МИНИСТЕРСТВО ОБЩЕГО И
ПРОФЕССИОНАЛЬНОГО ОБРАЗОВАНИЯ
РОССИЙСКОЙ ФЕДЕРАЦИИ
АЛТАЙСКИЙ ГОСУДАРСТВЕННЫЙ УНИВЕРСИТЕТ
МАТЕМАТИЧЕСКИЙ ФАКУЛЬТЕТ, ЭКОНОМИЧЕСКИЙ
ФАКУЛЬТЕТ
КАФЕДРА МАТЕМАТИЧЕСКОГО АНАЛИЗА

МНОГОМЕРНЫЙ

СТАТИСТИЧЕСКИЙ АНАЛИЗ

Учебное пособие

**Издательство Алтайского
государственного университета
Барнаул, 2003**

УДК 517.0
ББК 22,16

Работа выполнена при поддержке Национального фонда подготовки кадров. Инновационный Проект Развития Образования.

Дронов С.В. Многомерный статистический анализ. : Учебное пособие. Барнаул: Изд-во Алт. гос. ун-та, 2003. 213 с.

Учебное пособие создано на основе опыта преподавания автором курсов многомерного статистического анализа и эконометрики. Содержит материалы по дискриминантному, факторному, регрессионному анализу, анализу соответствий и теории временных рядов. Изложены подходы к задачам многомерного шкалирования и некоторым другим задачам многомерной статистики.

Рецензент:

Глава 1

Предварительные сведения

Курс "Многомерный статистический анализ"является логическим продолжением и развитием таких традиционных математических курсов, как "Теория вероятностей"и "Математическая статистика". Он отчасти перекликается с курсом "Эконометрика", но содержит более универсальные и продвинутые методы. В процессе обработки числовых данных часто возникает необходимость проанализировать наличие связей и взаимную зависимость большого количества показателей. При этом отмечаются эффекты, которые обусловлены многомерным характером данных и отсутствовали в том случае, когда просто изучались связи двух одномерных величин. Одна из специфических задач, часто возникающая на практике - задача снижения размерности, суть которой в том, что нужно выяснить, какие из характеристик изучаемого объекта являются существенными для решаемой задачи, а какие несущественны и лишь неоправданно усложняют, засоряют рассматриваемую картину. Имеется множество других задач и методов их решения, которые можно применять в многомерной ситуации. Еще одним из разделов, традиционно относимых к курсам, аналогичным нашему, является набор приемов, применяемых при работе с нечисловыми данными, их обработка и придание им наглядного характера. Сегодня грамотный практик обязан учитывать множество факторов, влияющих на тот процесс, который он исследует с целью принятия оптимальных решений, а поэтому многомерный статистический анализ приобретает особую актуальность.

В последующих нескольких разделах мы вкратце напомним и разберем содержание тех разделов предшествующих математических курсов, которые особенно важны для того, чтобы успешно понимать все, напи-

санное в следующих главах.

1.1 Высшая математика: математический анализ и линейная алгебра

Нам потребуется умение дифференцировать и искать экстремумы функций, в том числе и функций многих переменных, что подразумевает знакомство с частными производными и дифференциалом. Неоднократно будет применяться аппарат поиска так называемых условных экстремумов, кратко именуемый методом Лагранжа. Задача ставится так: найти экстремумы функции $f(x_1, ..., x_n)$ многих переменных, если переменные $x_1, ..., x_n$ обязаны подчиняться набору ограничений

$$\varphi_j(x_1, ..., x_n) \;=\; 0, \quad j = 1, ..., m. \tag{1.1}$$

Решается эта задача путем введения так называемой функции Лагранжа

$$L(x_1, ..., x_n, \lambda_1, ..., \lambda_m) \;=\; f(x_1, ..., x_n) \;-\; \sum_{j=1}^{m} \lambda_j \varphi_j(x_1, ..., x_n),$$

зависящей от первоначальных переменных и набора искусственно вводимых переменных $\lambda_1, ..., \lambda_m$, соответствующих каждому из накладываемых ограничений (1.1). Далее функция L исследуется на экстремумы обычным методом, но к условию равенства нулю всех частных производных по $x_i,\ i = 1, ..., n$ при поиске критических точек добавляются условия (1.1), в результате чего получается система с $n+m$ неизвестными и таким же количеством уравнений. При исследовании второго дифференциала d^2L как функции $x_1, ..., x_n$ в критической точке $\vec{x}_0$ на знакоопределенность привлекаются дополнительные соотношения между $dx_1, ..., dx_n$

$$\sum_{i=1}^{n} \frac{\partial \varphi_j}{\partial x_i}(\vec{x}_0) \;=\; 0, \;\; j = 1, ..., m,$$

следующие из (1.1).

Конечно же, необходимо отчетливо представлять себе, что такое определенный интеграл и каков его геометрический смысл.

Нам понадобятся понятия матрицы и вектора, как частного случая матрицы, одно из измерений которой равно 1. Если у матрицы число

строк равно единице, то мы будем говорить, что перед нами вектор-строка, если же число столбцов равно единице – то вектор-столбец. Как правило, записывая $\vec{a}$, мы будем подразумевать, если не оговорено иное, что это вектор-столбец. Конечно же, геометрическая интерпретация вектора, как направленного отрезка и вектора единичной длины как направления остается в силе.

