ТАРТУСКИЙ
ГОСУДАРСТВЕННЫЙ
УНИВЕРСИТЕТ

ТРУДЫ

ВЫЧИСЛИТЕЛЬНОГО ЦЕНТРА

44

ТАРТУ
1980

ТАРТУСКИЙ ГОСУДАРСТВЕННЫЙ
УНИВЕРСИТЕТ

МЕТОДЫ МНОГОМЕРНОГО СТАТИСТИЧЕСКОГО АНАЛИЗА

ТРУДЫ
ВЫЧИСЛИТЕЛЬНОГО
ЦЕНТРА

ВЫПУСК 44

ТАРТУ 1980

Утверждено на заседании совета математического факультета ТГУ 15 ноября 1979 года.

КЛАССИФИКАЦИЯ И КАНОНИЧЕСКИЕ МОДЕЛИ СБАЛАНСИРОВАННЫХ ПОЛНЫХ ФАКТОРНЫХ КОМПЛЕКСОВ С ПРОИЗВОЛЬНЫМ ЧИСЛОМ ФАКТОРОВ

Т. Мелс

1. Введение

В настоящей статье описаны в матричной символике некоторые модели сбалансированных дисперсионных анализов (ДА). Поскольку такое исследование в хорошо проработанной классической области может показаться странным, необходимо некоторое обоснование тематики статьи.

Начнем обращением внимания на то, что при большом числе публикаций по линейным статистическим методам, только небольшое число из них занимается систематикой и описанием целых классов ДА. В руководствах (например [3] – [7]) подробно описаны многие частные схемы ДА, однако о многих других возможных схемах иногда даже не упоминается. Поэтому нередко трудно найти в готовом виде подходящую схему ДА для обработки имеющихся данных. Такая ситуация хорошо известна каждому консультанту по математической статистике. Нам кажется, что хорошая систематика ДА может заметно повышать эффективность применения ДА и избегать ошибок, связанных с использованием неадекватных схем.

Есть у проблемы и чисто вычислительный аспект. В много-

факторных моделях ДА численное решение нормальных уравнений может оказаться трудоемкой задачей, особенно там, где нет достаточно мощных ЭВМ. Если, однако, известно аналитическое решение нормальных уравнений, как например в сбалансированных случаях ДА, то объем необходимых вычислений может быть заметно сокращен. Поэтому желательно иметь в библиотеке СП рядом с общими модулями ДА также и специальные модули для разных классов сбалансированных ДА. Здесь, конечно, необходимы специализированные средства описания схем ДА, что требует соответствующих теоретических исследований, в первую очередь по систематике схем ДА.

В настоящей статье сделана попытка дать общее определение и описание т.н. полных факторных комплексов (ПФК). Для сбалансированных ПФК предлагается также теоретический вывод таблицы ДА. Развита соответствующая матричная техника.

В класс ПФК входят хорошо известные "факторные планы", планы с группировкой и их разные гибриды. Неполные факторные комплексы (латинские квадраты и др.) в статье не рассматриваются, хотя и они могут быть довольно элегантно включены в общий формализм факторных комплексов (см. 2.7).

2. Полные факторные комплексы

В этом параграфе определяем полные факторные комплексы и некоторые смежные понятия. Даем также таблицу всевозможных типов ПФК, если число факторов не превышает пяти.

2.1. Любой фактор будем представлять множеством уровней. Например, если фактор A имеет уровни $a_1,\dots,a_s$, то A =

$= \{a_1, \ldots, a_s\}$. Каждому индивиду (= наблюдению, = измерению) $\omega \in \Omega$ (Ω – совокупность индивидов) соответствует некоторый элемент $\langle a, b, \ldots, d, e \rangle$ во множестве $M = A \times B \times \ldots \times D \times E$, где $A, \ldots, E$ – факторы, а $a, \ldots, e$ – их уровни, потенциально оказывающие влияние на индивида ω. Совокупность Ω всех индивидов определяет таким путем в M некоторое подмножество S, которое называем структурой (совокупности индивидов).

2.2. Будем говорить, что две уровни комбинируются (в выбранной совокупности Ω), если они являются координатами некоторой точки структуры S. Если уровни комбинируются, то в Ω найдется индивид, на котором эти уровни реализуются одновременно.

Если совокупность индивидов не пуста, то у всех факторов найдутся уровни, которые комбинируются. Однако мы будем всегда предполагать, что у каждого фактора все уровни комбинируются с некоторыми уровнями других факторов: уровни, которые не комбинируются, можно просто отбросить. В силу сделанного соглашения проекция структуры S на любое координатное множество совпадает с этим множеством.

2.3. Рассматриваем два специальные типа связи между двумя факторами A и B. Будем говорить, что фактор A подчиняется фактору B (символически $A > B$ или $B < A$), если каждый уровень фактора A комбинируется (в Ω) только с одним уровнем фактора B. Если $A < B$ и $B < A$, то факторы A и B являются копиями друг друга и мы называем их эквивалентными. Все эквивалентные факторы могут быть отождествлены без всякого ущерба для теории. Поэтому далее будем всегда предполагать,

что любой фактор эквивалентен только самому себе. В силу этого допущения множество факторов вместе с бинарным отношением $\prec$ превратится в частично упорядоченное множество.

Обозначим $\varphi_1,\ldots,\varphi_\tau$ всевозможные частичные упорядочения фиксированного множества факторов. Каждое упорядочение φ_i удобно интерпретировать как предикат: $\varphi_i(S)$ означает, что структура S порождает на множестве факторов упорядочение φ_i. Предикаты φ_i называем предикатами подчинения.

2.4. Другой специальный тип связи между двумя факторами - это перекрестная связь. Будем говорить, что факторы A и B связаны перекрестно (символически $A \times B$ или $B \times A$), если каждый уровень фактора A комбинируется (в Ω) с каждым уровнем фактора B. Предикаты всевозможных перекрестных связей на некотором фиксированном множестве факторов обозначим $\psi_1,\ldots,\psi_\tau$. Здесь $\psi_i(S)$ означает, что в структуре S между факторами имеются перекрестные связи ψ_i.

2.5. При фиксированном множестве факторов множество всевозможных структур частично упорядочено теоретико-множественным включением $\subset$. Включение $S_1 \subset S_2$ означает, что все комбинации уровней, составляющие структуру S_1, имеются также в структуре S_2.

Пусть φ - некоторый предикат подчинения. Рассматриваем максимальные элементы подмножества $\mathcal{S}_\varphi = \{S : \varphi(S)\}$ в частично упорядоченном множестве всех структур (на M). Каждый максимальный элемент (структура) множества $\mathcal{S}_\varphi$ устанавливает между факторами связи, где имеются всевозможные комбинации уровней, не противоречащие подчинениям φ. Мы будем называть

эти максимальные структуры полными факторными комплексами (ПФК).

2.6. Приведем пример трехфакторного ПФК. Пусть $A = \{a_1, a_2\}$, $B = \{b_1, b_2, b_3\}$, $C = \{c_1, c_2, c_3, c_4\}$, а φ состоит из подчинений $A < B$ и $A < C$. Тогда структура

$$S = \{\langle a_1, b_1, c_1\rangle, \langle a_1, b_1, c_2\rangle, \langle a_1, b_2, c_1\rangle, \langle a_1, b_2, c_2\rangle, \langle a_2, b_3, c_3\rangle, \langle a_2, b_3, c_4\rangle\}$$

является ПФК. Действительно, структура S порождает как подчинение $A < B$ (поскольку b_1 и b_2 соответствует только a_1, а b_3 соответствует только a_2), так и подчинение $A < C$ ($c_1, c_2 \to a_1$, $c_3, c_4 \to a_2$). Но никакая более мощная структура на $A \times B \times C$ этими подчинениями уже не обладает.

2.7. Аналогично ПФК могут быть определены некоторые дуальные им структуры как минимальные (в смысле отношения $\subset$) элементы множества $\mathcal{S}_\psi = \{S : \psi(S)\}$, где ψ – предикат перекрестных связей. Например, латинский квадрат может быть определен как минимальный элемент среди трехфакторных структур, где каждые два фактора связаны перекрестно. Комбинируя, далее, переходы к минимальным и максимальным элементам, можно получить огромное количество интересных факторных структур. Эти вопросы не являются, однако, предметом настоящей статьи.

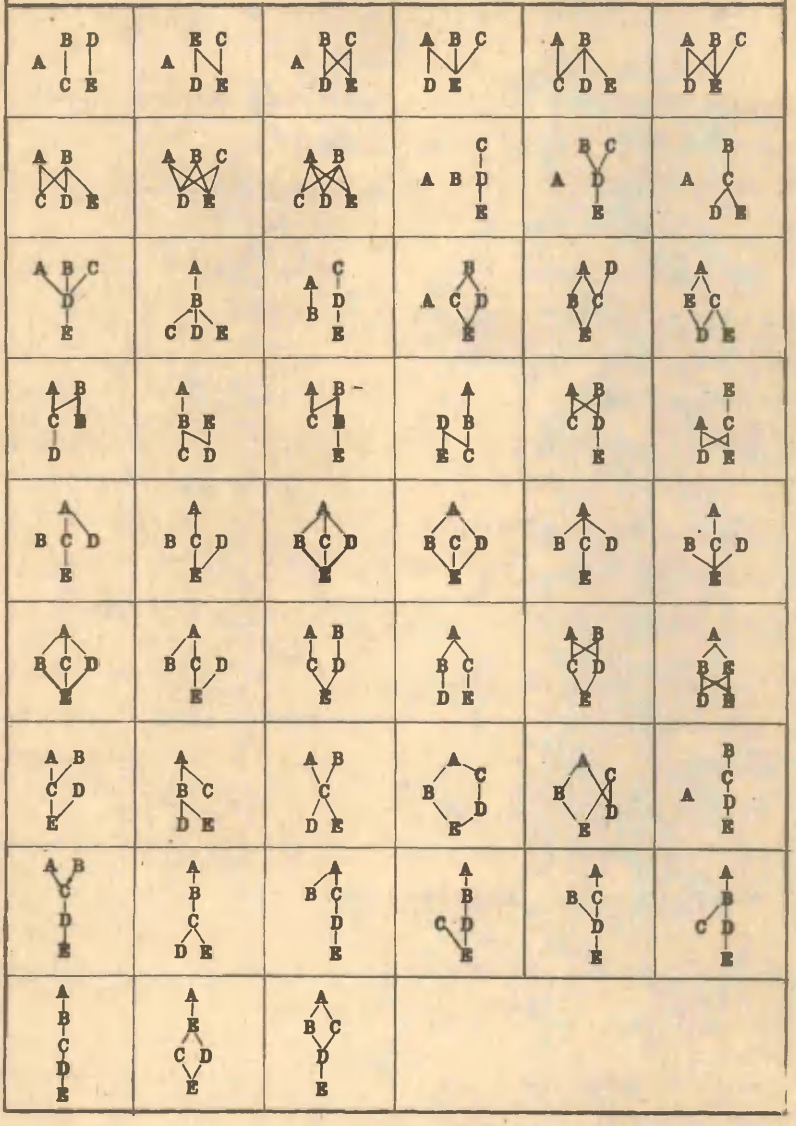
2.8. Переходим к понятию типа ПФК. Ясно, что в общих чертах тип ПФК определяется предикатом подчинения или (что равносильно) соответствующим частичным упорядочением это-

Таблица I

Структурные графы полных факторных комплексов

Двухфакторные комплексы					
					
Трехфакторные комплексы					
					
Четырехфакторные комплексы					
					
					
					
Пятифакторные комплексы					
					
					

Пятифакторные комплексы (продолж.)



го ПФК. Но разные упорядочения φ_1 на множестве факторов могут быть изоморфны в том смысле, что после подходящего переименования факторов они совпадают. Изоморфны, например, упорядочения $\varphi_1 = \{A<B\}$ и $\varphi_2 = \{B<A\}$. Поэтому типом ПФК естественно называть не само упорядочение φ этого ПФК, а класс эквивалентности упорядочения φ по изоморфности.

Непосредственным перебором вариантов можно убедиться, что в случае двух факторов A и B имеется два типа ПФК – $\{\ \}$ и $\{A<B\}$. В случае трех факторов имеется уже пять типов ПФК: $\{\ \}$, $\{A<B\}$, $\{A<B,\ A<C\}$, $\{A<B,\ C<B\}$ и $\{A<B<C\}$. Для четырех- и пятифакторных ПФК различных типов соответственно 16 и 63.

Конкретные примеры на все типы двух-, трех- и четырехфакторных ПФК приведены в [2]. Эти примеры показывают, что всевозможные типы ПФК могут вполне естественным образом встречаться в исследованиях по биологии, технике и медицине.

2.9. Различные типы ПФК наглядно и удобно представлять ориентированными графами: вершины графа изображают факторы, а каждому подчинению $A<B$ соответствует ребро, направленное из A в B. Если начертить эти графы таким образом, чтобы в них горизонтальных ребр нет, а все ребра считать направленными сверху вниз (тогда на ребрах стрелки указать не надо), то получим "графы", которые называем структурными графами. Структурные графы всевозможных типов ПФК с числом факторов до пяти изображены в таблице 1.

2.10. Выделяем теперь ПФК, обладающие определенной симметрией. Пусть $A = \{a_1,\dots,a_s\}$ – фактор в ПФК. Обозначим $\mathcal{A}$

класс всех факторов, которым A подчиняется. Тогда любой уровень $a_i \in A$ определяет однозначно некоторый кортеж $f(a_i)$ уровней факторов класса $\mathcal{A}$. Введем в A отношение эквивалентности ("эквивалентность подчинения") $\sim$, считая $a_i \sim a_j$ если $f(a_i) = f(a_j)$. Если, например, фактор A не подчиняется другим факторам (т.е. $\mathcal{A} = \emptyset$), то $f(a_i)$ - пустой кортеж и при всех $a_i, a_j \in A$ имеем $a_i \sim a_j$.

Разбиение A на классы эквивалентности по $\sim$ называем естественным разбиением фактора. ПФК называем симметричным, если классы естественного разбиения любого фактора всегда имеют одинаковую мощность (зависящую от фактора). Симметричны, например, все ПФК без подчинений.

Симметричный ПФК называем сбалансированным, если все имеющиеся в нем комбинации факторных уровней реализуются на одинаковом числе (≥ 1) индивидов. Основным объектом дальнейшего изучения в статье и являются сбалансированные ПФК.

2.11. В случае ПФК удобно пользоваться т.н. сокращенной нумерацией уровней факторов, которая широко используется в разных вариантах ДА. Техника применения этой методики заключается в следующем. Каждому фактору в ПФК сопоставляется (латинский) индекс, называемый коренным индексом фактора. Внутри каждого естественного класса эквивалентности уровни фактора нумеруются коренным индексом фактора. Класс эквивалентности в свою очередь характеризуется однозначно значениями коренных индексов всех факторов, которым данный фактор подчиняется. Эти индексы указываются в скобках за коренным индексом фактора и называются дополнительными индексами.

Если ПФК симметричен, то у всякого фактора все классы эквивалентности имеют равную мощность, которую обозначаем n_A у фактора A, n_B у фактора B и т.д. и называем сокращенной размерностью фактора. Коренный индекс фактора A в симметричном ПФК пробегает значения от 1 до n_A.

В сбалансированном ПФК постоянным является также и число индивидов при каждой комбинации факторных уровней; это число (называемое повторностью измерения) обозначим n, а коренный индекс отдельного индивида (наблюдения) обозначим ν ($\nu=1,\ldots,n$). Общее число индивидов в совокупности Ω, структура которой является сбалансированным ПФК, равно произведению всех сокращенных размерностей факторов ПФК на повторность n.

3. Линейные модели сбалансированных ПФК

3.1. Переходим к конкретизации линейной модели

$$Y = K\Psi + \varepsilon \qquad (1)$$

для некоторого сбалансированного k-факторного ПФК. Как обычно, Y является здесь $(N \times q)$-матрицей наблюдений (N – число индивидов в Ω, q – размерность признак-вектора), а матрицы K и Ψ называются соответственно плановой матрицей и параметрической матрицей. Остаток ε считаем случайной центрированной матрицей.

Конкретизация модели (1) состоит в установлении порядка, в котором стоят численные значения признаков в матрице Y, и в выборе параметрической матрицы Ψ и соответствующей плановой матрицы K. В случае сбалансированных ПФК некоторый кано-

нический метод этих конкретизаций подсказывается структурой ПФК. Несколько следующих пунктов посвящены описанию этого метода.

3.2. Результаты измерения q исследуемых признаков у индивида образуют строку в $У$. Поэтому строки матрицы наблюдений можно упорядочить как индивиды в $\mathfrak{Q}$. Каноническим является следующий способ. Сначала упорядочиваем некоторым образом (обычно в алфавитном порядке) множество всех коренных индексов, включая ν, который занимает последнее место. Выбранный порядок индексов фиксируем и будем называть каноническим. Набор $k+1$ индексов, взятых в каноническом порядке, определяет однозначно индивид, а значит, и строку в $У$. Если теперь упорядочить значения индексных наборов лексикографически, то и получим искомый канонический вид матрицы наблюдений.

3.3. Чтобы конкретизировать правую часть (1), будем выделить в $К$ некоторые блоки столбцов, $К = (К_1 \vdots \ldots \vdots К_r)$, а в $\mathcal{Y}$ выделим соответствующие им блоки строк $\mathcal{Y}_1$. Модель (1) может быть тогда представлен в виде

$$У = \sum_{i=1}^{r} К_i \mathcal{Y}_i + \varepsilon . \qquad (2)$$

Рассматриваем естественный (канонический) выбор блоков $К_1$ и $\mathcal{Y}_1$ в случае сбалансированного ПФК. Все блоки в параметрической матрице $\mathcal{Y}$ разделяются на три класса. Первый класс состоит только из одного блока $\mathcal{Y}_1$ – вектора общих средних всех q признаков:

$$\mathcal{Y}_1 = (\mu^{(1)}, \ldots, \mu^{(q)}) .$$

Второй класс образуют блоки т.н. главных эффектов уровней факторов: главные эффекты каждого фактора образуют отдельный блок. Каждому уровню фактора соответствует q главных эффектов - на каждый признак один эффект. Эти эффекты образуют строку блока.

Главные эффекты обозначим одинаково с уровнями фактора, используя сокращенную нумерацию. При необходимости указываем верхним индексом еще номер признака (т.е. номер столбца в блоке). Например, $a^{(t)}_{i(k)}$ есть главный эффект факторного уровня $a_{i(k)}$ на t-тый признак. Строки в блоке упорядочиваем лексикографически по набору нижних индексов. Например, блок главных эффектов вида $a^{(t)}_{i(k)}$ может выглядеть как

$$\Psi_2 = \begin{pmatrix} a^{(1)}_{1(1)} & a^{(2)}_{1(1)} \\ a^{(1)}_{1(2)} & a^{(2)}_{1(2)} \\ a^{(1)}_{2(1)} & a^{(2)}_{2(1)} \\ a^{(1)}_{2(2)} & a^{(2)}_{2(2)} \end{pmatrix}.$$

Третий класс образуют блоки из эффектов взаимодействий уровней разных факторов. Каждому набору факторов, где ни один фактор не подчиняется никакому другому фактору набора, соответствует блок взаимодействий. Блоки взаимодействий построены аналогично блокам главных эффектов. Например, эффект взаимодействия уровней $a_{i(k)}$ и $b_{j(kl)}$ на t-тый признак обозначим $ab^{(t)}_{ij(kl)}$ и этот эффект находится в t-том столбце и ij(kl)-ой строке блока взаимодействий факторов A и B.

3.4. Описываем блоки K_i, соответствующие каноническому разложению параметрической матрицы. Введем для этого обозначения

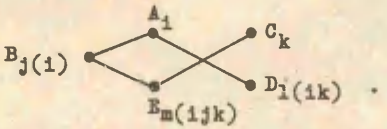
1_p – единичная матрица порядка p,

$\dot{1}_p$ – столбец из p единиц,

$\breve{1}_p$ – специальная матрица, $\breve{1}_p^T = (1_{p-1} : -\dot{1}_{p-1})$.

Каждый блок K_i имеет вид тензорного произведения $k+1$ матриц 1_p, $\dot{1}_p$ или $\breve{1}_p$ (k – число факторов в ПФК). Состав сомножителей в K_i определяется нижними индексами элементов блока ψ_i, а порядок сомножителей – установленным каноническим порядком индексов. Каждому индексу, который не встречается у элементов блока ψ_i, соответствует в K_i (на позиции, фиксированной для этого индекса) сомножитель $\dot{1}_p$. Коренным индексам ψ_i соответствуют в K_i сомножители $\breve{1}_p$, а дополнительным индексам – сомножители 1_p. Размерность p каждого сомножителя в K_i равно сокращенной размерности фактора, соответствующего этому индексу.

3.5. Рассмотрим пример модели пятифакторного сбалансированного ПФК. Пусть факторы A – E связаны в ПФК согласно структурному графу



(3)

На графе указаны также коренные и дополнительные индексы факторов. Фиксируем следующий порядок индексов: i,j,k,l,m,ν.

В данном случае параметрическая матрица разлагается на 10 блоков $\psi_1,\ldots,\psi_{10}$ с элементами соответственно $\mu^{(t)}$, $a_i^{(t)}$,

$b^{(t)}_{j(i)}$, $c^{(t)}_{k}$, $d^{(t)}_{l(ik)}$, $e^{(t)}_{m(ijk)}$, $ac^{(t)}_{ik}$, $bc^{(t)}_{jk(i)}$, $bd^{(t)}_{jl(ik)}$ и $ed^{(t)}_{lm(ijk)}$. Взаимодействия трех или большего числа факторов в модели данного ПФК отсутствуют.

Канонический вид плановой матрицы получается следующий:

$$K = (1\,1\,1\,1\,1\,1\,1 \;:\; 1\,0\,1\,0\,1\,0\,1\,0\,1\,0\,1\,0\,1 \;:\; 1\,0\,1\,0\,1\,0\,1\,0\,1\,0\,1\,0\,1 \;:$$
$$1\,0\,1\,0\,1\,0\,1\,0\,1\,0\,1\,0\,1 \;:\; 1\,0\,1\,0\,1\,0\,1\,0\,1\,0\,1\,0\,1 \;:\; 1\,0\,1\,0\,1\,0\,1\,0\,1\,0\,1\,0\,1 \;:$$
$$1\,0\,1\,0\,1\,0\,1\,0\,1\,0\,1\,0\,1 \;:\; 1\,0\,1\,0\,1\,0\,1\,0\,1\,0\,1\,0\,1 \;:\; 1\,0\,1\,0\,1\,0\,1\,0\,1\,0\,1\,0\,1), \qquad (4)$$

где сомножители в каждом блоке имеют соответственно n_A, n_B, n_C, n_D, n_E и n строк.

Если по найденным $\hat{\mu}$ и K вычислить произвольный элемент правой части (1), то получим для него равенство

$$y^{(t)}_{ijklm\nu} = \mu^{(t)} + a^{(t)}_{i} + b^{(t)}_{j(i)} + c^{(t)}_{k} + d^{(t)}_{l(ik)} + e^{(t)}_{m(ijk)} +$$
$$+ ac^{(t)}_{ik} + bc^{(t)}_{jk(i)} + bd^{(t)}_{jl(ik)} + ed^{(t)}_{lm(ijk)} + \varepsilon^{(t)}_{\nu(ijklm)}, \qquad (5)$$

где $i=1,\dots,n_A$; $j=1,\dots,n_B$; ... ; $\nu=1,\dots,n$. Равенство (5) дает представление модели (2) в компонентах.

