. 25–35.

24. *Панде П., Холп Л.* Что такое «Шесть сигм»?: Революционный метод управления качеством / Пер. с англ. М.: Альпина Бизнес Букс, 2004.

25. *Толстова Ю.Н.* Математические методы в социологии // Социология в России / Под ред. В.А. Ядова. 2-е изд., перераб. и дополн. М.: Издательство Института социологии РАН, 1998. С. 83–89, 98–103.

26. *Краскал Дж.* Взаимосвязь между многомерным шкалированием и кластер-анализом // Классификация и кластер. М.: Мир, 1980. С. 20–41.

27. *Kruskal J.B., Wish M.* Multidimensional Scaling // Sage University Paper Series: Qualitative Applications in the Social Sciences. 1978. No. 1.

28. *Харман Г.* Современный факторный анализ. М.: Статистика, 1972.

29. *Терехина А.Ю.* Анализ данных методами многомерного шкалирования. М.: Наука, 1986.

30. *Перекрест В.Т.* Нелинейный типологический анализ социально-экономической информации: Математические и вычислительные методы. Л.: Наука, 1983.

31. *Тюрин Ю.Н., Литвак Б.Г., Орлов А.И., Сатаров Г.А., Шмерлинг Д.С.* Анализ нечисловой информации. М.: Научный совет АН СССР по комплексной проблеме «Кибернетика», 1981.

32. *Орлов А.И.* Общий взгляд на статистику объектов нечисловой природы // Анализ нечисловой информации в социологических исследованиях. М.: Наука, 1985. С. 58–92.