Знак ^{t}, расположенный справа от обозначения матрицы, будет означать транспонирование, т.е. такую операцию, при которой строки матрицы становятся ее столбцами и наоборот. В частности, матрица A симметрична тогда и только тогда, когда $A^t = A$, а $\vec{a}^t$ – вектор-строка.

Символом $\cdot$ условимся обозначать скалярное произведение векторов, т.е. если $\vec{a} = (a_1, ..., a_n)^t$, $\vec{b} = (b_1, ..., b_n)^t$, то

$$\vec{a} \cdot \vec{b} = \sum_{j=1}^{n} a_j b_j = \|\vec{a}\| \|\vec{b}\| \cos\psi,$$

где ψ – угол между векторами $\vec{a}$ и $\vec{b}$, а символом $\|\vec{a}\|$ обозначена длина вектора:

$$\|\vec{a}\| = \sqrt{\sum_{j=1}^{n} a_j^2}.$$

Векторы называются ортогональными, если их скалярное произведение равно 0. Заметим также, что всегда

$$\vec{a} \cdot \vec{b} = \vec{a}^t \vec{b},$$

где слева векторы перемножаются по правилу умножения матриц.

Используя тот легко проверяемый (например, сравнением каждого из элементов левой и правой части равенства) факт, что при транспонировании порядок умножаемых матриц меняется, т.е.

$$(AB)^t = B^t A^t,$$

легко доказать следующее утверждение, позволяющее "перебрасывать"матричный множитель с одного аргумента скалярного произведения на другой:

Лемма 1 *Для произвольных векторов $\vec{a}, \vec{b}$ и квадратных матриц A, B справедливо*

$$A\vec{a} \cdot \vec{b} = \vec{a} \cdot A^t \vec{b}, \quad \vec{a} \cdot B\vec{b} = B^t \vec{a} \cdot \vec{b}.$$

Действительно,

$$A\vec{a} \cdot \vec{b} \;=\; (A\vec{a})^t\vec{b} \;=\; \vec{a}^t(A^t\vec{b}) \;=\; \vec{a} \cdot A^t\vec{b},$$

и первое утверждение доказано. Второе утверждение легко получается из первого, если мы вспомним, что скалярное произведение коммутативно и $(B^t)^t = B$.

Весьма полезным является также и следующий факт.

Лемма 2 (Неравенство Коши-Буняковского) *Для любых векторов*

$$|\vec{a} \cdot \vec{b}| \;\leq\; \|\vec{a}\|\|\vec{b}\|,$$

причем равенство достигается в том и только том случае, когда найдется такое число λ, что

$$\vec{a} \;=\; \lambda\vec{b}.$$

Доказательство. Введем в рассмотрение

$$\vec{a}^* \;=\; \frac{\vec{a}}{\|\vec{a}\|},\;\; \vec{b}^* \;=\; \frac{\vec{b}}{\|\vec{b}\|}.$$

Тогда $\|\vec{a}^*\| = \|\vec{b}^*\| = 1$ и

$$\|\vec{a}^* \pm \vec{b}^*\|^2 \;=\; (\vec{a}^* \pm \vec{b}^*) \cdot (\vec{a}^* \pm \vec{b}^*) \;=\; 2 \pm 2\vec{a}^* \cdot \vec{b}^* \geq 0,$$

откуда

$$|\vec{a}^* \cdot \vec{b}^*| \;\leq 1,$$

и достаточно обе части этого неравенства умножить на $\|\vec{a}\|\|\vec{b}\|$, чтобы получить неравенство Коши-Буняковского.

Если нашлось такое λ, что $\vec{a} = \lambda\vec{b}$, то равенство проверить нетрудно. Обратно, если в неравенстве Коши-Буняковского достигнуто равенство, то из приведенной выше выкладки следует, что $\|\vec{a}^* \pm \vec{b}^*\| = 0$, откуда $\vec{a}^* \;=\; \mp\vec{b}^*$, и можно выбрать

$$\lambda \;=\; \mp\frac{\|\vec{a}\|}{\|\vec{b}\|}.$$

Лемма доказана.

Говорят, что векторы $\vec{a}_1, ..., \vec{a}_n$ образуют ортонормированный базис n-мерного пространства, если $\|\vec{a}_j\| = 1,\ j = 1, ..., n$ и при $i \neq j\ \vec{a}_i \cdot \vec{a}_j = 0$.

С геометрической точки зрения умножение квадратной матрицы на вектор можно рассматривать как поворот и растяжение, т.е. умножение на матрицу представляет собой своего рода преобразование n-мерного пространства. Для некоторых из векторов это преобразование является "чистым"растяжением. Такие векторы называются собственными векторами матрицы, а соответствующие коэффициенты растяжения – ее собственными числами. Формально: пусть A – квадратная матрица. Вектор $\vec{a}$ называется ее собственным вектором, отвечающим собственному числу λ, если

$$A\vec{a} = \lambda\vec{a}.$$

Известно, что симметричная матрица имеет ортонормированный базис из собственных векторов.

Матрица называется положительно определенной, если для любого ненулевого вектора

$$A\vec{a} \cdot \vec{a} > 0.$$

Из положительно определенной матрицы можно извлечь квадратный корень, т.е. можно возвести ее в степень со знаменателем 2. По определению, $A^{1/2}$ – это такая матрица, что

$$A^{1/2}A^{1/2} = A.$$

Понятие определителя квадратной матрицы мы здесь напоминать не будем. Введем лишь обозначение $|A|$ для него. Если $|A| \neq 0$, то у матрицы имеется обратная, т.е. такая A^{-1}, что

$$AA^{-1} = A^{-1}A = I,$$

где через I обозначена единичная матрица. Иногда нам будет необходимо подчеркнуть, каков порядок этой матрицы, тогда он будет фигурировать при ней в качестве индекса (I_n).