3.6. С помощью канонической плановой матрицы нетрудно найти проектор Q, порождающий т.н. остаточную сумму квадратов y^TQy. Известно [5], что

$$Q = 1 - K(K^TK)^-K^T, \qquad (6)$$

где $^-$ обозначает обобщенную обратную матрицу. Вычислить K^TK помогает следующая лемма.

Лемма 1. Если K_i и K_j – разные канонические блоки в плановой матрице сбалансированного ПФК, то $K_i^TK_j = 0$.

Доказательство. Обозначим $K_i = \alpha_1 \otimes \ldots \otimes \alpha_{k+1}$ и $K_j = \beta_1 \otimes \ldots \otimes \beta_{k+1}$, где $\alpha_1, \ldots, \beta_{k+1}$ равны 1, $\dot{1}$ или $\breve{1}$ (размерность матрицы зависит от порядкового номера сомножителя). Тогда

$$K_i^T K_j = \alpha_1^T \beta_1 \otimes \ldots \otimes \alpha_{k+1}^T \beta_{k+1} . \qquad (7)$$

Выражение (7) равно нулю, коль скоро в нем некоторый сомножитель $\alpha_1^T \beta_1$ имеет вид $\breve{1}^T \dot{1}$, так как

$$\breve{1}^T \dot{1} = 0 . \qquad (8)$$

Обозначим множества коренных и дополнительных индексов блока Ψ_i соответственно через G и J_G, а у блока Ψ_j соответственно через F и J_F. Пусть, кроме того, $F = F_1 \cup H$ и $G = G_1 \cup H$, где $F_1 \cap G = G_1 \cap F = \emptyset$. Тогда, в силу (8), $K_i^T K_j$ может не равняться нулю лишь при условии, что каждому F_1-сомножителю ($= \breve{1}$) в K_i соответствует в K_j сомножитель 1, т.е. J_G-сомножитель. Другими словами, необходимым условием для $K_i^T K_j \neq 0$ является $F_1 \subset J_G$.

Включение $F_1 \subset J_G$ означает, что каждому F_1-фактору подчиняется некоторый G-фактор. Поскольку никакой F-фактор не подчиняется никакому отличному от него F_1-фактору ($F_1 \subset F$, а F – множество коренных индексов эффекта !), то G-факторы, которые подчиняются F_1-факторам, не входят в F. Значит, они входят в $G - F = G_1$. Увидим, что если $K_i^T K_j \neq 0$, то каждому F_1-фактору подчиняется некоторый G_1-фактор. Но наше рассуждение вполне симметрично относительно G_1 и F_1. Поэтому верно также, что при $K_i^T K_j \neq 0$ каждому G_1-фактору подчиняется некоторый F_1-фактор. Стало быть $F_1 \cap G_1 \neq \emptyset$, что, однако, невозможно. Лемма доказана.

На основе доказанной леммы, K^TK имеет блочно-диагональный вид, где на диагонали находятся блоки $K_i^TK_i$, являющиеся, очевидно, тензорными произведениями матриц $\dot{1}^T\dot{1}$, $\check{1}^T\check{1}$ или 1. Эти матрицы регулярные, так что $(K_i^TK_i)^- = (K_i^TK_i)^{-1}$. Поэтому $(K^TK)^- = (K^TK)^{-1}$ и $(K^TK)^{-1}$ также имеет блочно-диагональный вид, где диагональные блоки являются тензорными произведениями матриц $(\dot{1}^T\dot{1})^{-1}$, $(\check{1}^T\check{1})^{-1}$ или 1. Теперь легко найти, что

$$Q = 1 - \sum_{t=1}^{r} K_t(K_t^TK_t)^{-1}K_t^T \ . \qquad (9)$$

3.7. Если, согласно некоторой нулевой гипотезе $H_{0,t}$, t-тый эффект отсутствует (= нулевой), то блоки ϑ_t и K_t в (2) следует пропустить. Проектор Q принимает при $H_{0,t}$ вид

$$Q_t = Q + K_t(K_t^TK_t)^{-1}K_t^T \ ,$$

а остаточная сумма квадратов увеличивается на

$$y^T(Q_t - Q)y = y^TK_t(K_t^TK_t)^{-1}K_t^Ty \ . \qquad (10)$$

Выясним более подробно структуру проектора $K_t(K_t^TK_t)^{-1}K_t^T$. Мы уже установили, что эта матрица является тензорным произведением k + 1 сомножителей трех возможных типов, соответствующих трем возможным ролям индексов в элементах ϑ_t. Рассмотрим некоторый индекс и пусть p - сокращенная размерность его фактора. Если t-тый блок ϑ_t не содержит данного индекса, то индексу соответствует в тензорном произведении сомножитель

$$\dot{1}_p(\dot{1}_p^T\dot{1}_p)^{-1}\dot{1}^T = p^{-1}U_p \equiv \bar{U}_p \ , \qquad (11)$$

где U_p – $(p \times p)$–матрица из единиц. Далее, если данный индекс является коренным индексом эффекта, то ему соответствует сомножитель вида

$$\check{1}_p(\check{1}_p^T\check{1}_p)^{-1}\check{1}_p^T = \check{1}_p(1_{p-1} - \frac{1}{p}U_{p-1})\check{1}_p^T = 1_p - \bar{U}_p, \qquad (12)$$

а любому добавочному индексу соответствует сомножитель 1_p. Например, для блока $\vartheta_t = \vartheta_9 = (bd_{jl(ik)})$ в модели (5) проектор в (10) имеет строение

$$Q_t - Q = 1_{n_A} \otimes (1_{n_B} - \bar{U}_{n_B}) \otimes 1_{n_C} \otimes (1_{n_D} - \bar{U}_{n_D}) \otimes \bar{U}_{n_E} \otimes \bar{U}_n. \qquad (13)$$

3.8. Возвращаемся к вычислению (10). Основная трудность здесь заключается в несовместимости тензорного характера проектора Q и нетензорной природы матрицы У. Чтобы обойти это затруднение, преобразуем матрицу наблюдений к тензорному виду, используя формальные факторные операторы $\hat{A},\ldots,\hat{E}$ и $\hat{\eta}$, а также формальный индивид σ. Будем предполагать, что факторные операторы и индивид σ являются элементами некоторых векторных пространств, которые имеют соответственно размерности $n_A,\ldots,n_E$, n и q (q – размерность исследуемого в ДА вектора признаков).

Компоненты (нечисловые "координаты") операторов $\hat{A},\ldots,\hat{\eta}$ в некотором ортогональном базисе обозначим $\hat{A}_1,\ldots,\hat{A}_{n_A};\ldots;$ $\hat{\eta}_1,\ldots,\hat{\eta}_n$. Сочетания $\hat{A}_i\hat{B}_j\ldots\hat{\eta}_\nu$ называем операторами измерения. Мы определяем их правилом

$$(\hat{A}_i\ldots\hat{E}_m\hat{\eta}_\nu)(\sigma_1,\ldots,\sigma_q) = (y^{(1)}_{i\ldots m\nu},\ldots,y^{(q)}_{i\ldots m\nu}), \qquad (14)$$

где $(\sigma_1,\ldots,\sigma_q) = \sigma$, а $\sigma_1,\ldots,\sigma_q$ удобно называть обобщенными

признаками. Оператор измерения, действуя на обобщенный признак, порождает число. Эту операцию естественно интерпретировать как генерирование значения признака зависящим от оператора способом.

3.9. Алгебраические и матричные операции над результатами измерения сводятся теперь к операциям над операторами измерения. Линейная комбинация операторов измерения переводит по нашему определению объект σ в линейную комбинацию значений исходных операторов от σ. В частности, из (14) вытекает представление матрицы наблюдений

$$y = (\hat{A} \otimes \ldots \otimes \hat{\eta})\sigma . \qquad (15)$$

Введем еще векторные операторы

$$\bar{A} = \bar{U}_{n_A} \hat{A} = \begin{pmatrix} \hat{A}_\cdot \\ \vdots \\ \hat{A}_\cdot \end{pmatrix}, \ldots, \bar{\eta} = \bar{U}_n \hat{\eta} = \begin{pmatrix} \hat{\eta}_\cdot \\ \vdots \\ \hat{\eta}_\cdot \end{pmatrix} ,$$

где

$$\hat{A}_\cdot = \frac{1}{n_A} \sum_{i=1}^{n_A} \hat{A}_i, \ldots, \hat{\eta}_\cdot = \frac{1}{n} \sum_{\nu=1}^{n} \hat{\eta}_\nu .$$

Тогда имеем, например,

$$(\hat{A}_i \ldots \hat{E}_m \hat{\eta}_\cdot)\sigma = n^{-1} \sum_{\nu=1}^{n} (\hat{A}_i \ldots \hat{E}_m \hat{\eta}_\nu)\sigma = (y^{(1)}_{i \ldots m_\cdot}, \ldots, y^{(q)}_{i \ldots m_\cdot}), \qquad (16)$$

где в правой части использована известная [1] в ДА операция пунктирования индекса (усреднение по индексу). Матрица из строк (16) имеет вид $(\hat{A} \otimes \ldots \otimes \hat{E} \otimes \bar{\eta})\sigma$.

3.10. Переходим к вычислению (10). Используя (15), немедленно получим равенство

$$y^T(Q_t - Q)y = \sigma^T(\hat{A}^T\psi_1\hat{A} \otimes \ldots \otimes \hat{\eta}^T\psi_{k+1}\hat{\eta})\sigma , \qquad (17)$$

где (согласно 3.7) ψ_1 равно 1, $1-\bar{U}$ или $\bar{U}$, в зависимости от того, является ли i-ый индекс дополнительным, коренным или отсутствующим в ψ_t. Например,

$$\hat{A}^T\psi_1\hat{A} = \begin{cases} \sum_i \hat{A}_i^T\hat{A}_i, & \text{если } i \text{ - дополнительный,} \\ \sum_i \hat{A}_\cdot^T\hat{A}_\cdot, & \text{если } i \text{ не используется в } \psi_t, \\ \sum_i (\hat{A}_i^T\hat{A}_i - \hat{A}_\cdot^T\hat{A}_\cdot), & \text{если } i \text{ - коренный.} \end{cases} \qquad (18)$$

Применяя правило (18) для каждого $\psi_1, \ldots, \psi_{k+1}$ в (17), находим окончательный вид произведения (10). Например, для эффекта $\psi_t = (bd_{jl(ik)})$ в (5) получим

$$y^T(Q_t - Q)y = \sum_{i,\ldots,m,\nu} \sigma^T[\hat{A}_i^T\hat{A}_i \otimes (\hat{B}_j^T\hat{B}_j - \hat{B}_\cdot^T\hat{B}_\cdot) \otimes \hat{C}_k^T\hat{C}_k \otimes (\hat{D}_l^T\hat{D}_l - \hat{D}_\cdot^T\hat{D}_\cdot) \otimes$$

$$\otimes \hat{E}_\cdot^T\hat{E}_\cdot \otimes \hat{\eta}_\cdot^T\hat{\eta}_\cdot]\sigma = \sum_{i,\ldots,m,\nu} (y_{ijkl\ldots}^T y_{ijkl\ldots} - y_{ijk\ldots}^T y_{ijk\ldots} -$$

$$- y_{i\cdot kl\ldots}^T y_{i\cdot kl\ldots} + y_{i\cdot k\ldots}^T y_{i\cdot k\ldots}), \qquad (19)$$

где y с индексами обозначает строку q чисел. Напомним, что (19) выражает прирост остаточной суммы квадратов от игнорирования эффекта ψ_t в модели (2). Формула (19) непосредственно пригодна для вычисления этого прироста.

3.11. Резюмируя сделанное отметим, что ДА л ю б о г о сбалансированного ПФК можно выполнить стандартной методикой, которая хорошо известна для разных частных случаев ПФК, например в [7]. При доказательстве правил ДА ПФК удобно поль-

зоваться канонической формой линейной модели и операторами измерения.

Сформулируем, наконец, основные технические приемы ДА ПФК в виде теоремы.

Теорема (дисперсионного анализа сбалансированных ПФК). В случае произвольного k-факторного сбалансированного ПФК полная квадратная сумма $y^T y$ ($= (q \times q)$-матрица) разлагается в квадратные суммы отдельных эффектов и остаточную квадратную сумму $y^T Q y$. Эффекты соответствуют общему среднему μ и всевозможным наборам $k \geq 1$ взаимно не подчиняющихся факторов. Квадратную сумму любого эффекта можно вычислить суммированием по всем $k+1$ индексам выражения, которое является суммой всевозможных матриц вида $(-1)^{\varrho} y^T_{\_\_} y_{\_\_}$. Здесь знак $\_\_$ обозначает $(k+1)$-местное индексное поле, где точками заменены все индексы, которые в эффекте отсутствуют, и еще $\varrho \geq 0$ коренных индексов, а все дополнительные индексы эффекта обязательно указаны.

Число степеней свободы квадратной суммы эффекта равно произведению $k+1$ сомножителей, где каждому пропущенному в эффекте индексу соответствует сомножитель 1, каждому дополнительному индексу соответствует сомножитель p, а каждому коренному индексу эффекта соответствует сомножитель $p-1$ (p обозначает сокращенную размерность индекса).

Доказательство в основном уже дано. Разложение суммы $y^T y$ вытекает из равенства

$$y^T y = y^T Q y + \sum_{t=1}^{T} y^T (Q_t - Q) y ,$$

которое следует из (9). Число степеней свободы квадратной суммы легко вычислить как $sp(Q_t - Q)$, учтя что след тензорного произведения матриц (например, произведения (13)) равен произведению следов сомножителей.

Литература

1. Ahrens, H., Varianzanalyse. Berlin, 1967.
2. Mõls, T., Matemaatika bioloogidele I. Tartu, 1978.
3. Длин А.М., Факторный анализ в производстве. М., 1975.
4. Кендалл М., Стьюарт А., Многомерный статистический анализ и временные ряды. М., 1976.
5. Рао С.Р., Линейные статистические методы и их применения. М., 1968.
6. Хикс Ч., Основные принципы планирования эксперимента. М., 1967.
7. Шеффе Г., Дисперсионный анализ. М., 1963.

НОВОЕ СЕМЕЙСТВО ПОКАЗАТЕЛЕЙ СТАТИСТИЧЕСКОЙ ЗАВИСИМОСТИ И ЕГО ПРИМЕНЕНИЯ

Ю. Вардья, Т. Мелс

В работе введено семейство показателей тесноты взаимной статистической зависимости двух случайных величин на конечном множестве и указаны некоторые применения этих показателей. Доказано, что рассматриваемые показатели являются корреляционными функционалами в смысле [1]. То, что они имеют тривиальную характерную группу, позволяет применять их для проверки независимости против зависимости заданного вида. Методом Монте-Карло моделирования показано, что для определенных альтернативов соответствующий критерий имеет большую мощность чем критерий χ^2.

1. Определения и соглашения

В данной работе рассматриваются случайные величины на конечном множестве $M = \{a_1, \ldots, a_n\}$. Класс всех таких случайных величин обозначим X, а подкласс невырожденных случайных величин в нем обозначим Y.

Следуя [1] будем говорить, что отображение $k : Y \times Y \to [0,1]$ является корреляционным функционалом (КФ), если оно удовлетворяет следующим условиям 1^0–8^0 (аксиомам КФ): (1^0) сущест-

вует такая группа G преобразований $M \to M$, что $k(x,y) = 1$ тогда и только тогда, когда $P\{x = gy\} = 1$ при некотором $g \in G$ (G называется характерной группой); (2^0) значение $k(x,y)$ зависит от x и y через их совместное распределение $P_{x,y}$, т.е. $k(x,y) = k(P_{x,y})$; (3^0) $k(x,y) = k(y,x)$; (4^0) если $k(x_i,x) \to 1$, то $k(x_i,y) \to k(x,y)$ при любом $y \in Y$; (5^0) если x и y статистически независимы, то $k(x,y) = 0$; (6^0) если $k(x_i,x) \to 1$, то можно найти $g_1, g_2, \ldots$ в G таким образом, чтобы $g_i x_i \to x$ по вероятности; (7^0) если x и z независимы и $0 \le t \le 1$, то

$$k(P_{x,y}) \ge k(tP_{x,y} + (1-t)P_{x,z}); \qquad (1)$$

(8^0) при $t \to 1$ неравенство (1) переходит в равенство.

КФ называется чувствительным, если для него верно обращение аксиомы 5^0, и широким, если из сходимости по вероятности $x_i \to x$ следует сходимость $k(x_i,x) \to 1$. Известно [2], что для случайных величин на конечном множестве существуют только широкие КФ, а для любой группы преобразований $M \to M$ существует чувствительный КФ, имеющий эту группу своей характерной группой. Однако, до сих пор мало описано применений введенных понятий. В следующем пункте статьи мы вводим одно семейство чувствительных широких КФ, имеющих тривиальную характерную группу, а затем (в п. 3) описываем некоторые возможные применения этого семейства.

2. Семейство метрических корреляционных функционалов

Пусть $M = \{a_1, \ldots, a_n\}$. Дадим функционал $k_{\alpha,\sigma,\nu,\varkappa}(x,y)$ для всех случайных величин x и y из множества Y невырожденных случайных величин правилом

$$k_{\alpha,c,\nu,\varkappa}(x,y) = \frac{d_\alpha^{\nu}(P_{x,y};P^{\otimes}_{x,y})}{d_\alpha^{\nu}(P_{x,y};P^{\otimes}_{x,y}) + c d_\alpha^{\varkappa}(P_{x,y};P^{\Delta}_{x,y})}, \qquad (2)$$

где $P_{x,y}$ – распределение вектора (x,y), $P^{\otimes}_{x,y} = P_x \otimes P_y$, d_α – метрика вида

$$d_\alpha(P,Q) = \sqrt{\sum_{i,j=1}^{n} \alpha_{ij}(p_{ij} - q_{ij})^2}, \qquad (3)$$

($\alpha_{ij} = \alpha_{ji} > 0$, $i,j = 1,\dots,n$), $P^{\Delta}_{x,y}$ – ближайшее в метрике d_α к $P_{x,y}$ распределение класса $\Delta_2 = \{P: p_{ij} = 0, \text{ если } i \neq j\}$, c – константа. В частном случае, где $\alpha_{ij} = 1$ для всех $i,j = 1,\dots,n$, сокращаем обозначение $k_{\alpha,c,\nu,\varkappa}$ до $k_{c,\nu,\varkappa}$.

Отметим, что определение функционала (2) корректно, поскольку

$$d_\alpha^{\nu}(P_{x,y};P^{\otimes}_{x,y}) + c d_\alpha^{\varkappa}(P_{x,y};P^{\Delta}_{x,y}) = 0$$

тогда и только тогда, когда $P_{x,y} = P^{\otimes}_{x,y} = P^{\Delta}_{x,y}$, откуда следует, что пара (x,y) не входит в $Y \times Y$, т.е. в область определения функционала (2).

Теорема. Функционал $k_{\alpha,c,\nu,\varkappa}(x,y)$ является при $0 < \varkappa \leq \nu$ и $c > 0$ широким чувствительным корреляционным функционалом с тривиальной характерной группой.

Доказательство. Покажем сначала, что функционал $k_{\alpha,c,\nu,\varkappa}$ удовлетворяет аксиомам корреляционного функционала. Удовлетворенность аксиом $1^0, 2^0$ и 5^0 видно непосредственно. Выполненность аксиомы 3^0 легко вытекает из (3) и из симметрии $\alpha_{ij} = \alpha_{ji}$. Проверим аксиомы 4^0 и 6^0.

Если $k_{\alpha,c,\nu,\varkappa}(x_i,x) \to 1$, то $c d_\alpha^{\varkappa}(P_{x_i,x};P^{\Delta}_{x_i,x}) \to 0$ или $d_\alpha(P_{x_i,x};P^{\Delta}_{x_i,x}) \to 0$. Это возможно только при $P\{x_i = x\} \to 1$

или $P_{x_i,x} \to P_{x,x}$, откуда, очевидно, следует сходимость по вероятности $x_i \to x$. Тем самым $k_{\alpha,c,\nu,\varkappa}(x,y)$ удовлетворяет аксиоме 6°. Далее, из сходимости $x_i \to x$ следует сходимость $P_{x_i,y} \to P_{x,y}$ и в силу непрерывности диагонализации $P \to P^{\Delta}$ и непрерывности метрики d_α имеем

$$d_\alpha(P_{x_i,y};P^{\Delta}_{x_i,y}) \to d_\alpha(P_{x,y};P^{\Delta}_{x,y}).$$

Аналогично, из непрерывности рандомизации $P \to P^{\otimes}$ следует сходимость

$$d_\alpha(P_{x_i,y};P^{\otimes}_{x_i,y}) \to d_\alpha(P_{x,y};P^{\otimes}_{x,y}).$$

Таким образом,

$$k_{\alpha,c,\nu,\varkappa}(x_i,y) \to k_{\alpha,c,\nu,\varkappa}(x,y),$$

т.е. $k_{\alpha,c,\nu,\varkappa}$ удовлетворяет аксиоме 4°.

Проверим аксиомы 7° и 8°. Покажем сначала, что если x и z независимы, $0 < \varkappa \leq \nu$ и $0 \leq t \leq 1$, то имеют место равенство

$$d^{\nu}_{\alpha}((tP_{x,y}+(1-t)P_{x,z});(tP_{x,y}+(1-t)P_{x,z})^{\otimes}) = t^{\nu} d^{\nu}_{\alpha}(P_{x,y};P^{\otimes}_{x,y}) \quad (4)$$

и неравенство

$$d^{\varkappa}_{\alpha}((tP_{x,y}+(1-t)P_{x,z});(tP_{x,y}+(1-t)P_{x,z})^{\Delta}) \leq t^{\varkappa} d^{\varkappa}_{\alpha}(P_{x,y};P^{\Delta}_{x,y}). \quad (5)$$

Пусть распределения случайных величин (x,y), x, y и z задаются соответственно матрицами (p_{ij}), (p_i), (q_j) и (s_j). Тогда

$$d^{2}_{\alpha}((tP_{x,y}+(1-t)P_{x,z});(tP_{x,y}+(1-t)P_{x,z})^{\otimes}) =$$

$$= \sum_{i,j=1}^{n} \alpha_{ij}(tp_{ij}-tp_i q_j)^2 = t^2 \sum_{i,j=1}^{n} \alpha_{ij}(p_{ij}-p_i q_j)^2 = t^2 d^2_{\alpha}(P_{x,y};P^{\otimes}_{x,y}),$$

откуда и вытекает (4). Пользуясь далее формулой

$$\delta_k = p_{kk} + \sum_{i\neq j} p_{ij} \,/\, \alpha_{kk} \sum_{I} \alpha_{ll}^{-1} ,$$

где δ_k – диагональный элемент матрицы $P^{\Delta}_{x,y}$, получим

$$d^2_{\alpha}(P_{x,y};P^{\Delta}_{x,y}) = \sum_{k} \alpha_{kk} \frac{(\sum_{i\neq j} p_{ij})^2}{\alpha^2_{kk}(\sum_{I} \alpha^{-1}_{ll})^2} + \sum_{i\neq j} \alpha_{ij} p^2_{ij} =$$

$$= (\sum_{i\neq j} p_{ij})^2 \,/\, A + \sum_{i\neq j} \alpha_{ij} p^2_{ij} ,$$

где $A = \sum_{I} \alpha^{-1}_{ll}$. Следовательно,

$$d^2_{\alpha}((tP_{x,y}+(1-t)P_{x,z});(tP_{x,y}+(1-t)P_{x,z})^{\Delta}) =$$

$$= (\sum_{i\neq j} (tp_{ij}+(1-t)p_i s_j))^2 / A + \sum_{i\neq j} \alpha_{ij}(tp_{ij}+(1-t)p_i s_j)^2 \geq$$

$$\geq t^2(\sum_{i\neq j} p_{ij})^2 / A + t^2 \sum_{i\neq j} \alpha_{ij} p^2_{ij} = t^2 d^2_{\alpha}(P_{x,y};P^{\Delta}_{x,y}) ,$$

что и доказывает (5).