Принципиально важной для дальнейшего является следующая теорема, обычно не включаемая в курсы линейной алгебры:

Теорема 1 *Пусть матрица A положительно определена. Тогда*

$$\max_{\|\vec{x}\|=1} A\vec{x} \cdot \vec{x} = \lambda_1,$$

где λ_1 – наибольшее собственное число матрицы, причем этот максимум достигается при $\vec{x} = \vec{e}_1$, на собственном векторе, отвечающем этому собственному числу.

Как известно, матрицы – это прямоугольные таблицы, и положения элемента матрицы можно указать, задавая пару индексов. В силу этого матрицы можно называть двухвходовыми таблицами. Иногда (и нам встретятся такие случаи) характеризация некоторых объектов производится по трем и более параметрам и для того, чтобы указать расположение объекта в пространстве этих параметров, нужно указать три или более индексов. Тогда мы говорим, что задана трех- или многовходовая таблица. Представлять себе трехвходовую таблицу можно как прямоугольный параллелепипед, у которого по каждому из "слоев"располагается таблица двухвходовая, т.е. матрица. Чаще всего, при необходимости трехвходового табулирования, приводят именно эти матрицы - "слои"(см. таблицы распределения Фишера в приложении).

1.2 Теория вероятностей

Как уже упоминалось выше, курс теории вероятностей нужен нам будет целиком. Его понятия будут постоянно встречаться и использоваться, при этом содержание некоторых из них напоминается по мере их использования. Здесь вкратце упомянем лишь наиболее важные.

Случайной величиной ξ мы называем действительную функцию исхода случайного эксперимента, обладающую тем свойством, что для произвольного вещественного числа x возможно вычислить вероятность события $\{\xi < x\}$. (Это свойство математики называют измеримостью). Функция $F(x) = \mathbf{P}(\xi < x)$ называется функцией распределения случайной величины.

Мы будем часто использовать понятие квантили случайной величины или ее функции распределения. Пусть $\alpha \in (0, 1)$. Число t_α называется квантилью уровня α распределения F, если

$$F(t_\alpha) = \alpha.$$

Пусть число α достаточно близко к единице. Тогда можно интерпретировать последнее соотношение так: рассматриваемая случайная величина

ξ почти наверняка окажется меньше, чем t_α, и если вдруг она оказывается больше, значит, произошло что-то экстраординарное. Поэтому число $t(\beta)$, определяемое соотношением

$$\mathbf{P}(\xi \geq t(\beta)) \;=\; \beta$$

называют критической точкой распределения F. Нетрудно выписать связь между понятиями квантили и критической точки одного и того же распределения:

$$t_\alpha \;=\; t(1-\alpha).$$

Если множество значений случайной величины конечно или счетно, т.е. ее значения можно перечислить, то говорят, что она имеет дискретный тип распределения. Дискретный тип принято характеризовать рядом распределения - таблицей, в которой перечислены все значения случайной величины и указаны их вероятности.

Если нашлась такая неотрицательная функция p, что при произвольном x справедливо

$$F(x) \;=\; \int\limits_{-\infty}^{x} p(t)dt,$$

то она называется плотностью распределения ξ, а сама случайная величина называется величиной абсолютно непрерывного типа. Качественной интерпретацией, заодно несколько объясняющей терминологию, может служить представление о случайной величине, как о механизме, распределяющем единичную массу по точкам числовой прямой. С этой точки зрения функция распределения есть масса, расположенная левее точки x, а плотность распределения представляет собой (линейную) плотность в физическом понимании этого термина.

Исходя из этого представления, можно ввести некое подобие плотности в дискретном случае. Такая "плотность"должна быть равна нулю во всех точках числовой прямой кроме точек-значений ξ, в которых равна бесконечности, но так, что интеграл от нее по всей прямой равен 1. Более строго такую "плотность"можно построить при помощи δ-функции Дирака. В этом смысле два главных типа распределений похожи.

Если функция распределения непрерывна, но почти всюду постоянна, то случайная величина называется сингулярной. В одномерном случае сингулярность встречается достаточно редко, но в многомерной ситуации

этот феномен приобретает гораздо большую важность. Подробнее см. в разделе 2.1.

Математическим ожиданием дискретной случайной величины называют число

$$\mathbf{M}\xi = \sum_i x_i \mathbf{P}(\xi = x_i),$$

где сумма берется по всем значениям ξ, а для абсолютно непрерывного случая с плотностью p

$$\mathbf{M}\xi = \int_{-\infty}^{\infty} xp(x)dx.$$

Во всех случаях с точки зрения "массовой"интерпретации математическое ожидание является координатой центра масс, получающейся при действии ξ системы. Вместо понятия "математическое ожидание"часто пользуются понятием "среднее".

Величина

$$\mathbf{D}\xi = \mathbf{M}(\xi - \mathbf{M}\xi)^2$$

называется дисперсией и служит мерой изменчивости случайной величины, точнее, мерой разброса ее значений вокруг среднего. Нулевая дисперсия бывает только у постоянных величин. С точки зрения "массовой"интерпретации дисперсия - момент инерции системы масс относительно центра масс.

1.3 Математическая статистика

В математической статистике исходным материалом служат результаты n независимых наблюдений над случайной величиной ξ. Набор результатов этих наблюдений $X = (x_1, ..., x_n)$ называют выборкой объема n из распределения ξ (или из генеральной совокупности). С теоретической точки зрения выборка X - n-мерный случайный вектор с независимыми координатами, каждая из которых имеет то же распределение, что и наблюдаемая величина.

В самой общей формулировке задач математической статистики больше ничего не предполагается известным. Тогда принято заменять величину ξ на ее эмпирический (выборочный) аналог ξ^* – дискретную случайную величину, принимающую только значения, оказавшиеся в выборке, с равными вероятностями. При этом, конечно же, если какое-то

из значений в выборке встречается несколько раз, то соответствующие вероятности увеличиваются. Например, если $X = (1, 2, 1, 3, 1, 1, 3)$, то ξ^* определяется следующим рядом распределения:

ξ^*	1	2	3
	4/7	1/7	2/7

Все характеристики ξ в этом контексте называют теоретическими, а соответствующие характеристики ξ^* - эмпирическими или выборочными. В частности, эмпирическое математическое ожидание и дисперсия задаются, соответственно, формулами

$$\bar{X} = \frac{1}{n}\sum_{i=1}^{n} x_i; \quad S^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{X})^2.$$

Для получения оценки некоей теоретической характеристики чаще всего достаточно произвести в алгоритме ее вычисления замену всех вхождений ξ на ξ^* (метод подстановки). Отметим, что оценкой обычно называют достаточно произвольную функцию выборки, предназначенную для использования вместо соответствующей теоретической характеристики. Таким образом, оценка всегда является случайным объектом, поскольку выборка является таковым. При этом крайне желательными свойствами оценок считаются их состоятельность, т.е. приближение к истинным значениям теоретических характеристик при увеличении объема выборки, и несмещенность, т.е. отсутствие систематической ошибки. Формально свойство несмещенности записываем так. Пусть θ – теоретическая (неизвестная) характеристика, θ^* – ее оценка. Она называется несмещенной, если

$$\mathbf{M}\theta^* = \theta.$$

Для проверки различного рода гипотез о распределениях строятся критерии, т.е. правила, позволяющие по выборке принять решение о справедливости одной из выдвигаемых гипотез. Если гипотез всего две - основная и альтернативная, то вместо критерия можно строить критическое множество - набор тех выборок, на которых основную гипотезу следует отвергнуть.

Чаще всего для построения критического множества выбирается мера отклонения наблюдаемой выборки от ожидаемой в предположении

основной гипотезы. Если выбранная мера в нашем эксперименте оказывается "слишком большой", то гипотезу отвергают, обосновывая это отсутствием согласования между наблюдаемым и ожидаемым. Термин "слишком большая"приобретает точный смысл при обращении к статистическим таблицам, в которых собраны предельно допустимые значения выбранных мер отклонения. Естественно, проверке каждой гипотезы должно предшествовать теоретическое исследование и выбор (или построение) соответствующей статистической таблицы.

Многие критерии, а также лучшие из оценок неизвестных параметров распределений, получаются на основе функции правдоподобия. Она определяется в случае, когда класс изучаемых распределений состоит либо из дискретных, либо из абсолютно непрерывных распределений.

Формально. Введем обозначение. Пусть $f(x, \theta)$ обозначает значение плотности в точке x для абсолютно непрерывного случая и равна значению вероятности того, что наблюдаемая величина приняла значение x, в дискретном случае. Функцией правдоподобия, построенной по выборке X объема n называется

$$L(X, \theta) \ = \ \prod_{i=1}^{n} f(x_i, \theta).$$

Оценкой максимального правдоподобия для θ называется то значение θ^*, для которого функция L достигает своего максимума как функция аргумента θ. В этих определениях все участвующие объекты могут быть и векторными величинами.

Глава 2

Общая теория многомерных распределений

2.1 Случайные векторы

Пусть имеется вектор $\vec{\xi} = (\xi_1, ..., \xi_n)$, каждая координата которого представляет из себя (одномерную) случайную величину. Тогда мы говорим, что $\vec{\xi}$ является случайным вектором или n-мерной случайной величиной. Распределение $\vec{\xi}$ по отношению к распределениям $\xi_1, ..., \xi_n$ называется совместным, а распределение каждой из координат по отношению к распределению случайного вектора - маргинальным. Уместно заметить здесь, что под словом "распределение"здесь и ниже будет пониматься либо (интегральная) функция распределения, либо плотность распределения в случае абсолютно непрерывных случайных величин, либо ряд распределения, если величины рассматриваются дискретные. Как известно из курса теории вероятностей, знания любой из перечисленных характеристик достаточно для восстановления другой. Отметим также, что совместное распределение полностью определяет маргинальные, а обратное неверно. Функции маргинальных распределений $F_j(x),\ j = 1, ..., n$ определяются по функции совместного распределения F формулами

$$F_j(t) \ = \ \lim_{*} F(x_1, ..., x_{j-1}, t, x_{j+1}, ..., x_n),$$

где предел берется в предположении, что все переменные, кроме j-й, стремятся к бесконечности. Контрпримером к обратному утверждению служит следующий: пусть подбрасываются две монеты, ξ_1 - число гербов,

выпавших на первой из них, ξ_2 - число решек на ней же, ξ_3 - число гербов, выпавших на второй монете. Тогда ряды распределений ξ_i, $i = 1, 2, 3$ совпадают, но совместное распределение ξ_1 и ξ_2 не такое, как у ξ_1 и ξ_3.