Возвращаемся к доказательству аксиом 7^0 и 8^0. Пользуясь формулами (2) и (4) и неравенством $0 \leq t \leq 1$, получим

$$k_{\alpha,o,\nu,\varkappa}(P_{x,y}) \frac{d^{\nu}_{\alpha}((tP_{x,y}+(1-t)P_{x,z});(tP_{x,y}+(1-t)P_{x,z})^{\otimes})}{d^{\nu}_{\alpha}((tP_{x,y}+(1-t)P_{x,z});(tP_{x,y}+(1-t)P_{x,z})^{\otimes})+ct^{\nu}d^{\varkappa}_{\alpha}(P_{x,y};P^{\Delta}_{x,y})}.$$

Учтя неравенство $t^{\nu} \leq t^{\varkappa}$ и формулу (5) увидим, что действительно

$$k_{\alpha,o,\nu,\varkappa}(P_{x,y}) \geq k_{\alpha,o,\nu,\varkappa}(tP_{x,y} + (1-t)P_{x,z}). \qquad (6)$$

Таким образом, аксиома 7^0 выполнена. Легко видеть, что при $t \to 1$ неравенство (6) превратится в равенство и поэтому

$k_{\alpha,o,\nu,\varkappa}$ удовлетворяет также и аксиоме 8^{o}.

Мы уже доказали, что $k_{\alpha,o,\nu,\varkappa}$ является КФ. Осталось установить его чувствительность, так как широкость вытекает из общей теории КФ. Допустим, что x и y статистически зависимы. Тогда $d_{\alpha}^{\nu}(P_{x,y};P_{x,y}^{\otimes})>0$ и, следовательно, $k_{\alpha,o,\nu,\varkappa}(x,y)>0$. Значит, $k_{\alpha,o,\nu,\varkappa}$ – чувствительный КФ. Теорема полностью доказана.

КФ, определенный формулой (2), будем в дальнейшем называть метрическим КФ. Укажем здесь некоторые простейшие свойства метрических КФ (в их истинности легко убедиться, представив КФ (2) в виде

$$k_{\alpha,o,\nu,\varkappa}(x,y) = \frac{1}{1 + o\, d_{\alpha}^{\varkappa}(P_{x,y};P_{x,y}^{\Delta}) / d_{\alpha}^{\nu}(P_{x,y};P_{x,y}^{\otimes})}\ ):$$

при $o_1 \geq o_2$

$$k_{\alpha,o_1,\nu,\varkappa}(x,y) \leq k_{\alpha,o_2,\nu,\varkappa}(x,y);$$

при $\nu_1 \geq \nu_2$

$$k_{\alpha,o,\nu_1,\varkappa}(x,y) \geq k_{\alpha,o,\nu_2,\varkappa}(x,y);$$

при $\varkappa_1 \geq \varkappa_2$

$$k_{\alpha,o,\nu,\varkappa_1}(x,y) \leq k_{\alpha,o,\nu,\varkappa_2}(x,y);$$

при $\varkappa_1/\nu_1 = \varkappa_2/\nu_2$

$$k_{\alpha,o_2,\nu_2,\varkappa_2}(x,y) = (1+o_2((k_{\alpha,o_1,\nu_1,\varkappa_1}^{-1}(x,y)-1)/o_1)^{\nu_2/\nu_1})^{-1};$$

при $\alpha = b \cdot \alpha'$

$$k_{\alpha,o,\nu,\varkappa}(x,y) = k_{\alpha',o\cdot b^{\frac{\varkappa-\nu}{2}},\nu,\varkappa}(x,y).$$

3. Применение метрических корреляционных функционалов к исследованию статистической зависимости

Рассматриваем канал связи с n символами $a_1,\ldots,a_n$, входом x и выходом y. Статистическую зависимость y от x интерпретируем как передачу информации через канал. Однако связь устроена так, что выходной символ принимается как входной, т.е. преобразования символов в канале не предполагается. Поэтому передаваемая (с помощью статистической зависимости) информация может иметь практическую ценность лишь при условии, что вероятность совпадения входных и выходных символов в некотором среднем смысле превышает вероятность их случайного совпадения. В таком случае будем говорить, что имеется прямая зависимость между входом и выходом.

При изучении описанного канала иногда возникает необходимость измерять силу прямой статистической зависимости между входом и выходом, а также проверить реальность этой зависимости. Мы предлагаем использовать для решения этих или аналогичных проблем КФ с тривиальной характерной группой, поскольку они более чувствительны к прямой зависимости аргументов.

Некоторые метрические КФ также имеют достаточно малую чувствительность к непрямой зависимости аргументов. Чтобы иллюстрировать возможности применения метрических КФ, рассматриваем ради простоты двоичный канал с передаваемыми символами a_1 и a_2. Если маргинальные распределения P_x и P_y фиксированы для такого канала, то вероятность $p_{11} = P\{x = a_1 \text{ и } y = a_2\}$ однозначно определяет совместное распределение $P_{x,y}$.

Поэтому в данном случае в качестве аргумента для КФ достаточно взять p_{11}.

На рисунке 1 показаны графики некоторых метрических КФ как функции от p_{11}. Эти графики показывают, как можно подобрать при фиксированных P_x и P_y параметры c и $\varkappa=\nu$, чтобы значение $k_{c,\varkappa,\nu}$ превышал бы заданный уровень (например, 0.09) лишь при достаточно сильной прямой зависимости (т.е. при $p_{11} \gg p_{1.}.p_{.1}$, где $p_{1.} = P\{x = a_1\}$ и $p_{.1} = P\{y = a_1\}$), и в этом случае значение $k_{c,\varkappa,\nu}$ близко к 1.

Теперь допустим, что распределение $P_{x,y}$ оценивается эмпирическим распределением (которое обозначаем $P^{э}_{x,y}$), а $k_{c,\varkappa,\nu}(P^{э}_{x,y})$ служит статистиком для проверки прямой зависимости между x и y. Тогда область нечувствительности $k_{c,\varkappa,\nu}$ охватывает и некоторую долю случаев (несильной) прямой статистической зависимости (см. рис. 1). Подбором параметров c и $\varkappa$ эта область может быть совмещена с областью принятия нулевой гипотезы H_0 об отсутствии прямой зависимости. В таком случае нулевая гипотеза принимается при $k_{c,\nu,\varkappa}(P^{э}_{x,y}) \approx 0$ и отклоняется при $k_{c,\varkappa,\nu}(P^{э}_{x,y}) \approx 1$. Иными словами, статистик $k_{c,\nu,\varkappa}(P^{э}_{x,y})$ является своего рода индикатором критической области этой гипотезы.

Применение описанного критерия было статистически моделировано (на микрокалькуляторе TI-58). По результатам этого моделирования построено эмпирическое распределение величины $k_{0.01,10,10}(P^{э}_{x,y})$, где каждое эмпирическое распределение $P^{э}_{x,y}$ получено по $N = 15$ измерениям – имитациям вектора (x,y) с маргинальными распределениями $P_x = P_y = (0.5,0.5)$ и заданным значением p_{11}. Сводка результатов этого моделирования

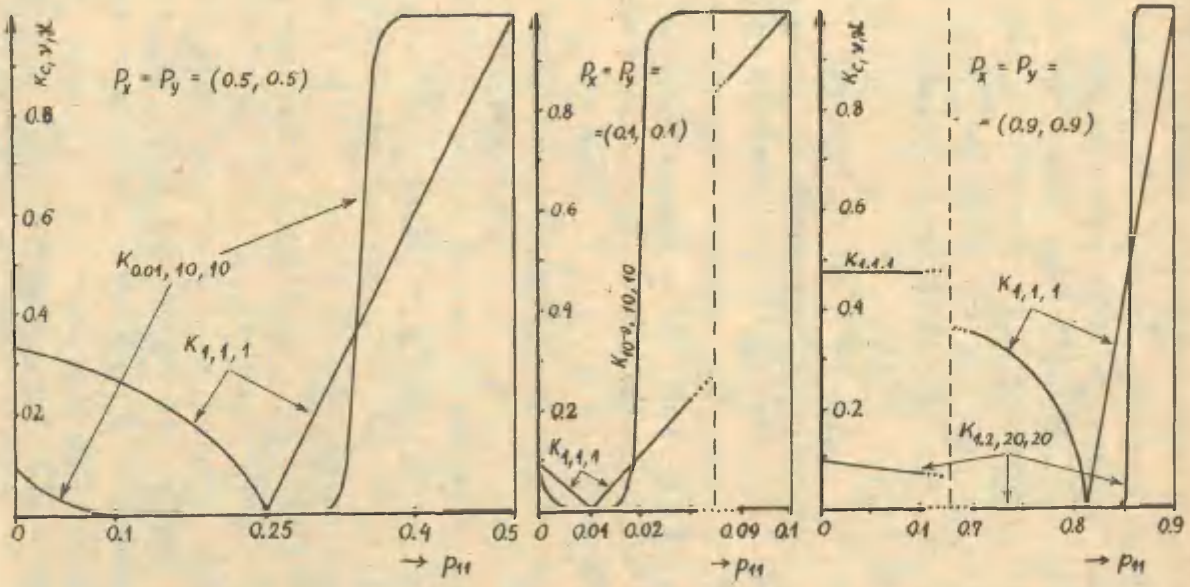


Рис. 1. Графики некоторых метрических корреляционных функционалов в двумерном случае.

Таблица 1.

Сводка результатов статистического моделирования применений критериев $k_{0.01,10,10}(P^{s}_{x,y})$ и χ^2 при $P_x=P_y=(0.5,0.5)$ и $N=15$.
В скобках даны результаты, касающиеся критерия χ^2.

Характер статистической зависимости	P_{11}	Эмпир. распред. $k_{0.01,10,10}(P^{s}_{x,y})$						Число моделир. $P^{s}_{x,y}$	Эмпир. уровень значимости	Эмпир. мощность критерия
		[0.0, 0.1)	[0.1, 0.2)	[0.2, 0.5)	[0.5, 0.8)	[0.8, 0.9)	[0.9, 1.0]			
непрямая зависимость	0.10							(105)	(0.63)	
	0.15							(121)	(0.32)	
	0.20	0.98	0.00	0.00	0.01	0.01	0.00	152	0.02	
								(116)	(0.09)	
независимость	0.25	0.94	0.00	0.00	0.01	0.03	0.00	259	0.058	
								(211)	(0.062)	
	0.30	0.75	0.01	0.00	0.07	0.17	0.00	257		0.25
								(110)		(0.15)
	0.35	0.48	0.01	0.00	0.12	0.39	0.01	124		0.52
прямая								(150)		(0.30)
зависимость	0.40	0.16	0.00	0.00	0.04	0.79	0.02	107		0.84
								(116)		(0.68)
	0.45	0.02	0.00	0.00	0.03	0.80	0.14	209		0.98
								(142)		(0.97)
	0.50							-		1.00

приведена в таблице 1.

Как видно, в данном случае значения $k_{0.01,10,10}$ редко попадают в интервал [0.1,0.8), что и следовало бы ожидать (см. рис. 1). Из таблицы 1 и построенного по ней рисунка 2 вытекает также, что для рассматриваемой пары конкурирующих гипотез мощность критерия, построенного по $k_{0.01,10,10}(P^{э}_{x,y})$, превышает мощность обычного критерия χ^2.

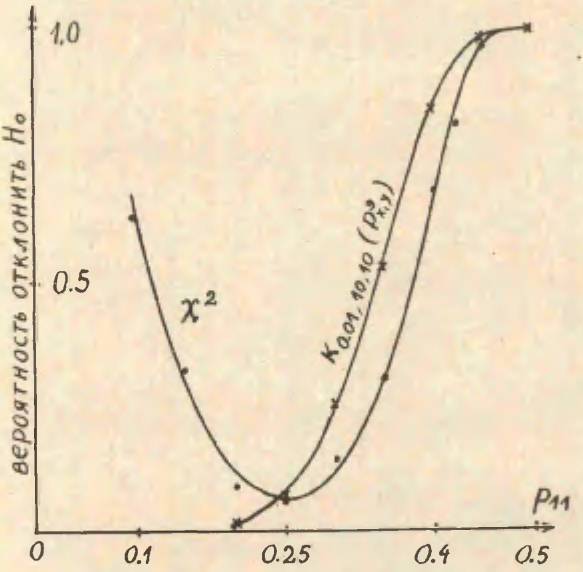


Рис. 2. Кривые мощности двух критериев.

Литература

1. Мелс Т.Э., Определение общих корреляционных коэффициентов. Труды ВЦ ТГУ, 1972, 25, 61-78.
2. Мелс Т.Э., Описание класса общих корреляционных коэффициентов для случайных элементов на конечном множестве. Труды ВЦ ТГУ, 1972, 26, 3-18.

АСИМПТОТИЧЕСКОЕ РАСПРЕДЕЛЕНИЕ КОЭФФИЦИЕНТОВ ПОЛИНОМИАЛЬНОЙ РЕГРЕССИИ

А.-М. Парринг

1. Введение

Пусть дан случайный вектор $x_*' = (x_0, x_1, \ldots, x_p)$. Функцией регрессии для компонента x_0 называется функция $g(x)$, выполняющая условие

$$E(x_0 - g(x))^2 = \min_{\varphi \in G} (x_0 - \varphi(x))^2,$$

где G – класс измеримых функций. Вычисление (оценивание) функции регрессии возможно в случае, когда известно распределение вектора x_* (распределение известно до точности неизвестных параметров, т.е. возможно определить аналитический вид функции $g(x)$). Меньше априорной информации требует определение наилучшего приближения к компоненту x_0 из некоторого класса функции $\mathcal{F}$, т.е. функции $h(x)$, которая удовлетворяет условию

$$E(x_0 - h(x))^2 = \min_{\varphi \in \mathcal{F}} E(x_0 - \varphi(x))^2.$$

Функцию $h(x)$ будем называть функцией регрессии из класса $\mathcal{F}$. Как известно (см. [3], стр. 126), функция $h(x)$ является наилучшим приближением функции регрессии из класса $\mathcal{F}$ в смысле

наименьших квадратов.

В дальнейшем посмотрим подробнее случай, где классом $\mathcal{F}$ выбран класс многочленов степени k из компонентов вектора x.

2. Функция полиномиальной регрессии

Пусть классом $\mathcal{F}_{k,I}$ является класс многочленов степени k из компонентов вектора x, т.е. если $f \in \mathcal{F}_{k,I}$, то

$$f(x) = \alpha_0 + \sum_{(l_1,\dots,l_p)\in I} \alpha_{l_1\dots l_p} x_1^{l_1}\dots x_p^{l_p},$$

где $I = \{(l_1,\dots,l_p) : 0 \le l_s \le k, \sum_{s=1}^{p} l_s \le k\}$ – некоторое фиксированное множество. Например, если p=2 и

$$I = \{(1,0),(0,1),(1,1)\}, \qquad (1)$$

то $f(x) = \alpha_0 + \alpha_{10}x_1 + \alpha_{01}x_2 + \alpha_{11}x_1x_2$. Функцию регрессии из класса $\mathcal{F}_{k,I}$ называем функцией полиномиальной регрессии.

Пусть множество I имеет s элементов. Обозначим вектор, компонентами которого являются упорядоченные некоторым образом произведения $x_1^{l_1}\dots x_p^{l_p}$ через[1] z, $z \in \mathcal{M}^{s\times 1}$, а вектор, компонентами которого являются таким же образом упорядоченные коэффициенты $\alpha_{l_1\dots l_p}$, через $\mathcal{A}$. Тогда функция полиномиальной регрессии имеет вид $f(x) = \mathcal{A}_0 + \mathcal{A}'z$. Предположим, что среди компонентов вектора x нет линейно зависимых и ни один из компонентов вектора x не является полиномиальной функцией от остальных. Так как функция полиномиальной регрессии явля-

---

[1] Символом $\mathcal{M}^{k\times l}$ обозначен класс (k×l)-мерных матриц.

ется линейной функцией вектора z, то вектор коэффициентов[2] $\alpha_*' = [\alpha_0 : \alpha']$ определяется системой линейных уравнений (см. напр. [3], стр. 129)

$$\begin{cases} \alpha = \Sigma^{-1} \sigma_0 \\ \alpha_0 = \mu_0 - \alpha'\mu , \end{cases} \tag{2}$$

где через $\Sigma_* = \begin{bmatrix} \sigma_{00} & \sigma_0' \\ \sigma_0 & \Sigma^{-1} \end{bmatrix}$ обозначена дисперсионная матрица, а через $\mu_*' = [\mu_0 : \mu']$ вектор средних вектора z_*, $z_*' = [z_0 : z']$, $z_0 = x_0$. Тесноту связи измеряет множественный коэффициент корреляции (МКК)

$$\rho^2 = \sigma_0' \Sigma^{-1} \sigma_0 \; / \; \sigma_{00} \; . \tag{3}$$

Выборочную оценку a_* вектора α_* получим, заменив в системе (2) дисперсионную матрицу и вектор средних их выборочными оценками S_* и $\bar{x}_*$. Асимптотическое распределение вектора оценок изучено в [2]. Оказывается, что независимо от распределения вектора z_*, асимптотическим распределением вектора оценок является нормальное распределение. Параметры этого распределения зависят от центральных моментов третьего и четвертого порядка.

3. Асимптотическое распределение коэффициентов полиномиальной регрессии

Введем следующую символику. Обозначим через ${}_*L_k$ случайную матрицу, определенную рекурсивным соотношением[3]

---

[2] Квадратные скобки указывают блочную структуру матрицы. Так блок $\alpha' = (\alpha_1, \ldots, \alpha_s)$.

$$ {}_{*}L_k = \begin{cases} {}_{*}L_{k-1} \otimes z_{*}', & \text{если } k \text{ четное} \\ {}_{*}L_{k-1} \otimes z_{*}, & \text{если } k \text{ нечетное} \end{cases} $$

и через ${}_{*}\bar{L}_k$ аналогичную матрицу, определенную для центрированного случайного вектора $z_{*} - \mu_{*}$

$$ {}_{*}\bar{L}_k = \begin{cases} {}_{*}\bar{L}_{k-1} \otimes (z_{*} - \mu_{*})', & \text{если } k \text{ четное} \\ {}_{*}\bar{L}_{k-1} \otimes (z_{*} - \mu_{*}), & \text{если } k \text{ нечетное,} \end{cases} $$

$k=1,2,\ldots$; ${}_{*}L_0 = {}_{*}\bar{L}_0 = 1$.

Матрицу ${}_{*}M_k = E_{*}L_k$ называем матрицей k-моментов случайного вектора z_{*}, матрицу ${}_{*}\bar{M}_k = E_{*}\bar{L}_k$ – матрицей центральных k-моментов случайного вектора z_{*}. Если все математические ожидания, являющиеся элементами матрицы ${}_{*}M_k$, существуют и конечны, то скажем, что вектор z_{*} имеет матрицу k-моментов. Кратко напишем в этом случае $z_{*} \in \mathcal{H}^k$.

Таким образом, матрицы ${}_{*}\bar{M}_3$ и ${}_{*}\bar{M}_4$ являются матрицами третьих и четвертых центральных моментов вектора z_{*}. В дальнейшем используем следующее расчленение этих матриц в блоки: ${}_{*}\bar{M}_4 = [({}_{*}\bar{M}_4)_{ij}]$, $i,j=0,1,\ldots,s$, где блок $({}_{*}\bar{M}_4)_{ij} = E((x_i-\mu_i)(x_j-\mu_j){}_{*}\bar{L}_2)$ и ${}_{*}\bar{M}_3 = [({}_{*}\bar{M}_3)_i]$, $i=0,1,\ldots,s$, где блок $({}_{*}\bar{M}_3)_i = E((x_i-\mu_i){}_{*}\bar{L}_2)$.

Параметры асимптотического распределения вектора оценок a_{*} определяются следующей теоремой, доказанной в [2].

Теорема 1. Пусть случайный вектор $z_{*} \in \mathcal{H}_4$ и среди компонентов вектора z_{*} нет линейно зависимых. Тогда[4]

---

[3] Символом $\otimes$ обозначено прямое произведение матриц, т.е. если $A \in \mathcal{M}^{m\times n}$, то $A \otimes B = [a_{ij}B]$, $i=1,\ldots,m$; $j=1,\ldots,n$.

[4] Символом $x \xrightarrow{\mathcal{L}} F$ обозначается сходимость по распределению.

$$\sqrt{n}\,(a_* - \varkappa_*) \xrightarrow{\mathcal{L}} N(0,\xi),$$

где

$$\xi = \begin{bmatrix} \sigma_{00}(1-\varrho^2) - 2\mu'\Sigma^{-1}f + \mu'\Pi\mu & f'\Sigma^{-1} - \mu'\Pi \\ \Sigma^{-1}f - \Pi\mu & \Pi \end{bmatrix}, \qquad (4)$$

а $\Pi = \Sigma^{-1}c\Sigma^{-1}$. Элементы c_{ij} матрицы $c \in \mathcal{M}^{s\times s}$ определены равенством $c_{ij} = \gamma'({}_*\bar{M}_4)_{ij}\gamma$, где $\gamma' = [1 \vdots -\varkappa]$. Элементы f_i вектора $f \in \mathcal{M}^{s\times 1}$ определены равенством $f_i = \gamma'({}_*\bar{M}_3)_i\gamma$.

В случае полиномиальной регрессии не все элементы матриц ${}_*\bar{M}_4$ и ${}_*\bar{M}_3$ являются центральными моментами первоначального вектора x_*. Например, если множество I определено равенством (1), т.е. $z = (x_1, x_2, x_1x_2)$, то элемент 1,2 блока $({}_*\bar{M}_4)_{33}$ $m_{3,3,1,2} = E((x_1x_2 - Ex_1x_2)^2(x_1 - Ex_1)(x_2 - Ex_2))$ и таким образом не является центральным моментом вектора $x' = (x_1, x_2)$. Но так как имеют место соотношения (см. [1])

$${}_*\bar{M}_3 = {}_*M_3 - \mu_* \otimes {}_*M_2 - {}_*M_2 \otimes \mu_* - \mathrm{vec}\,{}_*M_2 \otimes \mu_*' + 2\mu_* \otimes \mu_*' \otimes \mu_*, \qquad (5)$$

$$\begin{aligned} {}_*\bar{M}_4 = {}_*M_4 &- \mu_* \otimes {}_*M_3' - {}_*M_3' \otimes \mu_* - \mu_*' \otimes {}_*M_3 - {}_*M_3 \otimes \mu_*' + \\ &+ \mu_* \otimes \mu_* \otimes (\mathrm{vec}\,{}_*M_2)' + \mathrm{vec}\,{}_*M_2 \otimes \mu_*' \otimes \mu_*' + \mu_*' \otimes \mu_* \otimes {}_*M_2 + \\ &+ \mu_*' \otimes {}_*M_2 \otimes \mu_* + \mu_* \otimes {}_*M_2 \otimes \mu_*' + {}_*M_2 \otimes \mu_* \otimes \mu_*' - \\ &- 3\mu_* \otimes \mu_*' \otimes \mu_* \otimes \mu_*', \end{aligned} \qquad (6)$$

то элементы матриц ${}_*\bar{M}_4$ и ${}_*\bar{M}_3$ определяются через моменты первоначального вектора x_*. Предположения теоремы 1 выполнены, если $x_* \in \mathcal{H}^{2k}$.