Ряды распределений (ξ_1, ξ_2) и (ξ_1, ξ_3).

ξ_1/ξ_2	0	1
0	0	1/2
1	1/2	0

ξ_1/ξ_3	0	1
0	1/4	1/4
1	1/4	1/4

Приведем также формулы, позволяющие восстанавливать ряды распределений координат по рядам распределений векторов в дискретном случае и плотности распределений координат по плотности совместного распределения в абсолютно непрерывном случае для двумерных векторов. Дискретный вариант формул:

$$P(\xi_1 = a) \;=\; \sum_b P(\xi_1 = a, \xi_2 = b), \quad P(\xi_2 = b) \;=\; \sum_a P(\xi_1 = a, \xi_2 = b).$$

Здесь суммы берутся по всем возможным значениям ξ_2 в первом случае и по всем значениям ξ_1 во втором. В случае, когда координат больше двух, суммирование должно производиться по всем возможным наборам значений остальных (кроме извлекаемой) случайных величин.

Для абсолютно непрерывного случая формулы плотностей маргинальных распределений имеют вид:

$$p_{\xi_1}(x) \;=\; \int\limits_{-\infty}^{\infty} p_{\xi_1,\xi_2}(x,y)dy, \quad p_{\xi_2}(y) \;=\; \int\limits_{-\infty}^{\infty} p_{\xi_1,\xi_2}(x,y)dx.$$

Под интегралом здесь стоит плотность совместного распределения. Если координат больше двух, то формулы изменяются так же, как и в дискретном случае: интегрирование становится $(n-1)$-кратным, и интегралы берутся по всем переменным, кроме извлекаемой.

В многомерном случае наблюдается еще один интересный феномен, редко встречающийся в одномерных практических задачах. Сначала отметим, что дискретный тип распределений координат необходимо приводит к дискретному типу совместного распределения и наоборот. Если совместное распределение абсолютно непрерывно, то, как следует из формул, приведенных выше, маргинальные распределения также являются таковыми. А вот здесь обратное неверно. Более того, нетрудно привести пример, когда совместное распределение сосредоточено на некоторой кривой в пространстве и, несмотря на наличие плотностей каждой

из координат, является сингулярным. Простейший пример такого рода: пусть ξ_1 имеет равномерное распределение на отрезке $[0, 1]$, $\xi_2 = \xi_1$. Тогда вектор (ξ_1, ξ_2) распределен на диагонали единичного квадрата и имеет сингулярный тип распределения. Как видно из этого примера, подобного рода явления возникают, например, при наличии функциональной связи между координатами вектора.

Рассмотрим бросание точки в единичный квадрат. Очевидно, что при таком эксперименте с точки зрения классической теории вероятностей вероятность попадания точки на диагональ равна 0. Тем не менее, на практике такие точки иногда встречаются. Связано это с тем, что эксперимент производится человеком, а ему иногда (возможно, неосознанно) хочется, чтобы результат получился попроще, и точки, которые при ближайшем рассмотрении оказываются просто близки к точкам диагонали, принимаются им за точки диагонали. Итак, вмешательство субъективного фактора приводит к тому, что сингулярные распределения приобретают в многомерном случае большее практическое значение. Другой вывод, который следует из рассмотренного примера - события нулевой вероятности на практике иногда все же случаются.

2.2 Независимость

Самый простой для изучения случай возникает тогда, когда координаты случайного вектора независимы, т.е. для произвольного набора $x_1, ..., x_n$ функция совместного распределения распадается в произведение функций маргинальных распределений:

$$P(\xi_1 < x_1, ..., \xi_n < x_n) \;=\; P(\xi_1 < x_1)P(\xi_2 < x_2)...P(\xi_n < x_n).$$

Здесь, конечно же, маргинальные распределения полностью определяют распределение совместное. Независимость, которая только что была определена, и ее соотношение с понятием независимости в практическом смысле под которым обычно понимается отсутствие связи, является одним из наиболее содержательных понятий теории вероятностей. Наличие "теоретической"независимости обеспечивает независимость "практическую". Обратное, к сожалению, неверно. Дело в том, что практическая независимость требует только отсутствие четкой, ярко выраженной связи между величинами, теоретическая же предполагает отсутствие даже очень слабого, невыраженного влияния.

Проверка независимости наблюдаемых величин - одна из очень часто встречающихся задач, с которой нам придется иметь дело постоянно. Приведем "теоретические"критерии независимости дискретных и абсолютно непрерывных величин, поскольку проверка полного определения, приведенного выше, слишком сложна. Дискретные случайные величины независимы тогда и только тогда, когда для произвольного набора $x_1, ..., x_n$ их значений выполнено

$$P(\xi_1 = x_1, ..., \xi_n = x_n) \;=\; P(\xi_1 = x_1)P(\xi_2 = x_2)...P(\xi_n = x_n).$$

Абсолютно непрерывные случайные величины независимы тогда и только тогда, когда плотность совместного распределения распадается в произведение маргинальных плотностей, т.е. для любого набора $x_1, ..., x_n$ справедливо

$$p_{\xi_1,...,\xi_n}(x_1, ..., x_n) \;=\; p_{\xi_1}(x_1)p_{\xi_2}(x_2)...p_{\xi_n}(x_n).$$

Эти критерии охарактеризованы как "теоретические", поскольку они предполагают умение точно вычислять вероятности, связанные со случайными величинами (или знание плотностей распределений). На практике же у нас такой возможности нет, а так называемый статистический способ вычисления вероятностей, связанный с повторениями эксперимента и подсчетом количеств появлений изучаемого события, предполагает независимость упомянутых повторений, а значит, вновь упираясь в понятие независимости, мы получаем замкнутый круг, разумного выхода из которого в рамках этого подхода не существует.

Противоречие, описанное выше, было отмечено еще на ранних этапах создания аксиоматической теории вероятностей. Существует так называемая вторая аксиоматика Колмогорова, в основу которой положена аксиоматизация понятия независимости.

Естественным образом понятие независимости обобщается на случайные векторы. Векторы $\vec{\xi}, \vec{\eta}$ называются независимыми, если для достаточно произвольных подмножеств A, B n-мерного пространства справедливо

$$P(\vec{\xi} \in A,\ \vec{\eta} \in B) \;=\; P(\vec{\xi} \in A)P(\vec{\eta} \in B).$$

В рамках строго математического подхода к последнему определению класс подмножеств, о которых говорится в нем, естественно, указывается совершенно четко. Но нам достаточно знать, что все "обычные"подмножества,

т.е. те, что могут встретиться в практических задачах, в этот класс попадают. Можно доказать, что такая независимость эквивалентна тому, что каждая из координат вектора $\vec{\xi}$ не зависит от совокупности координат вектора $\vec{\eta}$.

2.3 Некоторые числовые характеристики случайных векторов

Как и в одномерном случае, для векторов вводится понятие математического ожидания $\mathbf{M}\vec{\xi} = (\mathbf{M}\xi_1, ..., \mathbf{M}\xi_n)^t$, которое обладает обычным набором свойств и, если представлять себе случайный вектор $\vec{\xi}$ как транспорт, распределяющий единичную массу по n-мерному пространству, то $\mathbf{M}\vec{\xi}$ будет радиус-вектором центра масс получившейся системы. Выпишем свойства математического ожидания в многомерном случае:

1. $\mathbf{M}(\vec{\xi} + \vec{\eta}) = \mathbf{M}\vec{\xi} + \mathbf{M}\vec{\eta}$;
2. Если λ - произвольное действительное число, то $\mathbf{M}\lambda\vec{\xi} = \lambda\mathbf{M}\vec{\xi}$;
3. Для независимых $\vec{\xi}, \vec{\eta}$ справедливо $\mathbf{M}(\vec{\xi} \cdot \vec{\eta}) = \mathbf{M}\vec{\xi} \cdot \mathbf{M}\vec{\eta}$, где, как обычно, символом $\cdot$ обозначено скалярное произведение;
4. Если распределение $\vec{\xi}$ абсолютно непрерывно, и f - измеримая функция, заданная в n-мерном пространстве и принимающая значения в области действительных чисел, то

$$\mathbf{M}f(\vec{\xi}) = \int_{-\infty}^{\infty} ... \int_{-\infty}^{\infty} f(\vec{x})p(\vec{x})dx_1...dx_n,$$

   Для дискретных распределений в последней формуле надо интегрирование заменить n-кратным суммированием по всем возможным наборам значений координат вектора $\vec{\xi}$;
5. $\|\mathbf{M}\vec{\xi}\| \leq \mathbf{M}\|\vec{\xi}\|$, где , как и выше, через $\|\vec{a}\|$ обозначена длина вектора $\vec{a}$, т.е. корень квадратный из суммы квадратов его координат.

Естественно, список свойств можно продолжить.

Аналогом понятия дисперсии в многомерном случае служит так называемая ковариационная матрица (или матрица ковариаций), элементы которой вычисляются так:

$$(\mathbf{cov}\vec{\xi})_{i,j} \;=\; \mathbf{cov}(\xi_i,\xi_j) \;=\; \mathbf{M}\xi_i\xi_j \;-\; \mathbf{M}\xi_i\mathbf{M}\xi_j,\; i,j=1,...,n.$$

Очевидно, если ξ_i, ξ_j независимы, то $(\mathbf{cov}\vec{\xi})_{i,j} \;=\; 0$.

Напомним, что равенство нулю элемента ковариационной матрицы не означает независимости координат. Если ковариация между двумя случайными величинами равна нулю, то такие величины называют некоррелированными. Свойство некоррелированности - более слабое свойство, чем независимость. Как известно, ковариации отвечают за линейную часть формулы, связывающей случайные величины (подробнее смотри раздел "Регрессионный анализ"), и "не замечают"зависимости, описываемые формулами более высоких степеней.

Свойства ковариационной матрицы:

1. ковариационная матрица симметрична;

2. на ее диагонали стоят дисперсии соответствующих координат;

3. ковариационная матрица неотрицательно определена. Действительно, если $C \;=\; \mathbf{cov}(\vec{\xi})$ и $\vec{t} \;=\; (t_1,...,t_n)^t$ - постоянный вектор, то

$$C\vec{t}\cdot\vec{t} \;=\; \sum_{i,j=1}^{n} C_{i,j}t_it_j \;=\; \mathbf{D}\left(\sum_{k=1}^{n} t_k\xi_k\right) \;\geq 0.$$

4. $\mathbf{cov}(\lambda\vec{\xi}) = \;\lambda^2\mathbf{cov}\vec{\xi}$ для произвольного числа λ.