4. Пример

Просмотрим конкретный пример использования приведенной теоремы. Пусть изучаемый случайный вектор двухмерный, $x_* = (x_0, x_1)$. Компонент x_1 – случайная величина с экспоненциальным распределением, т.е.

$$f_{x_1}(x) = \begin{cases} \lambda e^{-\lambda x}, & x \geq 0 \\ 0, & x < 0, \end{cases} \qquad \lambda > 0.$$

Компонент $x_0 = \frac{\nu+\lambda}{\lambda} e^{-\nu x_1}$, $\nu > 0$. Предположим, что связь между компонентами x_0 и x_1 неизвестно экспериментатору и информацию об этом векторе он может получить только в форме выборки, регистрируя конкретные значения вектора x_*. Он выбирает функцию регрессии из класса квадратных полиномов $f(x) = \alpha_0 + \alpha_1 x + \alpha_2 x^2$ (в данном случае $I = \{1,2\}$). Таким образом, $z_* = (x_0, x_1, x_1^2)$. Используя в формулах (2) и (3) выборочные оценки моментов, можно вычислить оценки коэффициентам полиномиальной регрессии и МКК.

Так как нам известно конкретный механизм изучаемого явления, можно вычислить истинные α_*, ϱ^2 и дисперсионную матрицу асимптотического распределения. Чтобы характеризовать поведение оценки a_* в данной ситуации, проделаем эти вычисления.

Из определения вектора x_* следует, что $x_* \in \mathcal{H}^k$ при всех конечных k. Имеет место соотношение

$$Ex_1^l x_0^s = \frac{(\nu+\lambda)^s}{\lambda^{s-1}} \int_0^\infty x^l e^{-(s\nu+\lambda)x} dx = \frac{(\nu+\lambda)^s}{\lambda^{s-1}} \frac{l!}{(s\nu+\lambda)^{l+1}}. \qquad (7)$$

Отсюда получим

$$\mu_*' = (Ex_0, Ex_1, Ex_1^2) = (1, 1/\lambda, 2/\lambda^2)$$

и

$$\Sigma_* = E(z_* - \mu_*)(z_* - \mu_*)' = \begin{pmatrix} \frac{\nu^2}{\lambda(2\nu+\lambda)} & -\frac{\nu}{\lambda(\nu+\lambda)} & -\frac{2\nu(\nu+2\lambda)}{\lambda^2(\nu+\lambda)^2} \\ -\frac{\nu}{\lambda(\nu+\lambda)} & \frac{1}{\lambda^2} & \frac{4}{\lambda^3} \\ -\frac{2\nu(\nu+2\lambda)}{\lambda^2(\nu+\lambda)^2} & \frac{4}{\lambda^3} & \frac{20}{\lambda^4} \end{pmatrix}.$$

Вектором коэффициентов квадратной регрессии является вектор $\varkappa_* = [\varkappa_0 \vdots \varkappa]$, где

$$\varkappa = -\lambda\nu/2(\nu+\lambda)^2 \begin{pmatrix} 2(3\nu+\lambda) \\ -\lambda\nu \end{pmatrix}$$

и

$$\varkappa_0 = (\lambda^2 + 3\nu\lambda + 3\nu^2)/(\lambda+\nu)^2,$$

таким образом

$$r(z) = \frac{\lambda^2+2\nu(\lambda+\nu)}{(\lambda+\nu)^2} - \frac{2\nu(3\nu+\lambda)}{(\lambda+\nu)^2} x_1 + \frac{\lambda^2\nu^2}{2(\lambda+\nu)^2} x_1^2 .$$

МКК имеет значение

$$\rho^2 = \lambda(2\nu+\lambda)(2\nu^2+2\nu\lambda+\lambda^2)/(\nu+\lambda)^4.$$

Чтобы избежать лишних сложностей в дальнейших вычислениях, ограничимся с случаем $\nu = \lambda = 1$. При таких значениях параметров $\rho^2 = 15/16 = 0{,}9375$ – значение, которое укажет, что выбранная функция является весьма хорошим приближением к компоненту x_0. Но посмотрим теперь дисперсионную матрицу асимптотического распределения. Вычислив при помощи формул (5), (6), (7) матрицы $_*\bar{M}_4$ и $_*\bar{M}_3$ и имея в виду, что при данных значениях параметров $\gamma' = (1, 1, -0{,}125)$, получим

$$\xi = \begin{pmatrix} 3{,}7657 & -8{,}8663 & 2{,}4969 \\ -8{,}8663 & 19{,}7837 & -5{,}5895 \\ 2{,}4969 & -5{,}5895 & 1{,}5850 \end{pmatrix}.$$

Если объем выборки n большой, дисперсионную матрицу оценки можно читать приблизительно равной $1/n\xi$.

Для сравнения вычислим еще дисперсионную матрицу асимптотического распределения коэффициентов регрессии вектора $y'_* = (y_0, y_1, y_2)$, первые и вторые моменты которого совпадают с соответствующими моментами вектора z_* при $\lambda = \nu = 1$, но который имеет нормальное распределение. Используя в [2] полученный результат, можем написать

$$\xi = \sigma_{00}(1-\rho^2)\begin{bmatrix} 1+\mu'\Sigma^{-1}\mu & \mu'\Sigma^{-1} \\ \Sigma^{-1}\mu & \Sigma^{-1} \end{bmatrix} = \begin{pmatrix} 0{,}0625 & 0{,}0625 & 0{,}0104 \\ 0{,}0625 & 0{,}1040 & -0{,}0208 \\ 0{,}0104 & -0{,}0208 & 0{,}0052 \end{pmatrix}.$$

Как увидим, при том самом объеме выборки приближенная оценка $1/n\xi$ дисперсионной матрицы коэффициентов значительно меньше, чем в предыдущем случае.

Но и при векторе x_* точность оценок резко повышается, если экспериментатор догадается провести вместо оценки функции квадратной регрессии преобразование данных и оценить для вектора $w'_* = (w_0, w_1) = (\ln x_0, w_1)$ функцию линейной регрессии. Действительно, в этом случае $w_0 = \ln \frac{\nu+\lambda}{\lambda} - \nu w_1$, $\rho^2 = 1$, и при объеме выборки $n \geq 2$ мы получим точные значения коэффициентов функции линейной регрессии. Вычислив в этом случае дисперсионную матрицу асимптотического распределения коэффициентов, увидим, что этой матрицей является нулевая матрица. Действительно, теперь вектор $\gamma' = (1, \nu)$. Так как $w_0 - Ew_0 =$

$= -\nu(x_1 - 1/\lambda)$, можем написать, обозначив $E(x_1 - 1/\lambda)^k$ через a^k,

$$({}_{w_*}\bar{M}_4)_{ij} = (-\nu)^{2-i-j}\, a^4 \begin{pmatrix} \nu^2 & -\nu \\ -\nu & 1 \end{pmatrix}$$

и

$$({}_{w_*}\bar{M}_3)_1 = (-\nu)^{1-i}\, a^3 \begin{pmatrix} \nu^2 & -\nu \\ -\nu & 1 \end{pmatrix},$$

$i,j=0,1$.

Следовательно

$$\gamma'({}_{w_*}\bar{M}_4)_{ij}\gamma = \gamma'({}_{w_*}\bar{M}_3)_i\gamma = 0 ,$$

$i,j=0,1$. Имея в виду, что $\varrho^2 = 1$, все элементы дисперсионной матрицы асимптотического распределения, определенной формулой (4), равны нулю.

Литература

1. Парринг А.-М., Вычисление асимптотических характеристик функции выборки. Уч. зап. ТГУ, № 429, 86-90.
2. Парринг А.-М., Оценка функции линейной регрессии и ее асимптотическое поведение. Уч. зап. ТГУ, № 429, 91-99.
3. Tiit, E., Parring, A., Mõls, T., Tõenäosusteooria ja matemaatiline statistika. Tallinn, 1977.

СРАВНЕНИЕ РЕГРЕССИИ В РАЗНЫХ ПОДСОВОКУПНОСТЯХ

А.Левисто, Э.Тийт

0. Одной из наиболее часто применяемых методик прикладной математической статистики является линейный регрессионный анализ. Это дает модель в виде формулы

$$y \approx Ey = \beta x \qquad (1)$$

для прогнозирования функционального признака y или функционального признак-вектора $y = (y_1,\dots,y_p)'$ по аргумент-признак-вектору $x = (1,x_1,\dots,x_m)'$. Параметрами прогноза являются коэффициенты регрессии, т.е. элементы матрицы коэффициентов регрессии β

$$\beta = \begin{pmatrix} \beta_{10} & \beta_{11} & \cdots & \beta_{1m} \\ \cdots & \cdots & \cdots & \cdots \\ \beta_{p0} & \beta_{p1} & \cdots & \beta_{pm} \end{pmatrix}. \qquad (2)$$

В одномерном случае матрица β заменяется вектор-строкой $\beta' = (\beta_0,\beta_1,\dots,\beta_m)$.

Довольно типичной с точки зрения анализа данных является следующая задача: Генеральная совокупность, для которой необходимо построить модель, состоит из некоторых подсовокупностей, пусть их число к. Исследователю неизвестно, является ли для всех этих подсовокупностей пригодной единая модель

(1), или необходимо для каждой подсовокупности построить отдельную модель, т.е., вычислить для каждой подсовокупности отдельную матрицу (2).

Аналогичная задача возникает и при исследованиях, проведенных в разные моменты времени: необходимо выяснить, изменяется ли существенно в течение времени модель, описывающая исследуемое явление.

В настоящей статье, исходя из общей теории наименьших квадратов (см. [1]), выведены формулы для эффективного решения поставленной задачи при предположении нормальности рассматриваемых признак-векторов.

Основная часть излагаемой методики реализована в ВЦ ТГУ на ЭВМ ЕС-1022 в виде опытного комплекта программ.

1. Пусть имеется K подсовокупностей генеральной совокупности (или K генеральных совокупностей). В каждой из них измерен признак-вектор[1] $z=(x' : y')'$, состоящий из m аргумент-признаков $x_1,\dots,x_m$ и p функциональных признаков $y_1,\dots,y_p$. Пусть объем k-той выборки – n_k. Обозначаем результаты измерения в k-той совокупности через две матрицы X^k и Y^k (k=1,2,...,K)

$$X^k=\begin{pmatrix} 1 & 1 & \dots & 1 \\ x^k_{11} & x^k_{12} & \dots & x^k_{1n_k} \\ \dots & \dots & \dots & \dots \\ x^k_{m1} & x^k_{m2} & \dots & x^k_{mn_k} \end{pmatrix}, \quad Y^k=\begin{pmatrix} y^k_{11} & y^k_{12} & \dots & y^k_{1n_k} \\ \dots & \dots & \dots & \dots \\ y^k_{p1} & y^k_{p2} & \dots & y^k_{pn_k} \end{pmatrix}.$$

---

[1] Знак ' обозначает транспонирование.

Если $p=1$, то матрица Y^k заменяется на соответствующий вектор $(Y^k)' = (y_1^k,\dots,y_{n_k}^k)$.

Кроме того рассматривают объединенную выборку объема n, $n = \sum_{k=1}^{K} n_k$. Заметим, что объемы подвыборок n_k произвольны, и тем объединенная выборка не должна быть представительным для генеральной совокупности.

Объединенная выборка описывается с помощью матриц X и Y,

$$X = (X^1 \vdots \dots \vdots X^K), \qquad Y = (Y^1,\dots,Y^K), \tag{3}$$

для которых применяется упрощенное обозначение

$$X = \begin{pmatrix} 1 & 1 & \dots & 1 \\ x_{11} & x_{12} & \cdots & x_{1n} \\ \dots & \dots & \dots & \dots \\ x_{m1} & x_{m2} & \cdots & x_{mn} \end{pmatrix}, \quad Y = \begin{pmatrix} y_{11} & y_{12} & \cdots & y_{1n} \\ \dots & \dots & \dots & \dots \\ y_{m1} & y_{m2} & \cdots & y_{mn} \end{pmatrix}.$$

Все статистики, вычисленные по k-той выборке и по объединенной выборке обозначаются соответственно с индексом k вверху или без индекса.

Пусть в k-той генеральной совокупности имеет место соотношение (1) с матрицей коэффициентов регрессии β^k $(k=1,2,\dots,K)$, а выборочная оценка матрицы β^k обозначается через B^k, $B^k = (b_{ij}^k)$ $(i=1,\dots,p;\ j=0,1,\dots,m)$,

$$B^k = Y^k(X^k)'(X^k(X^k)')^{-1}$$

(при предположении регулярности матрицы $X^k(X^k)'$).

Поставленная задача сводится к проверке нулевой гипотезы

$$H_0 : \beta^1 = \beta^2 = \dots = \beta^K.$$

2. Для проверки этой гипотезы вычислим в каждой выборке

по принципу наименьших квадратов остатки прогноза R_0^k и R_1^k. В случае $p>1$ они являются $(p\times p)$-квадратичными положительно определенными матрицами, на главной диагонали которых находятся ошибки прогноза отдельных компонент функционального признак-вектора. В случае $p=1$ R_0^k и R_1^k – квадратичные формы, для которых применяется традиционная символика $(R_0^k)^2$ и $(R_1^k)^2$. Имеем

$$R_0^k=(Y^k-B^kX^k)(Y^k-B^kX^k)'=Y^k(Y^k)'-Y^k(X^k)'(X^k(X^k)')^{-1}X^k(Y^k)'.$$

Если верна нулевая гипотеза, то $\beta^k=\beta$ $(k=1,2,\dots,K)$ и при регулярности матрицы XX' для β существует оценка B, $B=(b_{ij})$ $(i=1,\dots,p;\ j=0,1,\dots,m)$,

$$B=YX'(XX')^{-1}.$$

Остаток прогноза R_1^k при справедливости нулевой гипотезы вычисляется по матрице B:

$$R_1^k=(Y^k-BX^k)(Y^k-BX^k)' \qquad (k=1,2,\dots,K).$$

Отсюда найдутся и суммарные остатки прогноза, при упрощении которых пользуются определения (3) матриц X и Y:

$$R_0=\sum_{k=1}^{K}R_0^k=YY'-\sum_{k=1}^{K}[Y^k(X^k)'(X^k(X^k)')^{-1}X^k(Y^k)'], \qquad (4)$$

$$R_1=\sum_{k=1}^{K}R_1^k=YY'-YX'(XX')^{-1}XY'. \qquad (5)$$

Для получения критерия проверки гипотез предполагаем, что значения аргумент-признак-вектора X фиксированы, а функциональный признак-вектор Y имеет в каждой генеральной совокупности нормальное распределение, $Y^k\sim N_p(\beta^kX^k,\Sigma_Y)$.

При справедливости нулевой гипотезы признак-вектор y имеет и в объединенной генеральной совокупности такое же нормальное распределение, $y \sim N_p(\beta X, \Sigma)$.

3. Если функциональный признак-вектор y одномерен, то задача проверки гипотез точно решается при помощи F-распределения. Действительно, в таком случае отношение

$$\frac{R_1^2 - R_0^2}{R_0^2} \cdot \frac{n - r}{w},$$

где вместо R_1 и R_0 для сохранения традиционной записи написаны R_1^2 и R_0^2, имеет F-распределение с числами степеней свободы $n-r$ и w, где r – ранг матрицы X, n – число наблюдений и w – число независимых параметров, входящих в проверяемую гипотезу. В рассматриваемом случае $r = m+1$, $w = (K-1)(m+1)$.

4. Если же $p > 1$, то следует рассмотреть отношение детерминантов матриц остатков прогноза R_0 и R_1. Известно (см. [1]), что отношение

$$|R_0| \,/\, |R_1|$$

имеет т.н. Λ-распределение, характеризуемое параметрами p, m и n. Для практических целей применимо приближение Λ-распределения через F:

$$\frac{1 - \Lambda^{1/s}}{\Lambda^{1/s}} \cdot \frac{ws - 2v}{mp} \sim F(pm, ws - 2v),$$

где $w = n - (p+m+1)/2$, $s = [(p^2m^2-4)/(p^2+m^2-5)]^{1/2}$, $v = (mp-2)/4$; при нецелочисленных значений величины $ws - 2v$ пользуются округленными до целых чисел значениями.

5. Иногда представляет интерес проверить гипотезу, согласно которой в разных подсовокупностях (в разные моменты времени) модели совпадают до точности свободного члена.

Для формулировки этой гипотезы вводим следующие обозначения: $\beta = (\beta_0 : \beta^*)$, $B = (b_0 : B^*)$, $B^k = (b_0^k : B^{k*})$, где $\beta_0 = (\beta_{10},\dots,\beta_{p0})'$, $\beta^* = (\beta_{ij})$ $(i=1,\dots,p;\ j=1,\dots,m)$, а B, B^k, B^* и B^{k*} – оценки матриц β и β^* соответственно по объединенной и k-той выборке. Аналогично, пусть $x' = (1, x^{*'})$, $x^* = (x_1,\dots,x_m)'$ и соответственно выборки разделяются в блоки: $X = \begin{pmatrix} 1' \\ \dots \\ X^* \end{pmatrix}$, $X^k = \begin{pmatrix} 1' \\ \dots \\ X^{k*} \end{pmatrix}$, где 1 – вектор-столбец, состоящий из единиц, а $X^* = (x_{ij})$, $i=1,\dots,m;\ j=1,\dots,n$, $X^{k*} = (x_{ij}^k)$, $i=1,\dots,m;\ j=1,\dots,n_k$.

Проверяемая гипотеза в таком случае имеет вид

$$H_0^* : \beta^{1*} = \beta^{2*} = \dots = \beta^{K*} .$$

Для проверки этой гипотезы придется снова вычислить остаток прогноза при справедливости нулевой гипотезы.

Так как

$$R_1^{k*} = (Y^k - b_0^k - B^* X^{k*})(Y^k - b_0^k - B^* X^{k*})',$$

то

$$R_1^* = \sum_{k=1}^{K} R_1^{k*} = \sum_{k=1}^{K} (Y^k - b_0^k - B^* X^{k*})(Y^k - b_0^k - B^* X^{k*})'.$$

При проверке гипотезы H_0^* при помощи F- или Λ-распределений следует число m заменить числом $m^* = m - 1$.

6. Вторая модификация поставленной задачи состоит в сравнении подмоделей.

Пусть I – некоторое подмножество множества $(0,1,\dots,m)$, $I = (i_1, i_2,\dots, i_s)$. Обозначаем через $x(I)$ подвектор

$(x_{i_1},\ldots,x_{i_s})'$ вектора x и через $X(I)$ подматрицу $(x_{i_1}:\ldots:x_{i_s})'$ матрицы X, где $x_{i_h} = (x_{i_h1},\ldots,x_{i_hn})$ – i_h-ая строка матрицы X. Аналогично определяются и матрицы $X^k(I) = (x^k_{i_h,j})$ $(h=1,\ldots,s;\ j=1,\ldots,n_k,\ k=1,\ldots,K)$. Также вводится подматрица $\beta(I)$ матрицы β, $\beta(I) = (\beta_{i,j_h})$ $(i=1,\ldots,p;\ h=1,\ldots,s)$.

Подмоделью называется соотношение

$$y \approx E(y) = \beta(I)x(I),$$

в котором участвуют только аргумент-признаки $x_{i_1},\ldots,x_{i_s}$. Подмодель в k-той генеральной совокупности имеет вид:

$$y^k \approx E(y^k) = \beta^k(I)x^k(I).$$

Задача сравнения подматриц состоит в проверке нулевой гипотезы

$$H_0(I) : \beta^1(I) = \beta^2(I) = \ldots = \beta^K(I).$$

В случае $s=1$ данная методика позволяет проверять значимость каждого отдельного аргумент-признака в модели. Пошаговая реализация методики сравнения подмоделей позволяет выяснить оптимальный комплект аргумент-признаков, вызывающий различие моделей в разных подмножествах (в разные моменты времени).

При проверке гипотез следует отметить, что вместо $m+1$ придется применить число разных параметров s.

7. Если через J обозначается подмножество множества индексов всех исследуемых совокупностей, $J = (j_1,\ldots,j_t)$, $J \subset (1,\ldots,K)$, то простой модификацией описанной методики возможно проверять нулевую гипотезу

$$H_0(J) : \beta^{j_1} = \beta^{j_2} = \ldots = \beta^{j_t}.$$

При вычислении параметров F- и Λ-распределений следует к

заменить на t.

Пошаговое применение этой методики полезно для выяснения самых сильно отличающих друг от друга и от других подсовокупностей с точки зрения исследуемой модели.

8. В задачах анализа данных нередко целесообразно рассмотреть в качестве исходных данных не матрицы наблюдений X^k и X, а вычисленные по ним статистики - результаты первичной обработки. Такая ситуация имеет место, например, при весьма больших массивах данных, при подмассивах, исследуемых в разные времена или в разных местах.

Типичными результатами первичной обработки являются

а) вектор-столбец оценок средних значений исследуемых признаков в k-той выборке $\bar{z}^k$, $\bar{z}^k = (1, \bar{x}_1^k, \ldots, \bar{x}_m^k, \bar{y}_1^k, \ldots, \bar{y}_p^k)'$;

б) выборочная ковариационная матрица S^k признак-вектора z в k-той выборке,

$$S^k = \left(\begin{array}{c|c} S_x^k & S_{xy}^k \\ \hline S_{yx}^k & S_y^k \end{array}\right);$$

в) выборочная корреляционная матрица R^k признак-вектора z в k-той выборке,

$$R^k = \left(\begin{array}{c|c} R_x^k & R_{xy}^k \\ \hline R_{yx}^k & R_y^k \end{array}\right).$$

Соответствующие оценки $\bar{z}$, S^k и R^k по объединенной выборке, как правило, не известны.

Их легко найти по оценкам в отдельных выборках:

$$\bar{z} = \frac{1}{n} \sum_{k=1}^{K} n_k \bar{z}^k ,$$

$$S = \frac{1}{n}\sum_{k=1}^{K} n_k S^k + \mathcal{E},$$

где

$$\mathcal{E} = \frac{1}{n}\sum_{k=1}^{K-1}\sum_{s=k+1}^{K} n_k n_s (\bar{z}^k - \bar{z}^s)(\bar{z}^k - \bar{z}^s)',$$

и $$R = (\operatorname{diag} S)^{-1/2}\, S (\operatorname{diag} S)^{-1/2}$$

(здесь $\operatorname{diag} A$ обозначает диагональную матрицу, в которой на диагонали находятся диагональные элементы матрицы A, а D^a при диагональной D есть матрица, в которой на диагонали элементы d_i^a, где d_i – соответствующие элементы матрицы D).

9. При помощи выборочных ковариационных или корреляционных матриц остатки регрессии (4) и (5) вычисляются следующим образом.

В k-той подсовокупности имеем:

$$b^k = S_{yx}^k (S_x^k)^{-1}$$

и

$$R_0^k = S_y^k - S_{yx}^k (S_x^k)^{-1} S_{xy}^k,$$

откуда получается суммарный остаток прогноза R_0:

$$R_0 = \sum_{k=1}^{K} \left[S_y^k - S_{yx}^k (S_x^k)^{-1} S_{xy}^k \right]. \qquad (6)$$

При справедливости нулевой гипотезы получим R_1:

$$R_1 = S_y - S_{yx}(S_x)^{-1} S_{xy}. \qquad (7)$$

Аналогичный результат можно получить и при помощи корреляционных матриц:

$$\begin{cases} \tilde{R}_0 = \sum_{k=1}^{K} \left[R_y^k - R_{yx}^k (R_x^k)^{-1} R_{xy}^k \right], \\ \tilde{R}_1 = R_y - R_{yx} (R_x)^{-1} R_{xy} . \end{cases} \qquad (8)$$

Хотя остатки прогноза $\tilde{R}_0$ и $\tilde{R}_1$, вычисленные по формулам (8) не совпадают с величинами R_0 и R_1, найденными по формулам (6) и (7), отношение детерминантов, используемое при проверке гипотез, сохраняется.