Иногда вместо ковариационных матриц рассматриваются матрицы корреляционные, у которых на $(i,\; j)$-м месте стоит коэффициент корреляции между i-й и j-й координатой случайного вектора. Коэффициенты корреляции

$$\rho(\xi,\eta) \;=\; \frac{\mathbf{cov}(\xi,\eta)}{\sqrt{\mathbf{D}\xi\mathbf{D}\eta}}$$

в отличие от ковариаций замечательны тем, что могут принимать только значения от -1 до 1, причем, поскольку независимым величинам соответствует нулевой коэффициент корреляции, а крайние значения (+1 и -1)

коэффициента корреляции соответствуют наличию линейной зависимости, которая является самой сильной из всех возможных форм зависимости, принято считать эти коэффициенты мерой зависимости изучаемых величин. Как уже было указано выше, эта мера на самом-то деле описывает только линейную часть зависимости, но в некоторых важных частных случаях (смотри раздел 2.4) этого бывает достаточно. Приведем шкалу, по которой качественно оценивается линейная зависимость случайных величин (она не раз нам пригодится ниже).

Качественная оценка связи по коэффициенту корреляции

Интервал значений	Характеристика связи
$[-1, -2/3]$	сильная отрицательная
$(-2/3, -1/3]$	умеренная отрицательная
$(-1/3, 1/3)$	слабая
$[1/3, 2/3)$	умеренная положительная
$[2/3, 1]$	сильная положительная

Удобно бывает привести рассматриваемый вектор $\vec{\xi}$ к такому виду, когда математическое ожидание его будет равно $\vec{0}$. Соответствующий процесс называется центрированием. Полагаем $\vec{\xi'} = \vec{\xi} - \mathbf{M}\vec{\xi}$. Можно также добиться того, чтобы результирующий вектор имел единичную матрицу ковариаций (нормирование). Для этого полагают $\vec{\xi''} = V^{-1/2}\vec{\xi'}$, где $V = \mathbf{cov}(\vec{\xi})$. Вектор, подвергнутый процедурам центрирования и нормирования, называют стандартизованным.

2.4 Нормальное распределение в многомерном случае

Пусть A - квадратная $n \times n$ матрица, являющаяся положительно определенной, через $|A|$ обозначен ее определитель. Тогда n-мерное распределение с плотностью

$$p(\vec{x}) = \sqrt{\frac{|A|}{(2\pi)^n}} \exp\left\{-\frac{1}{2}A(\vec{x}-\vec{a})\cdot(\vec{x}-\vec{a})\right\}$$

называется нормальным со средним $\vec{a}$ и ковариационной матрицей $V = A^{-1}$. Если вектор среднего нулевой, то соответствующее распределение называют центрированным.

Нетрудно убедиться (например, интегрированием плотности в пределах от $-\infty$ до ∞ по всем переменным, кроме выбранных), что распределение каждого подвектора нормального вектора ξ также нормально. В частности, каждая из его координат $\xi_i, i = 1, ..., n$ имеет нормальное распределение, а значит, диагональные элементы A^{-1} равны дисперсиям соответствующих координат.

Теорема 2 *Пусть вектор $\vec{\xi}$ имеет нормальное распределение. Если для произвольной пары индексов i, j справедливо $\rho(\xi_i, \xi_j) = 0$, то координаты $\xi_1, ..., \xi_n$ независимы.*

Доказательство. Из условия следует, что матрица $\mathbf{cov}\vec{\xi}$ диагональна, а значит и матрица A тоже диагональна. Обозначим ее диагональные элементы $\sigma_1^{-2}, ..., \sigma_n^{-2}$. Тогда в силу замечания, сделанного перед формулировкой теоремы, $\sigma_j^2 = \mathbf{D}\xi_j$, $j = 1, ..., n$. Очевидно, что

$$|A| = \prod_{j=1}^{n} \sigma_j^2, \quad \frac{1}{2}A(\vec{x} - \vec{a}) \cdot (\vec{x} - \vec{a}) = \sum_{j=1}^{n} \frac{1}{2\sigma_j^2}(x_j - a_j)^2.$$

Отсюда

$$p(\vec{x}) = \prod_{j=1}^{n} \frac{1}{2\sigma_j^2} \exp\left\{-\frac{1}{2}(x_j - a_j)^2\right\} = \prod_{j=1}^{n} p_{\xi_j}(x_j),$$

что и означает независимость.

Итак, в этом случае понятия независимости и некоррелированности совпадают.

Лемма 3 *Если $\vec{\xi}$ имеет n-мерное нормальное распределение с вектором средних $\vec{a}$ и ковариационной матрицей V, то и вектор $\vec{\eta} = V^{-1/2}(\vec{\xi} - \vec{a})$ имеет стандартное n-мерное распределение, т.е. нулевой вектор средних и единичную ковариационную матрицу. В частности, его координаты – стандартные нормальные случайные величины.*

Доказательство этой леммы легко получается заменой переменных в n-кратном интеграле, выражающем вероятность попадания $\vec{\eta}$ в измеримое множество.