10. Применение выборочных ковариационных или корреляционных матриц при вычислении регрессии полезно при неполных данных. Тогда все коэффициенты ковариации s_{ij}^k (корреляции) оценивают по всем наблюдениям из k-той совокупности, при которых i-ый и j-ый признак измерены. В таком случае получается оценка, которая максимально использует информацию, содержащуюся в выборке. При случайном расположении пробелов полученная оценка S^k является несмещенной, и полученные по ней оценки коэффициентов регрессии вполне применимы при анализе данных [2, 3].

Литература

1. Рао С.Р., Линейные статистические методы и их применения. Москва, 1968.
2. Тийт Э., Обработка неполных данных. Труды ВЦ ТГУ, 1977, 40, 14-20.
3. Koskel, S., Tiit, E., Regressioonanalüüs. Programme kõigile XIV, 1978, 43-49.

КЛАССИФИЦИРОВАНИЕ НАУЧНЫХ ДАННЫХ С ПОЛУЧЕНИЕМ ЧАСТИЧНО ПОКРЫВАЕМЫХ КЛАССОВ

Р. Ээремаа

В настоящей статье мы затрагиваем проблемы классифицирования эмпирического материала, накопленного в ходе научного исследования, как биолого-медицинского, социолого-психологического, экономического, педагогического и т.п. характера.

Целью классифицирования научных данных является раскрытие сущности и закономерностей исследуемого материала с помощью представления массив информации в сжатом виде так, чтобы потеря информации не была чрезмерной.

Характерными чертами классифицирования научного материала являются: отсутствие обучающих выборок и априорной информации о характере распределения исследуемых признаков внутри классов и отсутствие жестких требований на результат классифицирования; но существование некоторых интуитивных представлений о результате (таких, как: разумное число классов и принадлежность некоторых объектов в один и тот же класс) - такая интуиция базируется на теоретических и на профессиональных соображениях.

Отметим, что вышеописанную задачу классификации можно назвать и задачей таксономии, а получаемые классы таксонами.

В связи с тем, что цель классифицирования здесь недостаточно четко определена, формализировать такую задачу нелегкая проблема.

1. Общие требования на постановку и решение задачи автоматической классификации научного материала

1.1. Разделение задач автоматической классификации

На основе работы [2] все множество задач автоматической классификации можно разбить на несколько основных групп:

1) по их постановке - на детерминистские и стохастические (или вероятностные) задачи;

2) по способу решения - на иерархические и неиерархические;

3) по наличию или отсутствию априорной информации о числе классов;

4) по тому, основывается ли решающий алгоритм непосредственно на состояния объектов (т.е. на измерениях обследуемых признаков на каждом объекте) или на понятии меры близости (т.е. меры различия или сходства) между объектами.

Вышеприведенное разделение мы берем за основу при конкретизации постановки и решения задачи автоматической классификации научного материала.

1.2. Постановка задачи классифицирования научного материала

При задаче классифицирования научных данных обычно отсутствует такая достоверная априорная информация, которая нуждается в стохастическом подходе. Подходящим является здесь детерминистский подход, при котором исследуемый материал не

рассматривают как выборку из некоторой гипотической совокупности, а как самое совокупность. При таком подходе не существует эвристического этапа выбора гипотезы и "догадки" о закономерностях генеральной совокупности по конечной выборке.

О виде искомых таксонов надо сказать следующее. В ряде экспериментов было обнаружено, что при решении задачи таксономии человек использует простейшие (линейные) решающие функции. Поэтому, в работе [3] делается вывод, что если таксоны предназначены для непосредственного восприятия человеком, они должны иметь простую форму в виде гиперсферы.

Однако, то обстоятельство, что человек сам может осуществлять классифицирование в пространстве большой размерности только с помощью простейших решающих функций, не означает, что он не способен принять и понять более сложную классификацию.

Обычно эмпирический материал, накопленный в ходе научного исследования, обладает настолько "плохой" структурой, что неоправдан поиск классов простой формы с четко определенными классами. Хотя результат классифицирования научного материала должен быть упрощенным изложением материала, он кроме того должен довольно точно характеризовать и его внутреннюю структуру. Если на таком материале использовать методы нахождения непересекаемых классов простой формы, исследователь замечает при интерпретации получаемого результата, что расчленение некоторых объектов на разные таксоны с одной стороны является в соответствии с его теоретическим представлением о исследуемом материале, а с другой – эти таксоны заключают в себе "сомнительные" элементы, которые своим наличием

не поддерживают его гипотезу. Такое противоречие не позволяет ему ни принять новую гипотезу, ни опровергнуть старую.

При нахождении частично покрываемых таксонов произвольной формы такие "сомнительные" объекты обычно являются покрываемыми элементами между таксонами (на основе существующей информации они принадлежат к нескольким таксонам).

Классифицирование с получением пересекаемых классов соответствует и процессу мышления. Человек не всегда способен делать твердые, неопровержимые решения. На некотором уровне принятия решения он не в состоянии предпочесть одно решение другому, а при добавлении информации он иногда может такое предпочтение легко сделать. Отметим, что при классифицировании с получением частично покрываемых классов, добавочной информацией для решения вопроса о принадлежности "сомнительных" объектов к какому-то классу или нет, может быть и некоторая профессиональная информация исследователя (передать которую ЭВМ трудно).

Итак, задачу классифицирования научных данных поставим как детерминистскую задачу для нахождения классов произвольной формы, которые могут частично покрываться.

1.3. Решение задачи классифицирования научного материала

Эмпирический материал научных исследований по существу таков, что в нем разумно искать довольно однородные подмножества элементов, которые сравнительно изолированы от членов других подмножеств, и которые в то же время в свою очередь развиваются на сравнительно однородные, довольно изолированные друг от друга, подмножества.

Для того, чтобы получить точный и полный анализ исследуемого материала, с точки зрения учета его внутренней структуры, надо использовать иерархические процедуры. Иерархическое классифицирование дает исследователю возможность наглядной интерпретации проведенного анализа, но также и возможность выбирать уровень и соответствующее разбиение объектов в иерархическом дереве, исходя из своих теоретических соображений и из многих показателей, характеризующих разбиение.

При осуществлении иерархического классифицирования, которое основано на вышехарактеризованном представлении о классе как о множестве, не имеющего строго определенных границ, трудно представить иерархическое дерево традиционным образом. Поэтому, в таком случае результатом иерархического классифицирования представляются по очереди разбиения объектов на разных уровнях. Есть возможность выбирать на ЭВМ из таких уровней только один уровень (или несколько уровней) классификации соответственно каким-то критерием качества разбиения, в том числе соответственно наперед заданным числом классов. При частичной покрываемости между классами последняя значит задавание числа таких классов, в составе которых относительно мало покрываемых элементов.

1.4. Априорная информация о числе классов при задаче классифицирования научного материала

Ясно, что если мы реализуем классифицирование научного материала иерархическим процессом, то мы вместе с тем позволяем и задавание желаемого числа классов. Но в данном случае число классов является только одним из показателей, характе-

ризующих разбиение на каком-то уровне, а не представляет собой твердое условие на результат классифицирования, - т.е. процедур классифицирования не осуществляют, исходя из этого требования. Насколько произвольным может быть результат, когда в классифицировании исходят из наперед заданного числа классов, видно из примера, приведенного в следующей главе этой статьи.

1.5. Мера близости между объектами при задаче классифицирования научного материала

Мы еще не установили, должен ли решающий алгоритм основываться непосредственно на измерениях обследуемых признаков на объектах или на мере близости между объектами. На базе работ, в которых занимаются классифицированием первого вида, можно сказать, что обычно такое классифицирование осуществляется информационно-теоретическими средствами. Такая реализация сопровождается громоздкостью соответствующих процедур, большим удельным весом эвристических расчетов и необходимостью в емкой памяти ЭВМ.

При задаче классифицирования научного материала за основу берем подход, согласно которому описание любого объекта измерением m признаков интерпретируется геометрически в виде точки в m-мерном пространстве признаков. Естественно предположить, что геометрическая близость точек в этом пространстве означает сходство физических состояний соответствующих объектов. Поэтому, в данном подходе важное место приобретает понятие и мера близости объектов. Можно сказать, что, если имеется абстрактное множество, то введение понятия "близости"

между его элементами соответствует заданию в нем некоторой топологии. Тем самым пространство описаний превращается в топологическое пространство. Чтобы определить качество классификации, надо ввести функцию, которая указывала бы на степень согласованности данной классификации с введенной топологией.

Близость измерят либо мерой различия, либо мерой сходства (о дефиниции таких мер например [1]). Надо отметить, что выбор такой меры на практике носит субъективный характер. Но как в работах [4] и [11] подчеркивается, при использовании одного и того же самого алгоритма классифицирования, самые разные меры различия (сходства) дают в практике почти одинаковые результаты. Но все-таки надо предпочитать такие методы классифицирования, которые позволяют наложение разной меры различия (сходства), определяющейся структурой признакового пространства, методам, позволяющим наложение только функции расстояния, как это при выработанных методах часто бывает.

1.6. Процесс классифицирования научного материала и требования к методу классифицирования

Подытоживая теперь вышеприведенные обсуждения о классифицировании научного материала, можем сказать следующее. Процесс классифицирования будем рассматривать двухэтапно: к первому этапу относится конструирование матрицы различия (сходства) между классифицируемыми объектами; на втором этапе происходит конструирование классифицируемой системы, в которой объединяются объекты при различных уровнях различия (сходства), при том получаемые классы могут быть как непе-

ресекаемыми, так и частично покрываемыми. Такой двухэтапный процесс мы будем называть кластер-анализом.

Пусть у нас для конкретности данные для конструирования кластер-системы представлены в виде коэффициента различия (КР) над всеми исследуемыми объектами. Пусть их число n. Переход от КР к требуемому виду кластер-системы осуществляется применением кластер-метода к КР. При кластеризации научного материала разумно требовать от кластер-метода удовлетворения следующих условий:

1) независимость кластеризации от упорядочения объектов;

2) независимость кластеризации от шкалы;

3) если в данных некоторые кластеры четко вырисованы, то они должны сохраниться и после применения кластер-метода к исходному КР;

4) получаемый результат должен быть в некотором смысле самым лучшим при наложенных требованиях;

5) малым изменениям в данных должны соответствовать малые изменения в результатах.

Отметим, что в работах [9] и [12] представлена модель таксономии, в рамках которой можно математически корректно поставить и решить задачу классифицирования научного материала, так как все наши требования к классифицированию в нем удовлетворены.

2. Пример применения двух различных алгоритмов классифицирования, один из которых найдет классы, могущие частично покрываться

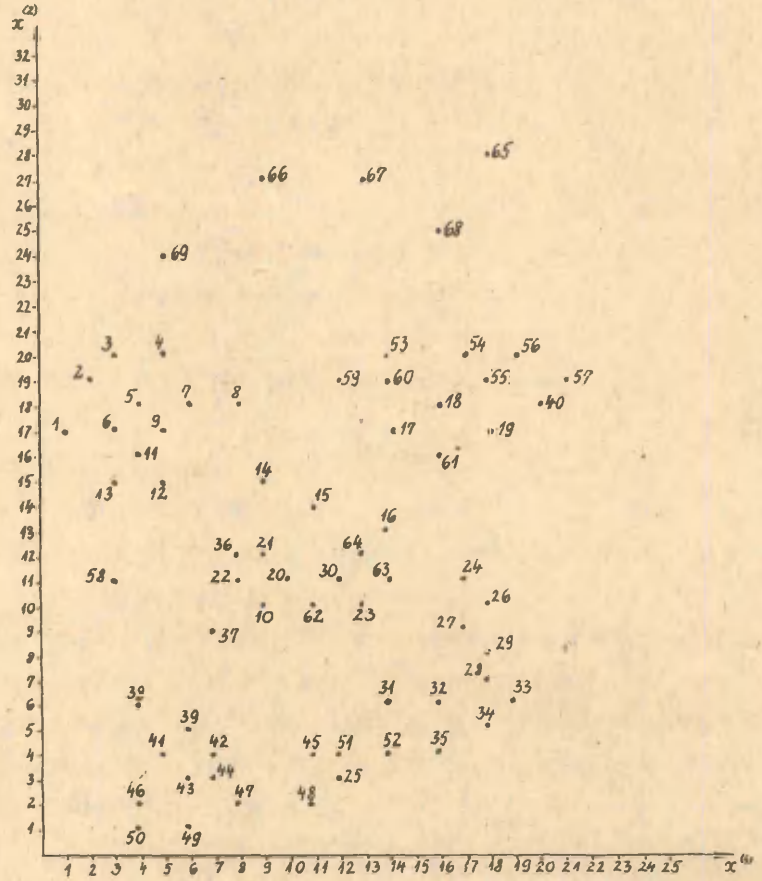
2.1. Исследуемая совокупность объектов и применяемые алгоритмы классифицирования

Пусть задана 69-элементная совокупность объектов, каждый из которых представлен в виде 2-мерного вектора измерений $X = (x^{(1)}, x^{(2)})$. В данном случае каждый объект можно считать точкой двумерного пространства. Соответствующие точки на плоскости могут быть расчленены человеком даже невооруженным глазом. Поэтому, на базе данной совокупности можно сравнивать как результаты применения разных методов классифицирования между собой, так и получаемые результаты с "истинным" расчленением объектов.

На рис. 1 исследуемые 69 объектов представлены в виде точек на плоскости. На этом рисунке визуально видно 5 - 6 скоплений наиболее близко расположенных друг от друга точек.

Демонстрируем, как выполняемость или невыполняемость некоторых вышехарактеризованных требований к классифицированию влияют на результат классифицирования. Мы сравниваем результаты при применении двух различных алгоритмов классифицирования, реализирующих детерминистский подход с использованием меры близости между объектами. Первым является иерархический алгоритм ППК [6-7] для осуществления k-кластеризации, т.е. для получения кластеров, которые могут покрываться в объеме до k-1 объектов, принимая $1 \leq k \leq n-2$, где n - число исследуемых объектов. Вторым возьмем неиерархический алгоритм

Рис. 1. Представление 69 объектов точками на плоскости.



"ФОРЭЛЬ-2" [5]. В данном случае оба алгоритма используют мерой различия межобъектное эвклидовое расстояние. Отметим, что вообще алгоритм ШК может работать с любой мерой различия или сходства, в отличие от алгоритма "ФОРЭЛЬ-2", в котором исходными данными для построения классификации используют только межобъектные эвклидовые расстояния. При использовании алгоритма "ФОРЭЛЬ-2" надо задавать желаемое число классов, а при алгоритме ШК такая информация не нужна.

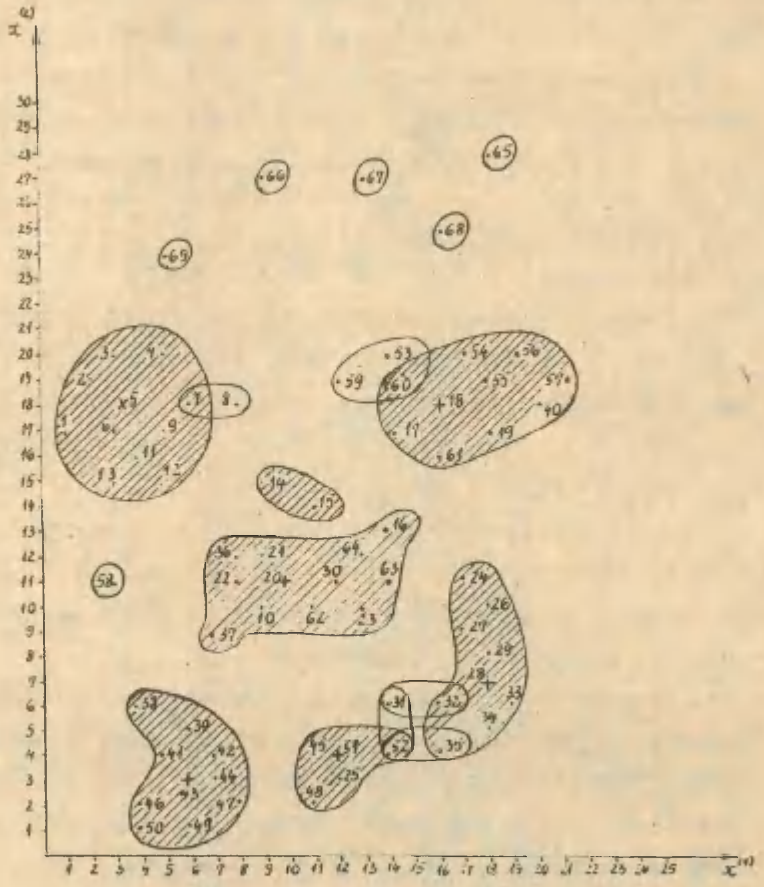
2.2. Применение алгоритма ШК

Применяя алгоритм ШК, можно получить кластер-систему, в которой объединяются объекты на нескольких уровнях различия (сходства). Есть возможность, что на ЭВМ выбирается из них т.н. уровень классификации, при котором получаемая кластер-структура в некотором смысле наиболее точно характеризует структуру данных.

На рис. 2 мы представляем результат 2-кластеризации алгоритмом ШК на уровне классификации, выбранном на ЭВМ. Отметим, что при 2-кластеризации получаемые кластеры могут частично покрываться до одного элемента.

В алгоритме ШК для каждого кластера вычисляется определенная мера внутренней связанности, которая может принимать значения в полуотрезке (0,1]. При том значение 1 соответствует случаю, когда в полученном кластере ни один элемент не является покрываемым. На рис. 2 сильно-связанные кластеры (для которых в данном случае численные значения меры связанности превышают 0,833) представлены заштрихованными. Также на рис. 2 крестиками отмечены т.н. типичные представители

Рис. 2. Получаемые кластеры 2-кластеризации алгоритмом ППК на уровне классификации, выбранном на ЭВМ.



кластеров (число элементов которых больше двух). В качестве типичного представителя выбирается элемент, максимально связанный с остальными в данном кластере (на базе исходной матрицы различия или сходства). При применении алгоритма ШК результатом получается 6 одноэлементных кластеров и 12 кластеров, число элементов которых больше одного.

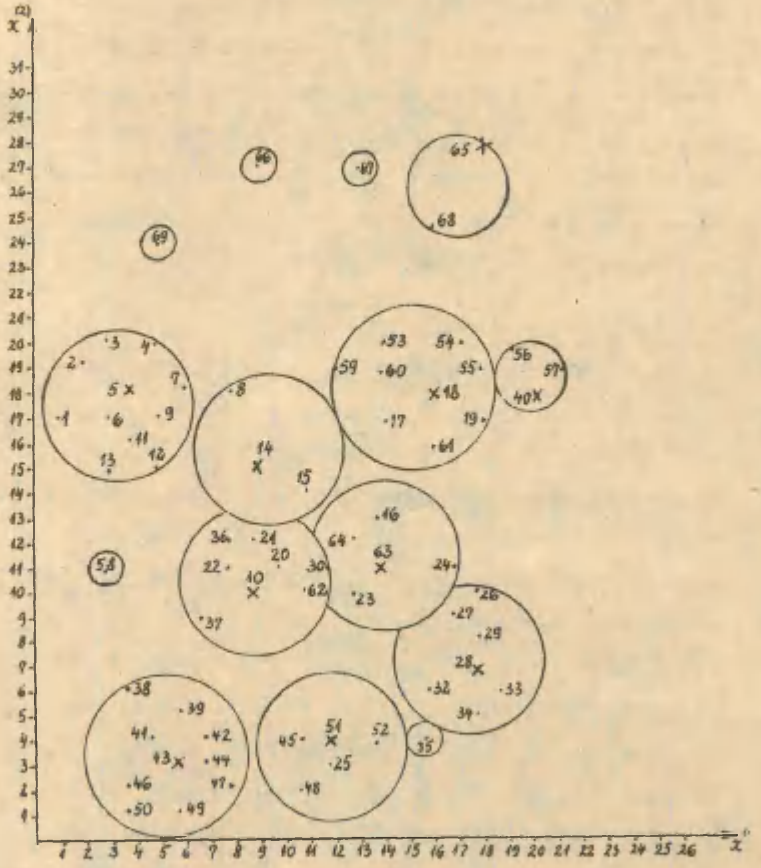
Отметим, что алгоритм ШК реализует метод классифицирования, который удовлетворяет всем требованиям, приведенным на стр. 62. Особое внимание обратим на то, что четковырисованные кластеры на рис. 1 сохранены после применения кластер-метода (см. рис. 2), и на то, что упорядочение объектов не изменит результат кластеризации.

2.3. Применение алгоритма "ФОРЭЛЬ-2"

Для сравнения результатов разных методов классификации применяем одну программу из хорошоизвестных алгоритмов группы "ФОРЭЛЬ", а именно "ФОРЭЛЬ-2". Алгоритм "ФОРЭЛЬ-2" произведет расчленение совокупности объектов гиперсферами, - здесь реализуется характеризованная выше идея о конструировании таксонов простой формы. В информации к этой программе надо задавать желаемое число таксонов.

Кажется, что результат классифицирования, похожий на разбиение, найденное алгоритмом ШК, можно получить при задавании наперед числа (непересекаемых) таксонов 15. Исходя из такого задавания, применение алгоритма "ФОРЭЛЬ-2" дает результатом разбиение на 15 таксонов, представленное на рис. 3. Для каждого выделенного таксона происходит поиск его типичной точки, т.е. точки, являющейся наиболее близкой к цен-

Рис. 3. Результат классифицирования исследуемых объектов алгоритмом "ФОРЭЛЬ-2" при заданном числе таксонов 15.



тру таксона. Эти типичные представители таксонов отмечены крестиками.

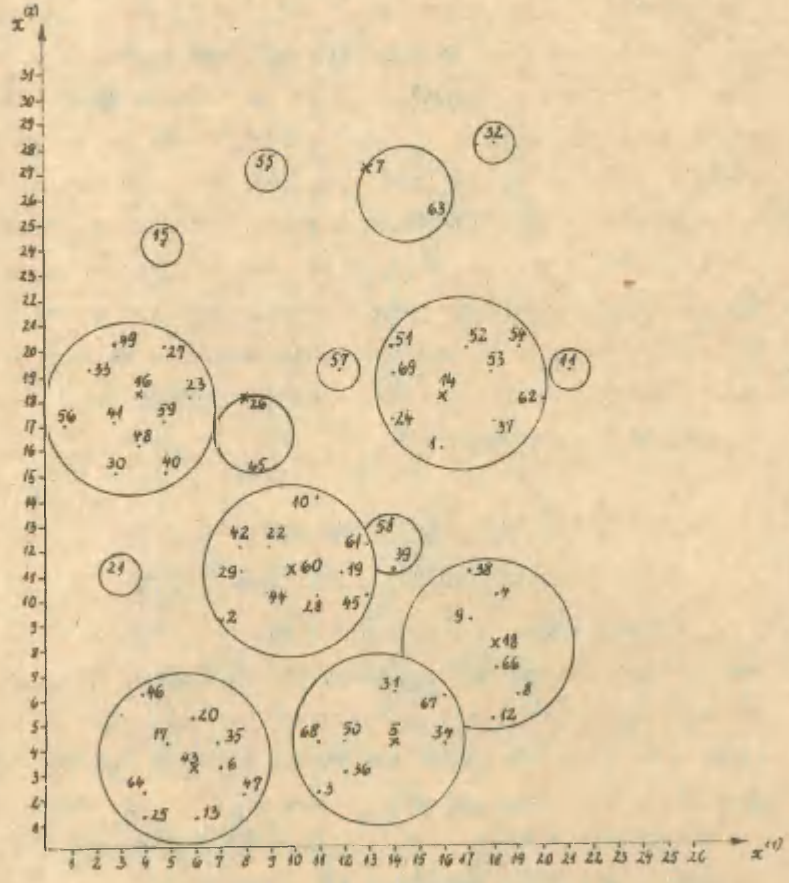
На рис. 3 видно, что таксоны, четко вырисованные на рис. 1, не сохранились после применения алгоритма "ФОРЭЛЬ-2". Это и естественно, потому что на метод классифицирования, который реализуется алгоритмом "ФОРЭЛЬ-2", не наложено требование о сохранении четко вырисованных таксонов в данных. Отметим, что на соответствующий метод не наложены ни требования, чтобы малым изменениям в данных соответствовали малые изменения в результатах, ни требования независимости результата от упорядочения объектов. Невыполняемость последнего условия иллюстрируем на рис. 4. На этом рисунке представлен результат классифицирования в случае изменения упорядоченности исследуемых объектов, и при неизмененном желаемом числе таксонов 15. Видно заметное различие между классификациями, представленные на рисунках 3 и 4.

3. Подходящий метод классифицирования при конкретной задаче классифицирования

При применении алгоритма "ФОРЭЛЬ-2" мы увидели, что субъективное задание наперед числа искомых таксонов, принесет за собой и результат, субъективный в своей сущности. Поэтому алгоритм, в исходной информации которого задается желаемое число классов, не всегда дает хороших результатов в смысле наиболее точного отражения структуры исследуемой совокупности объектов в найденной классификации.

Вообще, при использовании любого метода классифицирова-

Рис. 4. Результат классифицирования в случае изменения упорядоченности исследуемых объектов, алгоритмом "ФОРЭЛЬ-2" при заданном числе таксонов 15.



ния, найденного в литературе, надо иметь в виду цель, на достижение которой соответствующее классифицирование направлено. С поставленной целью тесно связаны и требования, выполнение которых от метода классифицирования требуется. Так, например, мы требуем от метода классифицирования, который мы применяем на научные данные, выполнения ряда естественных условий (см. стр. 62). Но такие же требования не выполнены при любом публицированном методе классифицирования. Изложенный выше пример указывает, что принесет за собой, например, невыполнение условия независимости результата от упорядочения объектов.

Наш пример также показал, что при применении алгоритма ШК получаемые классы (которые могли частично покрываться) довольно хорошо соответствовали нашим представлениям об искомой классификации. Отметим, что в литературе часто можно найти такие методы классифицирования, критерием качества которых служит только хорошая интерпретация найденных классов, согласованность полученной классификации с теоретическими представлениями при конкретной исследуемой совокупности данных. Довольно редко в литературе встречаем такой математически строгий подход к конструированию алгоритма классифицирования, который был основой для выработки алгоритма ШК. Этот подход содержит следующие этапы:

1) точное математическое определение данных и результата классифицирования, позволяющее рассматривать метод классифицирования как преобразование из одной структуры в другую;

2) определение требований на такое преобразование;

3) определение существования, единственности и свойств метода, удовлетворяющего наложенным требованиям;

4) нахождение эффективного алгоритма.

Итак, нахождение пересекаемых классов при классифицировании научного материала оправдано и возможно, - что свидетельствует о существовании такого нетрадиционного способа классифицирования рядом с традиционным способом (нахождением непересекаемых классов). Подчеркиваем еще раз, что при классифицировании научного материала в качестве исходной информации не используются интуитивные представления исследователя об ожидаемом результате классифицирования (в том числе о желаемом числе классов). Мы реализуем математически строго представляемый метод классифицирования, который соответствует всем наложенным нами требованиям. Но так как мы притом требуем сохранения четко вырисованных кластеров, то получаемый результат, отражающий наиболее точно структуру совокупности объектов, в большинстве случаях совпадает и с интуитивными ожиданиями исследователя.

Литература

1. Дюран Б., Оделл П., Кластерный анализ. М., "Статистика", 1977.
2. Елисеева И.И., Рукавишников В.О., Группировка, корреляция, распознавание образов. М., "Статистика", 1977.
3. Загоруйко Н.Г., Методы распознавания и их применение. М., "Сов. Радио", 1972.

4. Ивахненко А.Г., Самообучающиеся системы распознавания и автоматического управления. Киев, "Техника", 1965.

5. Математическое обеспечение ЕС ЭВМ. Минск, 1976, 10, 6, 141-143.

6. Ээремаа Р., Общая теория конструирования кластер-систем и алгоритмы для нахождения их численных представлений. Труды ВЦ ТГУ, 1978, 42, 53-77.

7. Ээремаа Р., Алгоритм опознания кластеров к-дендрограммы. Труды ВЦ ТГУ, 1978, 42, 78-93.

8. Jardine, C.J., Jardine, N., Sibson, R., The structure and construction of taxonomic hierarchies. Math. Biosciences, 1967, 1, 173-179.

9. Jardine, N., Sibson, R., A model for taxonomy. Math. Biosciences, 1968, 2, 465-482.

10. Jardine, N., Sibson, R., The construction of hierarchic and non-hierarchic classifications. The Computer Journal, 1968, 11, 177-184.

11. Lerman, J.C., Les bases de la classification automatique. Paris Collection "Programmation", 1970.

12. Sibson, R., A model for taxonomy II. Math. Biosciences, 1970, 6, 405- 430.

ДВУХМЕРНЫЙ СТАТИСТИЧЕСКИЙ АНАЛИЗ ДАННЫХ

Ю. Вилисмяэ

Важной частью системы статистического анализа данных в ВЦ ТГУ является т.н. двухмерный статистический анализ.

Комплект соответствующих программ работает с 1976 года как часть системы для ЭВМ "Минск-32". Теперь аналогичная методика реализована и для ЭВМ ЕС-1022 для работы в операционной системе ДОС.

Ниже излагается краткая характеристика этой методики, причем предполагается знание основных принципов системы статистической обработки данных (см. [1]).

1. Общие положения

Пусть задана совокупность исследуемых признаков $Z = \{X_1,\dots,X_m\}$. Двухмерный статистический анализ этих признаков состоит в совместном исследовании каждого признака из одной группы признаков

$$Z_1 = \{X_{i_1},\dots,X_{i_r}\}$$
$$(X_{i_s} \in Z;\quad 1 \le i_s \le m;\ s=1,\dots,r)$$

вместе с каждым признаком из второй группы

$$Z_2 = \{X_{k_1}, \ldots, X_{k_p}\}$$

$$(X_{k_t} \in Z; \quad 1 \leq k_t \leq m; \quad t=1,\ldots,p),$$

причем эти группы могут частично или полностью совпадать.

Основным элементом двухмерного статистического анализа является анализ одной пары признаков ($X_i \in Z_1$; $X_j \in Z_2$). Это означает образование таблицы частот и вычисление эмпирического и условных распределений рассматриваемых признаков X_i и X_j, а также вычисление некоторых параметров этого распределения, в том числе показателей связи.

Двухмерный статистический анализ всех исследуемых признаков осуществляется в ЭВМ при помощи последовательности анализов отдельных пар признаков. В конце анализа коэффициенты связи при надобности соединяются в (r x p)-матрицу и выпечатываются.

Для простоты в дальнейшем будем обозначать $X_i = X$ и $X_j = Y$.

2. Совместное исследование пары признаков X и Y

Пусть заданы выборки значений $(x_1, \ldots, x_N)$ и $(y_1, \ldots, y_N)$ признаков X и Y соответственно. Предположим, что признак X имеет k классов распределения: $A_1, \ldots, A_k$, а Y l классов: $B_1, \ldots, B_l$ $(k,l \geq 1)$.

Классом распределения в принципе может быть любое связное множество значений данного признака. Разделение области значений на классы задается при помощи данных классифицирования (см. [4], стр. 30-42). Классы выбираются непересекающими друг друга, причем ими необязательно охватывать все

элементы выборки.

Рассмотрим теперь отдельные части листинга, полученного в результате анализа пары признаков X и Y.

2.1. Заголовок. Выпечатываются имена и номера признаков X и Y, уровень значимости и другие общие параметры обработки.

2.2. Таблица частот. В таблице частот печатается матрица $\{n_{ij} \mid i=1,\dots,k;\ j=1,\dots,l\}$, где n_{ij} – количество объектов (индивидов), попавших одновременно в класс A_i признака X и в класс B_j признака Y.

2.3. Таблица совместного эмпирического распределения признаков. В таблице печатаются относительные частоты (в процентах) $\left\{\frac{n_{ij}}{n}\,100\right\}$ $(i=1,\dots,k;\ j=1,\dots,l)$, где $n=\sum_{i=1}^{k}\sum_{j=1}^{l} n_{ij}$. (Очевидно, $n \le N$.)

2.4. Параметры распределения. Вычисляются эмпирические средние:

$$\bar{x} = \frac{1}{n}\sum_{i} x_i, \qquad \bar{y} = \frac{1}{n}\sum_{i} y_i$$

и эмпирические стандартные отклонения:

$$s_X = \sqrt{\frac{1}{n-1}\sum_{i} (x_i-\bar{x})^2}, \qquad s_Y = \sqrt{\frac{1}{n-1}\sum_{i} (y_i-\bar{y})^2}$$

(в них и в формуле вычисления коэффициента линейной корреляции, суммирование производится по всем значениям индекса i при которых $x_i \in \bigcup_{s=1}^{k} A_s$ и $y_i \in \bigcup_{t=1}^{l} B_t$), а также доверительные границы эмпирических средних при заданном уровне значимости α:

$$\bar{x} - \frac{S_X t_\alpha}{\sqrt{n}}, \quad \bar{x} + \frac{S_X t_\alpha}{\sqrt{n}} \quad \text{и} \quad \bar{y} - \frac{S_Y t_\alpha}{\sqrt{n}}, \quad \bar{y} + \frac{S_Y t_\alpha}{\sqrt{n}},$$

где t_α – соответствующий квантиль t-распределения Стьюдента.

Кроме того вычисляются параметры связи:

коэффициент линейной корреляции:

$$r = \frac{\sum_i (x_i - \bar{x})(y_i - \bar{y})}{(n-1)\, S_X\, S_Y},$$

χ^2-статистика:

$$\chi^2 = n\left(\sum_{i=1}^{k}\sum_{j=1}^{l} \frac{n_{ij}^2}{n_{i.} n_{.j}} - 1\right),$$

где $n_{i.} = \sum_{j=1}^{l} n_{ij}$ и $n_{.j} = \sum_{i=1}^{k} n_{ij}$,

коэффициент связи Чупрова:

$$T = \sqrt{\frac{\chi^2}{n\sqrt{(k^*-1)(l^*-1)}}},$$

где k^* (соответственно l^*) – число непустых строк (столбцов) в таблице частот,

обе корреляционные отношения:

$$\eta(Y/X) = \left[\sum_{i=1}^{k} \frac{n_{i.}(\bar{y}_i - \bar{y})^2}{S_Y^2\,(n-1)}\right]^{1/2},$$

$$\eta(X/Y) = \left[\sum_{j=1}^{l} \frac{n_{.j}(\bar{x}_j - \bar{x})^2}{S_X^2\,(n-1)}\right]^{1/2},$$

где $\bar{y}_i$ $(i=1,\ldots,k)$ и $\bar{x}_j$ $(j=1,\ldots,l)$ – эмпирические условные средние признаков Y и X соответственно. (Формулы для их вы-

числения задаются ниже.)

При каждом параметре связи устанавливается и его статистическая значимость.

Отметим, что все перечисленные величины вычисляются в данном комплекте программ непосредственно по измеренным значениям признаков. В комплекте программ для ЭВМ "Минск-32" при вычислении этих величин все наблюдения признака, находящиеся в i-ом классе, заменяются т.н. представителем этого класса, которым в общем случае является взвешенное среднее тех элементов выборки, которые попадают в данный класс.

2.5. Информация о соединении классов. Для вычисления корреляционных отношений и коэффициента корреляции Чупрова в таблице частот последовательно соединяются те классы, в которых мало индивидов (меньше заданного числа ω), со соседними классами, пока не достигается достаточная величина класса. На листинге показывается, какие классы соединялись.

2.6. Условные распределения признаков. Условные распределения признака X определены для каждого класса B_j (j=1,...,l) признака Y. Эти условные распределения расположены в таблице условных распределений по столбцам, причем номер столбца указывает номер класса B_j, который определяет условие. Таким образом, в j-ом столбце (j=1,...,l) располагаются величины $100\,\frac{n_{ij}}{n_{\cdot j}}$ (i=1,...,k). Очевидно, сумма величин в каждом столбце равняется единице (100%).

Аналогично вычисляются условные распределения признака Y при условиях $X \in A_i$ (i=1,...,k). Они располагаются в соответствующей таблице по строкам.

2.7. Параметры условных распределений. Параметрами условных распределений признака X являются эмпирические условные средние $\bar{x}_j = \mathrm{E}X(Y \in B_j)$ (т.е. эмпирические средние признака X при условии, что значения признака Y относятся к классу B_j):

$$\bar{x}_j = \frac{1}{n_{\cdot j}} \sum_{y_i \in B_j} x_i \quad (j=1,\dots,l),$$

эмпирические условные стандартные отклонения:

$$s_X(j) = \sqrt{\frac{1}{n_{\cdot j}-1} \sum_{y_i \in B_j} (x_i - \bar{x}_j)^2} \quad (j=1,\dots,l).$$

Вычисляются также доверительные границы условных средних признака X:

$$\bar{x}_j - \frac{s_X(j)t_\alpha}{\sqrt{n_{\cdot j}}} \quad \text{и} \quad \bar{x}_j + \frac{s_X(j)t_\alpha}{\sqrt{n_{\cdot j}}} \quad (j=1,\dots,l).$$

Все эти параметры выпечатываются вместе с таблицей эмпирического распределения (см. 2.3) в виде дополнительных четырех строк под таблицей.

Аналогично определяются параметры условных распределений признака Y. Они выпечатываются в форме дополнительных столбцов к той же таблице распределения.

2.8. Регрессионные линии. Выпечатываются графики регрессионных линий (т.е. графики условных средних) признаков X и Y относительно признаков Y и X соответственно (см. [2]).

2.9. Корреляционное поле. Выпечатывается корреляционное поле признаков X и Y (см. [2]).

2.10. Номера индивидов. Выпечатываются номера тех индивидов, соответствующие значения которых попали в таблицу частот. Какой-то i-ый индивид может не попасть в таблицу, во-первых, потому что значение x_i не включается ни в одном из классов A_i или y_i ни в одном B_j; во-вторых, потому что данные неполны и значение X или Y не измерено при i-ом индивиде.

Все перечисленные части листинга при помощи ключа режима выпечатки можно либо выпечатать, либо опускать независимо друг от друга (за исключением параметров условных распределений, которые печатаются всегда вместе с таблицей эмпирического распределения). Это дает заказчику возможность заказывать только интересующую его часть информации.

3. Заказ двухмерного статистического анализа

Исходом двухмерного статистического анализа является массив значений признаков $X_1,\dots,X_m$ в стандартном виде на магнитном диске (см. [3]). Двухмерный анализ происходит на основе заказа, в котором нужно указывать:

1) перечень исследуемых признаков,

2) данные классифицирования,

3) общие параметры для анализа,

4) режим выпечатки.

Заказ вводится в ЭВМ через перфокарточное устройство ввода.

Перечень исследуемых признаков задается в виде

$$i_1,\dots,i_r * k_1,\dots,k_p) \ , \qquad (1)$$

где i_s (s=1,...,r) – номера признаков из группы Z_1, а k_t (t=1,...,p) – номера признаков из группы Z_2. Если $Z_1 = Z_2$, то перечень задается в виде

$$i_1,\ldots,i_r) \ . \qquad (2)$$

Признаки могут быть любого типа: A,B,C или R (см. [3]).

Данные классифицирования задаются только для тех признаков, для которых надо изменить имеющиеся в стандартном массиве признаков данные классифицирования.

Заказчик должен задавать некоторые общие параметры работы:

уровень значимости α,

имя массива значений признаков на магнитном диске,

количество индивидов N в этом массиве,

критерий пустого класса ω (см. 2.5),

критерий пустой таблицы ε.

Таблица частот считается пустой, если $n < \varepsilon$. В таком случае двухмерный анализ рассматриваемой пары признаков не осуществляется.

Все результаты двухмерного статистического анализа данных выводятся только на листе печатающего устройства. Листинг двухмерного анализа признаков $X_1,\ldots,X_m$ состоит из последовательности листингов двухмерных анализов отдельных пар признаков. Если перечень признаков задан в виде (1), то эта последовательность состоит из анализов всех пар признаков

$$(X_{i_s},X_{k_t}) \qquad (X_{i_s} \in Z_1;\ s=1,\ldots,r;\ X_{k_t} \in Z_2;\ t=1,\ldots,p).$$

Если перечень задан в виде (2), то эта последовательность состоит из анализов пар

$$(X_{i_s}, X_{i_t}) \qquad (X_{i_s}, X_{i_t} \in Z_1;\ i_s \geq i_t) .$$

Притом объем выпечатываемой информации при каждой паре определяется режимом выпечатки.

Литература

1. Труды ВЦ ТГУ, 1977, 40.
2. Вилисмяэ Ю., Выпечатка графиков первичного статистического анализа данных. Труды ВЦ ТГУ, 1978, 42, 117–124.
3. Коскель С., Структура данных на этапе ввода. Труды ВЦ ТГУ, 1978, 42, 105–111.
4. Programme kõigile XI, Tartu, 1976.

ЭМПИРИЧЕСКОЕ ИССЛЕДОВАНИЕ ГОРОДСКОЙ СРЕДЫ ПРИ ПОМОЩИ СИСТЕМЫ СТАТИСТИЧЕСКОЙ ОБРАБОТКИ ДАННЫХ В ВЦ ТГУ

Л.М. Тоодинг

В 1978–1979 году в вычислительном центре ТГУ проводилась совместная с горисполкомом г. Тарту научно-исследовательская работа по эмпирическому анализу городской среды. Целью исследования было составление математико-статистических моделей восприятия некоторых аспектов городской среды, а также анализ разрабатываемых моделей.

Обработка соответствующего эмпирического материала производилась на ЭВМ "Минск-32" и ЕС-1022, используя систему статистической обработки данных в ВЦ ТГУ (см. [6]) и пакет SSP (дополненная нами версия, см. [5]), притом практически в полном объеме. Поэтому, настоящему обзору одной конкретной прикладной работы, проведенной в ВЦ ТГУ, можно присвоить также определенное методологическое значение в смысле иллюстрации возможностей имеющегося у нас программного обеспечения.

1. Структура данных и выбор основных рабочих гипотез

1.1. Эмпирические данные рассматриваемого исследования собирались в виде анкет, заполнявшихся респондентами при по-

мощи интервьюеров. Респонденты выбирались случайно, притом попадание в выборку являлось равновероятным для всех тартуанцев в возрасте от 18 до 75 лет. Из 1100 разданных анкет пригодными для анализа оказались 1068.

За заполнением анкет следовала их проверка и кодирование, а затем ввод в фонд данных системы статистического анализа. Из 160 вопросов анкеты путем формализации образовался массив данных с 640 признаками.

Напомним, что основной структурной единицей данных при системе является прямоугольная матрица $X = (x_{ij}, i=1,\ldots,M; j=1,\ldots,N)$, представляющая выборку объема N M-мерного признак-вектора $\vec{X}_M = (X_1,\ldots,X_M)$, где элемент x_{ij} – значение i-го признака у j-го индивида (объекта). Данные могут быть неполными и иметь произвольный статистико-вероятностный характер (см. [7]).

В рассматриваемой задаче средняя доля пробелов образует 11% от объема выборки, а распределение признаков по типам следующее:

качественные – 19%

количественные упорядоченные – 66%

количественные – 15%.

Таким образом, в основном обработку можно было провести методами анализа количественной информации (осмысляя притом результаты упорядоченных признаков качественно), значит, данные соответствуют тому программному обеспечению, которым мы пользовались.

1.2. Содержательные гипотезы данной работы затрагивают в основном пространственно-деятельные соотношения в городской среде. В этих целях были выделены определенные пространственные зоны и типы действия. Рассматривалось следующее разделение городской среды:

1) квартира, жилое помещение;

2) жилище, жилой дом;

3) окружность дома;

4) соседство;

5) часть города, микрорайон;

6) город в целом;

7) загород.

В причинном плане, для установления соотношений детерминации, данное разделение осмысляется как некоторая иерархия и используется формальной базой для выведения основных рабочих гипотез. Притом есть смысл рассматривать оба направления - сверху вниз и снизу вверх, - считая все предыдущие ступени аргументами для данного.

Так, например, стереотип микрорайона мыслимо объяснить как характером связей с загородом и с городом в целом, так и непосредственно квартирой, отношениями к первичной жилой среде.

1.3. Большинство признаков рассматриваемого массива данных и служит для описания пространственных зон, но наряду с этим рассматриваются также следующего рода действия:

1) движение по маршруту "дом - место работы";

2) действия в домашнем хозяйстве;

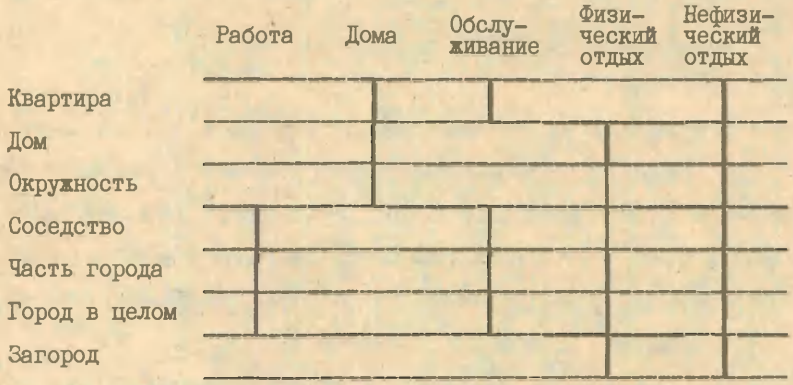
3) движение в целях обслуживания, закупок;

4) занятия физического отдыха;

5) занятия отдыха нефизического характера.

Основные рабочие гипотезы при данном разделении выражаются в зависимости 2–5 от 1 и во взаимосвязях 4 и 5, остальные возможные соотношения в признаках не отражаются.

1.4. Действия естественным образом прикрепляются к определенным зонам. В данной работе рассматривалось следующее их совместное распределение.



Приведенной схемой выражаются и выдвинутые и проверенные в данном исследовании рабочие гипотезы о связи (как правило, о взаимосвязи) между деятельным стереотипом и пространственными воображениями.

Фоном всех действий в соответствующих зонах являются социально-демографическая характеристика респондента и семьи, стереотип движения по городу и снабженность городской информацией.

1.5. Отметим еще один способ содержательного классифицирования данного материала. Именно, все названные выше блоки по существу распадаются на две части: на описание нынешнего положения ("существующее") и на описание идеального положения ("желаемое"). Различием в смысле этих двух аспектов характеризуется по соответствующим признакам степень удовлетворенности при каждой пространственной зоне и при каждом действии.

2. Методика статистического анализа

2.1. Эмпирический анализ городской среды методами математической статистики характеризуется довольно скромным практическим опытом, к тому же прибавляются и существенные локальные особенности. Проведенные до сих пор в г. Тарту эмпирические исследования (напр. [2,3]) хотя и пошли впрок своей ценной информацией об отдельных сторонах городского организма, но не затрагивали городской системы в целом, как это ожидается при рассматриваемой работе.

В методологическом плане такое отсутствие априорных знаний означает, с одной стороны, сравнительно формальную, по некоторому принципу систематическую постановку рабочих гипотез (как это и указывалось в 1), а с другой – намерение воспользоваться абсолютно всей мощностью имеющихся в нашем распоряжении средств обработки данных и соблюдения принципа максимального сохранения исходной информации на всех этапах анализа.

2.2. На этапе ввода в целях наименьшей потери информации проводилась сравнительно слабая формализация данных. Производилось лишь закодирование в численный вид, окончательная же формализация была отложена на этап обработки. Как правило, в течение анализа для одной и той же группы признаков использовалось несколько разных схем формализации.

Уточняем, что под формализацией в данном случае имеются в виду разнообразные преобразования признаков: перекодирование, вычисления по правилу типа арифметического выражения, индексирование и т.п. операции в системе статистической обработки данных (см. [4], стр. 109).

Например, на основе информации, представленной респондентом о структуре желаемой им квартиры (величина и функция каждого помещения), выводится функциональный тип квартиры. Притом количество названных помещений может быть произвольным.

В рассматриваемом массиве данных преобразованиям в целях формализации подвергалось 45% исходных признаков, притом в непреобразованном виде с ними был проведен в основном лишь первичный анализ.

2.3. В силу слабой формализованности и сложности изучаемого (объекта), а также с преднамерением затронуть "кое-что обо всем", в данном случае рассматривается пространство признаков сравнительно высокой размерности. Поэтому, общую схему анализа можно характеризовать как обработку исходных признаков отдельными блоками, притом связь между блоками рассматривается по сводным признакам блока. Точнее это значит

следующее.

Пусть задано некоторое разбиение множества элементов признак-вектора $\vec{X}$ на P непересекающихся множеств, которым соответствуют признак-векторы $\vec{X}^1_{M_1},\dots,\vec{X}^P_{M_P}$, где $\sum_{i=1}^{P} M_i = M$. Этим разбиением заодно определяется и P подматриц от выборки X, являющихся исходом для операций системы.

Разбиение может определяться, например, содержательной структурой признаков. В данном случае и рассматриваются совокупности признаков, соответствующие перечисленным в 1 зонам и действиям, к чему прибавляются еще блоки социально-демографических данных респондента и его семьи, блок о городской информации и способах движения.

В пределах каждого блока проводилась обработка с использованием, как правило, всех операций системы (см. [1]). Составлялись эмпирические распределения (одного признака, пары признаков, признак-вектора произвольной размерности), проводилось множественное сравнивание средних методом Шеффе, осуществлялся анализ корреляционных матриц, применялись множественные дисперсионный и регрессионный анализы.

В целях сжатия информации производились разные преобразования признаков и применялся факторный анализ. В результате этого i-ый блок $\vec{X}^i_{M_i}$ для дальнейшего анализа представлялся сводным признак-вектором $\vec{Y}^i_{m_i}$, где $m_i \ll M_i$. Анализом признак вектора $\vec{Y} = (\vec{Y}^1,\dots,\vec{Y}^P)$ осуществлялась проверка гипотез о взаимосвязях разных содержательных блоков, притом использовалась та же методика, что и внутри блоков.

Например, составлялись модели зависимости факторов квартиры от факторов способа проведения досуга и факторов соот-

ношений с загородом. Факторы желаемого типа района города прогнозировались через факторы семьи респондента и факторами жилища и т.д.

Таким подходом удалось в множественный анализ признаков включить косвенно и такие объемистые совокупности исходных признаков, совместный анализ которых был бы технически неосуществим. На уровне сводных признаков наблюдалась более высокая множественная связь в сравнении с любыми надлежащими моделями исходных признаков. Вдобавок облегчается в некоторых случаях (например, в регрессионных моделях) интерпретация результатов, так как получающиеся факторным анализом сводные признаки взаимоортогональны и отпадает надобность в осмыслении совлияний.

Переход на уровень сводных признаков в свою очередь можно осуществить и для некоторого разбиения $\tilde{Y}$ и т.д.

2.4. Наряду с естественной внутренней структурой совокупности признаков в данном исследовании рассматривались и некоторые способы априорной классификации индивидов. При этом имеется в виду не только анализ подматриц от X, соответствующих индивидам, при которых выполнено заданное логическое условие (см. [4], стр. 107), но и анализ данных об определенных сводных объектах. Хотя непосредственно измерялся респондент, житель города, и в отношении к нему и была устроена соответствующая статистическая выборка, согласно целям исследования, потребовалось также рассмотрение семьи и части города как самостоятельных объектов.

Данные для этих сводных объектов сгенерировались от вы-

борки X по информации соответствующих рассматриваемому суперобъекту респондентов.

Признаки семьи вычислялись операциями преобразования признаков: усреднением, индексированием, кодированием по логическому условию (см. [4]). Объекты "респондент" и "семья" находятся в однозначном соответствии, так как по правилу составления выборки весьма маловероятно попадание в нее нескольких представителей одной и той же семьи. Поэтому, получающиеся сводные данные о семье включались в первоначальный массив и обрабатывались вместе со всеми исходными признаками.

В качестве данных семьи вычислялись, например, демографический и социальный тип семьи, образовательный статус, средние показатели дохода, возрастный индекс и т.п.

Значения признаков для объекта "часть города" вычислялись операцией усреднения по значениям соответствующих признаков у респондентов рассматриваемой части города. Массив данных объекта "часть города" не совместим с массивами респондента и семьи, поэтому обрабатывался отдельно. В основном применялись методы классификации объектов. Так как передача информации в данном случае (в условиях нашего программного обеспечения) осуществляется повторным вводом, а не системно, то суперобъектов типа "часть города" следует рассматривать лишь в порядке исключения.

Вышесказанным охватываются основные характерные черты данной задачи. Опыт применения системы статистической обработки данных для решения ее убедил в том, что методологичес-

кие основы системы выбраны удачно и позволяют путем сравнительно формального подхода даже при скромной априорной информации получить содержательно хорошо осмысляемые результаты.

Литература

1. Прээм М., Тоодинг Л.М., Операции и принципы их следования. Труды ВЦ ТГУ, 1977, 40, 115-135.
2. Социальные предпосылки формирования пространственно-предметной среды. Тарту, 1972.
3. Типология уклада жизни городских и сельских семей на основе социологических и пространственно-предметных характеристик. Тарту, 1974.
4. Тоодинг Л.М., Преобразование стандартной матрицы. Труды ВЦ ТГУ, 1977, 40, 106-114.
5. Тоодинг Л.М., Применение пакета SSP в системе STP. Труды ВЦ ТГУ, 1978, 42, 147-156.
6. Тоодинг Л.М., Система статистической обработки данных в вычислительном центре ТГУ. Труды ВЦ ТГУ, 1977, 40, 3-7.
7. Тоодинг Л.М., Структура данных. Труды ВЦ ТГУ, 1977, 40, 75-85.

О ПРОЕКЦИОННЫХ МОДЕЛЯХ МНОГОМЕРНОГО ИНВАРИАНТНОГО ШКАЛИРОВАНИЯ

А.Б. Хмельницкая

При решении различных вопросов, связанных с обработкой экспериментальных данных, описываемых большим числом наблюдаемых характеристик, возникает проблема "сжатия" имеющейся эмпирической информации, т.е. задача адекватного описания изучаемых объектов посредством существенно меньшего набора интегральных признаков. Для решения подобного рода задач могут быть применены методы инвариантного шкалирования [1-4].

Наличие различных моделей инвариантного шкалирования позволяет изучать эмпирические данные различными способами и тем самым получать о них более объективную сжатую информацию.

В настоящей работе исследуются проекционные модели многомерного инвариантного шкалирования. В [4] показано, что все проекционные многомерные модели одной размерности при фиксированной функции близости в исходном пространстве признаков в случае, когда в качестве функции близости на выстраиваемой многомерной шкале рассматривается квадрат евклидовой метрики, эквивалентны и в следствии теоремы существования для соответствующей метрической модели [2] имеют решение (существуют функции шкалирования). В силу эквивалентности раз-

личных моделей достаточно рассмотреть только одну многомерную проекционную модель. Оказывается, что основные факты, справедливые для проекционных одномерных моделей [3-4] и для многомерной метрической модели [2], остаются в силе. А именно:

1) искать решение поставленной задачи можно двумя различными способами, что в связи с применением в вычислительных процедурах локальных методов дает возможность использовать решение, найденное одним путем, в качестве начального приближения для получения уточненного решения с помощью другой процедуры;

2) в случае вырожденности функции близости в исходном пространстве признаков вектор-функция шкалирования допускает два различных конечнопараметрических представления, причем число параметров обоих представлений не зависит от объема выборки, а лишь от "сложности" задачи;

3) для отыскания параметров, определяющих вектор-функцию шкалирования, для обоих представлений получены системы нелинейных уравнений;

4) в ряде случаев, в частности, в случае статистической независимости строящихся шкал, даются аналитические представления вектор-функции проекционного шкалирования.

Пусть (X,μ,r) - исходная генеральная совокупность объектов. Полагаем $r \in L^{\infty}(X \times X, \mu \times \mu)$. ρ - функция близости в R^n:

$$\rho(x,y) = \sum_{i=1}^{n} (x_i - y_i)^2 \quad \text{для любых} \quad x,y \in R^n. \qquad (1)$$

Для любого отображения $f : X \to R^n$ определим

$$\rho_f(z,w) = \rho(f(z),f(w)) \quad \text{для любых} \quad z,w \in X. \qquad (2)$$

Через $(\cdot,\cdot)$ и $\|\cdot\|$ обозначим соответственно скалярное произведение и норму в $L^2(X\times X,\mu\times\mu)$.

Так как ρ_f, определяемая формулами (1) – (2), положительно однородна степени 2 ([4]), то при исследовании проекционных моделей n-мерного инвариантного шкалирования достаточно рассмотреть задачу построения отображения $t : X \to R^n$ класса $L^4(X,\mu;R^n)$, доставляющего глобальный максимум функционалу

$$\psi_n^{(n)}(f) = \frac{(r,\rho_f)}{\|\rho_f\|} = \frac{\int\int r(z,w)\sum_{i=1}^{n}[f_i(z)-f_i(w)]^2 d\mu(z)d\mu(w)}{\sqrt{\int\int\left\{\sum_{i=1}^{n}[f_i(z)-f_i(w)]^2\right\}^2 d\mu(z)d\mu(w)}} \qquad (3)$$

на множестве отображений $f : X \to R^n$ класса $L^4(X,\mu;R^n)$, удовлетворяющих условию

$$\|\rho_f\|^2 = \int\int\left\{\sum_{i=1}^{n}[f_i(z)-f_i(w)]^2\right\}^2 d\mu(z)d\mu(w) = \|r\|^2. \qquad (4)$$

Заметим, что функционал $\psi_n^{(n)}(f)$ и условие (4) инвариантны относительно сдвига f на произвольный постоянный вектор. Поэтому исследуемую задачу проекционного многомерного инвариантного шкалирования достаточно рассматривать на подпространстве $L^{(n)} \subset L^4(X,\mu;R^n)$

$$L^{(n)} = \{f \in L^4(X,\mu;R^n) \mid \bar{f} = 0\},$$

где $\bar{f} = (\bar{f}_1,\ldots,\bar{f}_n)$; $\bar{f}_i = \int_X f_i d\mu \quad 1 \le i \le n.$

Кроме того, аналогично метрической модели многомерного инвариантного шкалирования, функционал $\psi_n^{(n)}(f)$ и условие (4) ин-

вариантны относительно вращения вектора f в R^n.

Как показано в [4], поставленную задачу можно решать двумя способами.

1. Т.к. для любого $a \neq 0$ $\Psi_{\Pi}^{(n)}(af) = \Psi_{\Pi}^{(n)}(f)$, то ищем любую вектор-функцию $t(z)$, на которой достигается глобальный максимум функционала $\Psi_{\Pi}^{(n)}(f)$ на множестве $f \in L^{(n)}$, а затем строим $t'(z) = at(z)$, причем величину a выбираем так, чтобы вектор-функция $t'(z)$ удовлетворяла условию (4).

Эту задачу будем кратко называть задачей $\Pi^{(n)}1$.

2. Решаем задачу на условный максимум функционала

$$\Psi_{\Pi Y}^{(n)}(f) = (r, \rho_f) = \int_X \int_X r(z,w) \sum_{i=1}^{n} [f_i(z) - f_i(w)]^2 d\mu(z) d\mu(w)$$

при условии (4).

Эту задачу будем называть задачей $\Pi^{(n)}2$.

Хотя так же как и в одномерном случае решения задач $\Pi^{(n)}1$ и $\Pi^{(n)}2$ совпадают, но т.к. они ищутся локальными методами и вычислительные схемы различны, параллельное решение обеих задач представляет интерес как в плане использования решения, полученного в рамках одной модели для получения уточненного решения в рамках другой, так и в плане сравнения результатов.

Для любой вектор-функции $f \in L^4(X, \mu; R^n)$ обозначим

$$\sigma_i^{(k)}(f) = \int_X f_i^k(w) d\mu(w) \qquad 1 \leq i \leq n; \; k = 1,2,3,4 \qquad (5)$$

$$\gamma_{ij}^{(k,l)}(f) = \int_X f_i^k(w) f_j^l(w) d\mu(w) \qquad 1 \leq i,j \leq n, \; i \neq j; \; k = 1,2,3; \; l = 1,2 \qquad (6)$$

$$\xi_{ijl}^{(k)}(f) = \int_X f_i^k(w) f_j(w) f_l(w) d\mu(w) \qquad 1 \leq i,j,l \leq n, \; i \neq j, \; i \neq l, \; j \neq l; \; k = 1,2 \qquad (7)$$

$$\varphi_{ijls}(f) = \int_X f_i(w)f_j(w)f_l(w)f_s(w)d\mu(w) \qquad 1\le i,j,l,s\le n,\ i\ne j,\ i\ne l,\ i\ne s,\ j\ne l,\ j\ne s,\ l\ne s \tag{8}$$

$$q_j(f;z) = \int_X r(z,w)f_j(w)d\mu(w) \qquad 1\le j\le n \tag{9}$$

Заметим, что $\psi_{ij}^{(k,l)}(f) = \psi_{ji}^{(l,k)}(f)$, $\xi_{ijl}^{(1)}(f) = \xi_{i_1i_2i_3}^{(1)}(f)$, $\varphi_{ijls}(f) = \varphi_{j_1j_2j_3j_4}(f)$, где (i_1,i_2,i_3), (j_1,j_2,j_3,j_4) соответственно произвольные перестановки чисел (i,j,l) и (i,j,l,s).

Положим

$$r(z) = \int_X r(z,w)d\mu(w).$$

В дальнейшем, как обычно, будем предполагать симметричность функции близости r. Это не умаляет общности результатов, однако упрощает их формулировку.

ТЕОРЕМА 1. 1. Если вектор-функция $t \in L^{(n)}$ является решением проекционной модели инвариантного шкалирования в смысле задачи $\Pi^{(n)}1$, то для п.в. $z \in (X,\mu)$ имеет место система равенств

$$\begin{cases} K_j^{(n)}(t;z) = 0 \qquad 1 \le j \le n \\ \sum_{i=1}^{n}\left\{\sigma_i^{(4)}(t)+3[\sigma_i^{(2)}(t)]^2\right\} + 2\sum_{j>i=1}^{n}[\sigma_i^{(2)}(t)\sigma_j^{(2)}(t)+ \\ \qquad +2[\psi_{ij}^{(1,1)}(t)]^2+\psi_{ij}^{(2,2)}(t)] = \frac{\|r\|^2}{2}, \end{cases} \tag{10}$$

где

$$K_j^{(n)}(t;z) = t_j^3(z)(r,\rho_t) + t_j(z)\left\{[3\sigma_j^{(2)}(t) + \right.$$
$$+ \sum_{\substack{i=1\\ i\ne j}}^{n}(t_i^2(z) + \sigma_i^{(2)}(t))](r,\rho_t) - \|r\|^2\bar{r}(z)\Big\} +$$
$$+ \left\{[-\sigma_j^{(3)}(t) + \sum_{\substack{i=1\\ i\ne j}}^{n}(2t_i(z)\psi_{ij}^{(1,1)}(t) - \right.$$
$$\left.- \psi_{ij}^{(2,1)}(t))](r,\rho_t) + \|r\|^2 q_j(t;z)\right\}.$$

2. Если вектор-функция $t \in L^{(n)}$ является решением проекционной модели инвариантного шкалирования в смысле задачи $\Pi^{(n)}2$, то для п.в. $z \in (X,\mu)$ имеет место система равенств

$$\begin{cases} \bar{K}_j^{(n)}(t;\lambda;z) = 0 \qquad 1 \le j \le n \\ \sum_{i=1}^{n} \left\{\sigma_i^{(4)}(t)+3\left[\sigma_i^{(2)}(t)\right]^2\right\} + 2 \sum_{j>i=1}^{n} \left[\sigma_i^{(2)}(t)\sigma_j^{(2)}(t)+ \right. \\ \left. +2\left[\tau_{ij}^{(1,1)}(t)\right]^2+\tau_{ij}^{(2,2)}(t)\right] = \frac{\|r\|^2}{2} , \end{cases} \tag{11}$$

где

$$\begin{aligned} \bar{K}_j^{(n)}(t;\lambda;z) = {} & 2\lambda t_j^3(z) + t_j(z)\left\{2\lambda\left[3\sigma_j^{(2)}(t) + \right.\right. \\ & \left. + \sum_{\substack{i=1 \\ i\neq j}}^{n} \left(t_i^2(z) + \sigma_i^{(2)}(t)\right)\right] - \bar{r}(z)\Big\} + \\ & + \left\{2\lambda\left[-\sigma_j^{(3)}(t) + \sum_{\substack{i=1 \\ i\neq j}}^{n} \left(2t_i(z)\tau_{ij}^{(1,1)}(t) - \right.\right.\right. \\ & \left.\left.\left. - \tau_{ij}^{(2,1)}(t)\right)\right] + q_j(t;z)\right\} , \end{aligned}$$

причем параметр λ является множителем Лагранжа и при выполнении системы равенств (11) имеет место равенство

$$\lambda = \frac{\psi_{\Pi X}^{(n)}(t)}{2\|r\|^2} . \tag{12}$$

Системы уравнений (10) и (11) будем называть системами уравнений многомерного проекционного инвариантного шкалирования соответственно в смысле модели $\Pi^{(n)}1$ или $\Pi^{(n)}2$.

Рассмотрим задачи $\Pi^{(n)}1$ и $\Pi^{(n)}2$, когда функция $r(z,w)$ вырождена, т.е. допускает конечное представление

$$r(z,w) = \sum_{\nu=1}^{q} P_\nu(z)Q_\nu(w), \tag{13}$$

где $P_\nu, Q_\nu \in L^\infty(X,\mu)$, $1 \le \nu \le q$. Представление (13) имеет место для ряда важных функций близости, встречающихся в приложениях при анализе как количественной так и качественной информации (например, для взвешенной евклидовой функции близости [1]).

Положим

$$\alpha_\nu = \int_X P_\nu(w)d\mu(w) \qquad 1 \le \nu \le q$$

$$\beta_\nu = \int_X Q_\nu(w)d\mu(w) \qquad 1 \le \nu \le q$$

Для любой вектор-функции $f \in L^4(X,\mu;R^n)$ определим два набора параметров

$$s_f^{(1)} = \Big\{\Big\{\sigma_i^{(k)}(f)\Big\}_{\substack{1\le i\le n\\ k=2,3,4}}, \Big\{\tau_{ij}^{(k,l)}(f)\Big\}_{\substack{1\le i,j\le n;\ i\ne j\\ k=1,2,\ l=1,2}}$$

$$\Big\{\theta_{i\nu}^{(k)}(f)\Big\}_{\substack{1\le i\le n,\ 1\le\nu\le q\\ k=1,2}}, \Big\{\eta_{i\nu}(f)\Big\}_{\substack{1\le i\le n\\ 1\le\nu\le q}}\Big\},$$

$$s_f^{(2)} = \Big\{\Big\{\sigma_i^{(k)}(f)\Big\}_{\substack{1\le i\le n\\ k=2,3,4}}, \Big\{\tau_{ij}^{(k,l)}(f)\Big\}_{\substack{1\le i,j\le n;\ i\ne j\\ k=1,2,\ l=1,2}}$$

$$\Big\{\theta_{i\nu}^{(1)}(f)\Big\}_{\substack{1\le i\le n\\ 1\le\nu\le q}}\Big\},$$

где $\sigma_i^{(k)}(f)$, $\tau_{ij}^{(k,l)}(f)$ определяются равенствами (5) – (6) и

$$\theta_{i\nu}^{(k)}(f) = \int_X f_i^k(w)Q_\nu(w)d\mu(w) \qquad 1 \le i \le n;\ 1 \le \nu \le q;\ k=1,2 \quad (14)$$

$$\eta_{i\nu}(f) = \int_X f_i(w)P_\nu(w)d\mu(w) \qquad 1 \le i \le n;\ 1 \le \nu \le q \quad (15)$$

Рассмотрим две совокупности $S^{(1)}$ и $S^{(2)}$ наборов соответственно

$$s^{(1)} = \Big\{\Big\{\sigma_i^{(k)}\Big\}_{\substack{1\le i\le n\\ k=2,3,4}}, \Big\{\gamma_{ij}^{(k,k)}\Big\}_{\substack{1\le i<j\le n\\ k=1,2}}, \Big\{\gamma_{ij}^{(2,1)}\Big\}_{\substack{1\le i,j\le n\\ i\ne j}},$$

$$\Big\{\theta_{iv}^{(k)}\Big\}_{\substack{1\le i\le n,\ 1\le v\le q\\ k=1,2}}, \Big\{\eta_{iv}\Big\}_{\substack{1\le i\le n\\ 1\le v\le q}}\Big\},$$

$$s^{(2)} = \Big\{\Big\{\sigma_i^{(k)}\Big\}_{\substack{1\le i\le n\\ k=2,3,4}}, \Big\{\gamma_{ij}^{(k,k)}\Big\}_{\substack{1\le i<j\le n\\ k=1,2}}, \Big\{\gamma_{ij}^{(2,1)}\Big\}_{\substack{1\le i,j\le n\\ i\ne j}},$$

$$\Big\{\theta_{iv}^{(1)}\Big\}_{\substack{1\le i\le n\\ 1\le v\le q}}\Big\}$$

вещественных чисел $\sigma_i^{(k)}$, $\gamma_{ij}^{(k,l)}$, $\theta_{iv}^{(k)}$, η_{iv}, удовлетворяющих условию

$$\sum_{i=1}^{n}[\sigma_i^{(4)}+3(\sigma_i^{(2)})^2]+2\sum_{j\le i=1}^{n}[\sigma_i^{(2)}\sigma_j^{(2)}+2(\gamma_{ij}^{(1,1)})^2+\gamma_{ij}^{(2,2)}] = \frac{\|r\|^2}{2}.$$

Отметим, что в дальнейшем под $\gamma_{ij}^{(1,2)}$ понимаем $\gamma_{ji}^{(2,1)}$ и, если $i > j$, то под $\gamma_{ij}^{(k,k)}$ понимаем всегда $\gamma_{ji}^{(k,k)}$.

Очевидно, что для любого $t \in L^{(n)}$, являющегося решением системы уравнений (10) (или (11)) $s_t^{(1)} \in S^{(1)}$ (соответственно $s_t^{(2)} \in S^{(2)}$).

Определим отображения

$M(s^{(1)}) : L^4(X,\mu;R^n) \to L^1(X,\mu;R^n)$ для любого $s^{(1)} \in S^{(1)}$,

$M_\lambda(s^{(2)}) : L^4(X,\mu;R^n) \to L^1(X,\mu;R^n)$ для любого $s^{(2)} \in S^{(2)}$

следующим образом. Для любых $f \in L^4(X,\mu;R^n)$ и $z \in X$ положим

$$M(s^{(1)})(f;z) = \Big\{[M(s^{(1)};f)_j](z)\Big\}_{j=1}^{n}$$

$$M_\lambda(s^{(2)})(f;z) = \Big\{[M_\lambda(s^{(2)};f)_j](z)\Big\}_{j=1}^{n}$$

$$[M(s^{(1)};f)_j](z)=f_j^3(z)+f_j(z)[\sum_{\substack{i=1\\ i\neq j}}^{n} f_i^2(z)-\delta_j^{(0)}(s^{(1)};z)]+$$

$$+2\sum_{\substack{i=1\\ i\neq j}}^{n} f_i(z)\gamma_{ij}^{(1,1)}+\delta_j(s^{(1)};z) \qquad (16)$$

$$[M_\lambda(s^{(2)};f)_j](z)=f_j^3(z)+f_j(z)[\sum_{\substack{i=1\\ i\neq j}}^{n} f_i^2(z)-\delta_{\lambda;j}^{(0)}(s^{(2)};z)]+$$

$$+2\sum_{\substack{i=1\\ i\neq j}}^{n} f_i(z)\gamma_{ij}^{(1,1)}+\delta_{\lambda;j}(s^{(2)};z)\ , \qquad (17)$$

где

$$\delta_j^{(0)}(s^{(1)};z)=\frac{\|r\|^2}{\varepsilon}\sum_{v=1}^{q}\beta_v P_v(z)-\sum_{i=1}^{n}\sigma_i^{(2)}-2\sigma_j^{(2)} \quad 1\le j\le n \qquad (18)$$

$$\delta_j(s^{(1)};z)=\frac{\|r\|^2}{\varepsilon}\sum_{v=1}^{q}\theta_{jv}^{(1)}P_v(z)-\sum_{\substack{i=1\\ i\neq j}}^{n}\gamma_{ij}^{(2,1)}-\sigma_j^{(3)} \quad 1\le j\le n \qquad (19)$$

$$\varepsilon = 2\sum_{v=1}^{q}\sum_{i=1}^{n}[\alpha_v\theta_{iv}^{(2)} - \theta_{iv}^{(1)}\eta_{iv}] \qquad (20)$$

$$\delta_{\lambda;j}^{(0)}(s^{(2)};z)= \frac{1}{2\lambda}\sum_{v=1}^{q}\beta_v P_v(z)-\sum_{i=1}^{n}\sigma_i^{(2)}-2\sigma_j^{(2)} \quad 1\le j\le n \qquad (21)$$

$$\delta_{\lambda;j}(s^{(2)};z)= \frac{1}{2\lambda}\sum_{v=1}^{q}\theta_{jv}^{(1)}P_v(z)-\sum_{\substack{i=1\\ i\neq j}}^{n}\gamma_{ij}^{(2,1)}-\sigma_j^{(3)} \quad 1\le j\le n \qquad (22)$$

Аналогично [2] для задач проекционного инвариантного шкалирования в смысле моделей $\Pi^{(n)}1$ и $\Pi^{(n)}2$ вводится понятие области разрешимости, соответственно $S_*^{(1)}$ и при фиксированном значении параметра λ $S_*^{(2)}(\lambda)$. $S_*^{(1)}$ ($S_*^{(2)}(\lambda)$) - это такое подмножество $S^{(1)}$ ($S^{(2)}$), что для любого набора $s^{(1)}\in S_*^{(1)}$

$(s^{(2)} \in S_*^{(2)}(\lambda))$ существует единственная вектор-функция $t \in L^4(X,\mu;R^n)$, удовлетворяющая для п.в. $z \in (X,\mu)$ системе уравнений

$$[M(s^{(1)};t)_j](z) = 0 \qquad 1 \le j \le n \qquad (23)$$

(соответственно

$$[M_\lambda(s^{(2)};t)_j](z) = 0 \qquad 1 \le j \le n\ ). \qquad (24)$$

В случае модели $\Pi^{(n)}_1$ в области разрешимости $S_*^{(1)}$ рассмотрим систему $n(2n+2+3q)$ уравнений относительно набора параметров $s^{(1)}$

$$\int_X [t_i(s^{(1)})](w)d\mu(w) = 0 \qquad 1 \le i \le n \qquad (25)$$

$$\int_X [t_i(s^{(1)})]^k(w)d\mu(w) = \sigma_i^{(k)} \qquad 1 \le i \le n;\ k=2,3,4 \qquad (26)$$

$$\int_X [t_i(s^{(1)})]^k(w)[t_j(s^{(1)})]^k(w)d\mu(w) = \gamma_{ij}^{(k,k)} \qquad 1 \le i < j \le n;\ k=1,2 \qquad (27)$$

$$\int_X [t_i(s^{(1)})]^2(w)[t_j(s^{(1)})](w)d\mu(w) = \gamma_{ij}^{(2,1)} \qquad 1 \le i,j \le n,\ i \ne j \qquad (28)$$

$$\int_X [t_i(s^{(1)})]^k(w)Q_v(w)d\mu(w) = \theta_{iv}^{(k)} \qquad 1 \le i \le n;\ 1 \le v \le q;\ k=1,2 \qquad (29)$$

$$\int_X [t_i(s^{(1)})](w)P_v(w)d\mu(w) = \eta_{iv} \qquad 1 \le i \le n;\ 1 \le v \le q\ , \qquad (30)$$

где вектор-функция $[t(s^{(1)})](z)$ является решением системы n кубических уравнений (23), которая в области разрешимости имеет единственное вещественное решение относительно вектор-функции $t(z)$.

Так как суть задачи $\Pi^{(n)}_2$ состоит в максимизации функционала $\Psi^{(n)}_{\Pi Y}$ то в силу равенства (12) ищем наибольшее λ, при котором в области разрешимости $S^{(2)}_*(\lambda)$ система $n(2n+2+q)$ уравнений (25)-(28) и (29) при $k=1$ (везде в (25)-(29) делаем замену $s^{(1)}$ на $s^{(2)}$) относительно набора параметров $s^{(2)}$, где вектор-функция $[t(s^{(2)})](z)$ является единственным решением системы кубических уравнений (24) с параметром λ, имеет решение.

ТЕОРЕМА 2. 1. Если существует такая вектор-функция проекционного n-мерного инвариантного шкалирования в смысле модели $\Pi^{(n)}_1$ ($\Pi^{(n)}_2$) $t \in L^{(n)}$, что имеет место включение $s^{(1)}_t \in S^{(1)}_*$ ($s^{(2)}_t \in S^{(2)}_*(\lambda)$ при некотором λ), то система уравнений (25)-(30) ((25)-(28) и (29) при $k=1$ и этом λ) всегда имеет решение $s^{(1)}_* \in S^{(1)}_*$ ($s^{(2)}_* \in S^{(2)}_*(\lambda)$).

2. Для любого решения $s^{(1)}_* \in S^{(1)}_*$ (при фиксированном λ $s^{(2)}_* \in S^{(2)}_*(\lambda)$) системы (25)-(30) ((25)-(28) и (29) при $k=1$) вектор-функция $t(s^{(1)}_*)$ ($t(s^{(2)}_*)$), определенная системой уравнений (23) ((24)), является решением системы уравнений (10) ((11) при этом λ).

ЗАМЕЧАНИЕ 1. Так же как и в случае метрической модели любое решение системы (25)-(30), (23) ((25)-(28), (29) при $k=1$, (24)) относительно вектор-функции $t(z)$ порождает $2^n n!$ различных решений рассматриваемой проекционной задачи на подпространстве $L^{(n)}$, которые соответствуют замене знаков и различным перестановкам координатных функций вектор-функции $t(z)$.

Аналогично метрической модели, если $t(z)$ является решением системы уравнений (25)-(30), (23) ((25)-(28), (29) при

k=1, (24)), $A = (a_{ij})_{i,j=1}^{n}$ – ортогональная матрица, то в силу инвариантности проекционной модели относительно вращений At(z) является решением соответствующей системы (25)–(30), (23) ((25)–(28), (29) при k=1, (24)), причем параметры новой системы $(\bar{\sigma}_i^{(k)}, \bar{\tau}_{ij}^{(k,l)}, \bar{\theta}_{iv}^{(k)}, \bar{\eta}_{iv})$ (соответственно $\bar{\sigma}_i^{(k)}, \bar{\tau}_{ij}^{(k,l)}, \bar{\theta}_{iv}^{(1)}$) пересчитываются по формулам

$$\bar{\sigma}_i^{(2)} = \sum_{j=1}^{n} a_{ij}\sigma_j^{(2)} + 2\sum_{k>j=1}^{n} a_{ij}a_{ik}\tau_{jk}^{(1,1)} \qquad 1\le i\le n \qquad (31)$$

$$\bar{\sigma}_i^{(3)} = \sum_{j=1}^{n} a_{ij}^3\sigma_j^{(3)} + 3\sum_{j=1}^{n}\sum_{\substack{k=1\\k\neq j}}^{n} a_{ij}^2a_{ik}\tau_{jk}^{(2,1)} +$$

$$+4\sum_{l>k>j=1}^{n} a_{ij}a_{ik}a_{il}\xi_{jkl}^{(1)}(t) \qquad 1\le i\le n \qquad (32)$$

$$\bar{\sigma}_i^{(4)} = \sum_{j=1}^{n} a_{ij}^4\sigma_j^{(4)} + 6\sum_{k>j=1}^{n} a_{ij}^2a_{ik}^2\tau_{jk}^{(2,2)} +$$

$$+4\sum_{j=1}^{n}\sum_{\substack{k=1\\k\neq j}}^{n} a_{ij}^3a_{ik}\tau_{jk}^{(3,1)}(t) + 12\sum_{j=1}^{n}\sum_{\substack{l>k=1\\k\neq j\\l\neq j}}^{n} a_{ij}^2a_{ik}a_{il}\xi_{jkl}^{(2)}(t) +$$

$$+24\sum_{p>l>k>j=1}^{n} a_{ij}a_{ik}a_{il}a_{ip}\varphi_{jklp}(t) \qquad 1\le i\le n \qquad (33)$$

$$\bar{\tau}_{ij}^{(1,1)} = \sum_{k=1}^{n} a_{ik}a_{jk}\sigma_k^{(2)} + \sum_{k=1}^{n}\sum_{\substack{l=1\\l\neq k}}^{n} a_{ik}a_{jl}\tau_{kl}^{(1,1)} \quad 1\le i<j\le n \qquad (34)$$

$$\bar{\tau}_{ij}^{(2,2)} = \sum_{k=1}^{n} a_{ik}^2a_{jk}^2\sigma_k^{(4)} + \sum_{k=1}^{n}\sum_{\substack{l=1\\l\neq k}}^{n}(a_{ik}^2a_{jl}^2 + 2a_{ik}a_{il}a_{jk}a_{jl})\tau_{kl}^{(2,2)} +$$

$$+2\sum_{i=1}^{n}\sum_{\substack{i=1\\l\neq k}}^{n}(a_{ik}^2a_{jk}a_{jl} + a_{ik}a_{il}a_{jk}^2)\tau_{kl}^{(3,1)}(t) +$$

$$+\sum_{k=1}^{n}\sum_{\substack{l=1\\l\neq k}}^{n}\sum_{\substack{p=1\\p\neq k\\p\neq l}}^{n}(a_{ik}^2a_{jl}a_{jp}+a_{il}a_{ip}a_{jk}^2+4a_{ik}a_{il}a_{jk}a_{jp})\xi_{klp}^{(2)}(t)+$$

$$+3\sum_{k=1}^{n}\sum_{\substack{l=1\\l\neq k}}^{n}\sum_{\substack{p=1\\p\neq k\\p\neq l}}^{n}\sum_{\substack{m=1\\m\neq k\\m\neq l\\m\neq p}}^{n}a_{ik}a_{il}a_{jp}a_{jm}\varphi_{klpm}(t) \quad 1\leq i<j\leq n \qquad (35)$$

$$\gamma_{ij}^{(2,1)}=\sum_{k=1}^{n}a_{ik}^2a_{jk}\sigma_k^{(3)}+\sum_{k=1}^{n}\sum_{\substack{l=1\\l\neq k}}^{n}(a_{ik}^2a_{jl}+2a_{ik}a_{il}a_{jl})\gamma_{lk}^{(2,1)}+$$

$$+2\sum_{k>l=1}^{n}\sum_{\substack{p=1\\p\neq k\\p\neq l}}^{n}a_{ik}a_{il}a_{jp}\xi_{klp}^{(1)}(t) \qquad 1\leq i,j\leq 1,\ i\neq j \qquad (36)$$

$$\bar{\theta}_{iv}^{(1)}=\sum_{j=1}^{n}a_{ij}\theta_{jv}^{(1)} \qquad 1\leq i\leq n;\ 1\leq v\leq q \qquad (37)$$

$$\bar{\theta}_{iv}^{(2)}=\sum_{j=1}^{n}a_{ij}^2\theta_{jv}^{(2)}+2\sum_{k>j=1}^{n}a_{ij}a_{ik}\psi_{jkv}(t) \qquad 1\leq i\leq n;\ 1\leq v\leq q \qquad (38)$$

$$\bar{\eta}_{iv}=\sum_{j=1}^{n}a_{ij}\eta_{jv} \qquad 1\leq i\leq n;\ 1\leq v\leq q \qquad (39)$$

где $\gamma_{ij}^{(3,1)}(t)$, $\xi_{ijk}^{(k)}(t)$, $\varphi_{ijkl}(t)$ определяются формулами (6) – (8), а

$$\psi_{ijv}(t)=\int_X t_i(w)t_j(w)Q_v(w)d\mu(w) \qquad 1\leq i<j\leq n;\ 1\leq v\leq q\,.$$

Ниже приводятся достаточные условия, обеспечивающие принадлежность наборов параметров, порождаемых вектор-функцией проекционного шкалирования в рамках моделей $\Pi^{(n)}{}_1$ и $\Pi^{(n)}{}_2$,

области разрешимости, и формулы, позволяющие вычислить значения вектор-функции шкалирования соответствующей этим наборам.

ТЕОРЕМА 3. Пусть $\tau(z) \in L^{(n)}$ является решением задачи $\Pi^{(n)}1$ ($\Pi^{(n)}2$). $A = (a_{ij})_{i,j=1}^{n}$ – какое-нибудь решение системы n^2 уравнений

$$\sum_{j=1}^{n} a_{ij}^2 = 1 \qquad 1 \le i \le n \qquad (40)$$

$$\sum_{k=1}^{n} a_{ik}a_{jk} = 0 \qquad 1 \le i < j \le n \qquad (41)$$

$$\sum_{k=1}^{n} a_{ik}\Big[a_{jk}\sigma_k^{(2)}(\tau) + \sum_{\substack{l=1 \\ l \ne k}}^{n} a_{jl}\mu_{kl}^{(1,1)}(\tau)\Big] = 0 \qquad 1 \le i < j \le n \qquad (42)$$

относительно вещественных a_{ij}.

(В случае модели $\Pi^{(n)}2$ $\lambda = \dfrac{\psi_{\Pi Y}^{(n)}(\tau)}{\|r\|^2}$.)

Тогда для того, чтобы набор параметров $s_t^{(1)}$ ($s_t^{(2)}$), порождаемый функцией проекционного n-мерного инвариантного шкалирования $t(z) = A\tau(z)$ принадлежал области разрешимости $S_*^{(1)}$ ($S_*^{(2)}(\lambda)$), достаточно, чтобы уравнение

$$T(z)\prod_{i=1}^{n}\left[T(z)-\Delta_j^{(0)}(z)\right]^2 - \sum_{i=1}^{n}\Delta_i^2(z)\prod_{\substack{j=1 \\ j\ne i}}^{n}\left[T(z)-\Delta_j^{(0)}(z)\right]^2 = 0, \qquad (43)$$

где $\Delta_j^{(0)}(z) = \delta_j^{(0)}(s_t^{(1)};z)$ $\quad(\Delta_j^{(0)}(z) = \delta_{\lambda;j}^{(0)}(s_t^{(2)};z))$,

$$\Delta_j(z) = \delta_j(s_t^{(1)};z) \qquad (\Delta_j(z) = \delta_{\lambda;j}(s_t^{(2)};z)) \quad , \qquad (44)$$

$\delta_j^{(0)}(s_t^{(1)};z)$, $\delta_j(s_t^{(1)};z)$, $\delta_{\lambda;j}^{(0)}(s_t^{(2)};z)$, $\delta_{\lambda;j}(s_t^{(2)};z)$ определяются равенствами (18) – (22) с помощью соотношений

(31)-(39) по вектор-функции $\tau(z)$ и набору параметров $s_\tau^{(1)}$ ($s_\tau^{(2)}$), для п.в. $z \in (X,\mu)$ имело единственное вещественное решение, удовлетворяющее условию

$$\sum_{i=1}^{n}[\sigma_i^{(4)}(f) + 3(\sigma_i^{(2)}(f))^2] + 2\sum_{j>i=1}^{n}[\sigma_i^{(2)}(f)\sigma_j^{(2)}(f) +$$

$$+ 2(\gamma_{ij}^{(1,1)}(f))^2 + \gamma_{ij}^{(2,2)}(f)] = \frac{\|r\|^2}{2},$$

где $f(z) = (f_1(z),\dots,f_n(z))$,

$$f_i(z) = -\frac{\Delta_i(z)}{T(z) - \Delta_i^{(0)}(z)} \qquad 1 \le i \le n,$$

а $T(z)$ - решение уравнения (43).

Более того, в этом случае для координатных функций вектор-функции $t(z) = (t_1(z),\dots,t_n(z))$ для п.в. $z \in (X,\mu)$ справедливы соотношения

$$t_i(z) = f_i(z) \qquad 1 \le i \le n.$$

ЗАМЕЧАНИЕ 2. Условия (40), (41) в совокупности суть условие ортогональности матрицы A, а смысл условия (42) состоит в том, что поворот вектор-функции проекционного инвариантного шкалирования $\tau(z)$, осуществляемый матрицей A гарантирует равенство нулю смешанных моментов вектор-функции $t(z) = A\tau(z)$, т.е.

$$\int_X t_i(w)t_j(w)d\mu(w) = 0 \qquad 1 \le i,j \le n,\ i \ne j.$$

В случае статистической независимости строящихся шкал, т.е. в случае равенства нулю смешанных моментов вектор-функции проекционного инвариантного шкалирования $t(z)$, и выпол-

нения дополнительного требования нормировки на шкалы

$$\int_X t_i^2(w)d\mu(w) = \text{const} \qquad 1 \le i \le n$$

справедлива

ТЕОРЕМА 4. Пусть решение задачи $\Pi^{(n)}1$ ($\Pi^{(n)}2$) таково, что

1) $t \in L^{(n)}$;

2) $\mathfrak{t}_{ij}^{(1,1)}(t) = 0, \qquad 1 \le i < j \le n;$

3) $\sigma_i^{(2)}(t) = \text{const} = \sigma_2(t), \qquad 1 \le i \le n.$

(В случае задачи $\Pi^{(n)}2$ $\lambda = \dfrac{\psi_{\Pi Y}^{(n)}(t)}{\|r\|^2}$.)

Тогда для того, чтобы $s_*^{(1)} \in S_*^{(1)}$ ($s_t^{(2)} \in S_*^{(2)}(\lambda)$), достаточно, чтобы для п.в. $z \in (X,\mu)$

$$\Delta(z) < 3\sqrt[3]{\frac{\sum_{i=1}^{n}\Delta_i(z)}{4}}$$

и $\sum_{i=1}^{n}\sigma_i^{(4)}(f) + n(2+n)[\sigma_1^{(2)}(f)]^2 + 2\sum_{j>i=1}^{n}\mathfrak{t}_{ij}^{(2,2)}(f) = \dfrac{\|r\|^2}{2}$,

где $\Delta_i(z)$ определяется соотношением (44),

$$\Delta(z) = \frac{\|r\|^2}{\varepsilon}\sum_{\nu=1}^{q}\beta_\nu P_\nu(z) - (n+2)\sigma_2(t),$$

$$\left(\ \Delta(z) = \frac{1}{2\lambda}\sum_{\nu=1}^{q}\beta_\nu P_\nu(z) - (n+2)\sigma_2(t)\ \right),$$

$$f_i(z) = \frac{3\Delta_i(z)}{5\Delta(z) + 3\left[\left(\frac{1}{2}v(z)+\sqrt{Q(z)}\right)^{1/3} + \left(\frac{1}{2}v(z)-\sqrt{Q(z)}\right)^{1/3}\right]}, \quad 1 \le i \le n,$$

$$Q(z) = \frac{2}{27} P^3(z) + \frac{1}{4} V^2(z),$$

$$P(z) = -\frac{1}{3} \Delta^2(z),$$

$$V(z) = \frac{2}{27} \Delta^3(z) - \sum_{i=1}^{n} \Delta_i^2(z),$$

причем в этом случае для координатных функций вектор-функции $t(z)$ для п.в. $z \in (X,\mu)$ справедливы соотношения

$$t_i(z) = f_i(z), \qquad 1 \le i \le n .$$

Отметим, что в силу справедливости теоремы 4 так же как и в случае метрической модели [2] разумно в общем случае задачу проекционного многомерного инвариантного шкалирования в рамках обеих моделей $\Pi^{(n)}_1$ и $\Pi^{(n)}_2$ рассматривать при дополнительном предположении выполнения условия статистической независимости строящихся шкал, т.е. для определения набора вектор-функции шкалирования, решать систему (25)-(30) ((25)-(28), (29) при $k=1$) при дополнительных условиях

$$r_{ij}^{(1,1)} = 0 \qquad 1 \le i < j \le n \qquad (45)$$

$$\sigma_i^{(2)} = \text{const} = \sigma_2 \qquad 1 \le i \le n . \qquad (46)$$

В этом случае условия 1), 2), 3) теоремы 4 оказываются выполнены. А тогда:

1) отпадает проблема решения системы (40)-(42), определяющей поворот начального приближения функции шкалирования, гарантирующий справедливость условия (45) для начального приближения решения системы (25)-(30) ((25)-(28), (29) при $k=1$);

2) упрощается проблема разрешимости, которая сводится к

условию существования единственного вещественного корня кубического уравнения, а не уравнения степени $2n+1$ в общем случае;

3) вектор-функция шкалирования в рамках обеих моделей $\Pi^{(n)}_1$ и $\Pi^{(n)}_2$ выписывается явно через набор параметров, ее определяющий.

Доказательство всех теорем проводится аналогично доказательствам соответствующих теорем в [1-3].

Литература

1. Перекрест В.Т., Об одной модели одномерного шкалирования. Автоматика и телемеханика, № 2, 1980.
2. Хмельницкая А.Б., Об оптимизационных моделях многомерного инвариантного шкалирования. Сб. "Моделирование и применение ЭВМ в социологии" под ред. Ю.Н. Толстовой. ИСИ АН СССР, 1980.
3. Перекрест В.Т., Хмельницкая А.Б., Проекционные модели одномерного инвариантного шкалирования. Сб. "Труды конференции по применению математических методов в социологии" под ред. Э.П. Андреева. М., "Наука", 1980.
4. Хмельницкая А.Б., Об эквивалентности некоторых моделей инвариантного шкалирования. Сб. "Дискретный анализ и обработка эмпирической информации № 2" под ред. Б.Г. Миркина. СО АН СССР, Новосибирск, 1980.

Содержание

МЕТОДЫ МНОГОМЕРНОГО СТАТИСТИЧЕСКОГО АНАЛИЗА. Труды вычислительного центра. Выпуск 44. На русском языке. Тартуский государственный университет. ЭССР, г. Тарту, ул. Юликооли, I8. Ответственный редактор С.Коскед. Сдано в печать II.I2.79. Бумага печатная 30x42 I/4. Печ. листов 7,0 (условных 6,5I).Учетно-нздат. листов 5,86. Тираж 500. МВ 09664. Типография ТГУ, ЭССР, г. Тарту, ул. Пялсона, I4. Зак. № I650. Цена 90 коп.