Теорема 3 *В условиях леммы 3 случайная величина* $V^{-1}(\vec{\xi}-\vec{a})\cdot(\vec{\xi}-\vec{a})$ *имеет распределение хи-квадрат с* n *степенями свободы.*

Доказательство. В силу свойств скалярного произведения и симметричности матрицы V имеем

$$V^{-1}(\vec{\xi}-\vec{a})\cdot(\vec{\xi}-\vec{a}) \;=\; V^{-1/2}(\vec{\xi}-\vec{a})\cdot(\vec{\xi}-\vec{a}) \;=\; \vec{\eta}\cdot\vec{\eta}.$$

При этом согласно лемме 3 и определению распределения хи-квадрат (приведено в приложении 2) утверждение теоремы следует из соотношения

$$\vec{\eta}\cdot\vec{\eta} \;=\; \sum_{i=1}^{n}\eta_i^2.$$

Сделаем следующее простое замечание. В случае, когда мы имеем двумерный нормальный вектор, плотность его распределения с точностью до констант перед экспонентой равна

$$\exp\{-\frac{1}{2}\,(A_{1,1}x_1^2+2A_{1,2}x_1x_2+A_{2,2}x_2^2)\},$$

где $A_{i,j}$ - элементы матрицы, обратной ковариационной:

$$A \;=\; V^{-1} \;=\; \begin{pmatrix} \frac{\sigma_1^2}{|V|} & -\frac{c_{1,2}}{|V|} \\ -\frac{c_{1,2}}{|V|} & \frac{\sigma_1^2}{|V|} \end{pmatrix},$$

$$c_{1,2} \;=\; \mathbf{cov}(\xi_1,\,\xi_2).$$

Отсюда, в частности, следует, что

$$\rho(\xi_1,\,\xi_2) \;=\; \frac{c_{1,2}}{\sigma_1\sigma_2} \;=\; -\frac{A_{1,2}}{\sqrt{A_{1,1}A_{2,2}}}. \tag{2.1}$$

Случай многомерного нормального распределения замечателен тем, что он очень часто встречается при изучении природных процессов. Одно время даже считалось, что практически все природные явления согласуют свое поведение именно с этим распределением. И даже сегодня, когда стало ясно, что такое предположение далеко от действительности, исследователи-практики, попав в незнакомую ситуацию, для начала пытаются применить к ней аппарат, разработанный для нормального случая. Вторая замечательная особенность рассматриваемой ситуации состоит в том, что для нормальных распределений многие сложные формулы приобретают относительно простой вид - все можно рассчитать в явном виде.

2.5 Корреляционная теория

В случае, когда изучаются только две случайные величины, их взаимосвязь (по крайней мере, ее линейная часть) описывается коэффициентом корреляции между ними. Если же размерность задачи больше двух, то можно ввести характеристики, похожие на этот коэффициент, которые описывают взаимодействие случайных векторов, координатами которых являются исследуемые величины. Эти коэффициенты называют коэффициентами частной и множественной корреляции.

К сожалению, простые формулы, определяющие эти коэффициенты, выписать невозможно, поэтому мы разовьем соответствующий подход на примерах.

Начнем с двумерного случая. Двумерное центрированное нормальное распределение (ξ_1, ξ_2) имеет плотность

$$p(x_1, x_2) \;=\; \sqrt{\frac{|A|}{2\pi}}\;\exp\{-\frac{1}{2}(A_{1,1}x_1^2 + 2A_{1,2}x_1x_2 + A_{2,2}x_2^2)\},$$

где A – матрица, обратная ковариационной. Т.е., если

$$V \;=\; \begin{pmatrix} \sigma_1^2 & \operatorname{cov}(\xi_1,\xi_2) \\ \operatorname{cov}(\xi_1,\xi_2) & \sigma_2^2 \end{pmatrix},$$

то $A_{i,j} \;=\; \frac{c_{i,j}}{|V|}$, $i,j = 1,2$, где $c_{i,j}$ – алгебраическое дополнение к месту (i,j) матрицы V. В нашем случае

$$c_{1,1} = \sigma_2^2, \quad c_{1,2} = -\operatorname{cov}(\xi_1,\xi_2), \tag{2.2}$$

и, с точностью до констант,

$$p(x_1, x_2) \;\asymp\; \exp\{-\frac{1}{2|V|}\,(c_{1,1}x_1^2 + 2c_{1,2}x_1x_2 + c_{2,2}x_2^2)\}. \tag{2.3}$$

Итак, в двумерном случае из (2.2) и (2.3) получаем, что

$$\rho(\xi_1,\,\xi_2) \;=\; -\;\frac{c_{1,2}}{\sqrt{c_{1,1}c_{2,2}}}. \tag{2.4}$$

Теперь рассмотрим случай трехмерного центрированного нормального вектора $\vec{\xi} \;=\; (\xi_1, \xi_2, \xi_3)$ и поставим задачу оценить связь между ξ_1 и ξ_2 при фиксированных остальных координатах - в данном случае ξ_3. У

условной плотности ξ_1, ξ_2 при фиксированном $\xi_3 = x_3$, вычисляемой по формуле

$$p_{\xi_1,\xi_2/\xi_3=x_3}(x_1, x_2) \;=\; p(x_1, x_2|x_3) \;=\; \frac{p(x_1, x_2, x_3)}{\int\limits_{-\infty}^{\infty} p(x_1, x_2, x_3)dx_3},$$

показатель степени экспоненты имеет вид

$$-\frac{1}{2|V|}\left(c_{1,1}(x_1 - b_1)^2 \;+\; 2c_{1,2}(x_1 - b_1)(x_2 - b_2) \;+\; c_{2,2}(x_2 - b_2)^2\right), \quad (2.5)$$

где b_1, b_2 – решения системы

$$c_{1,1}b_1 + c_{1,2}b_2 \;=\; -c_{1,3}x_3, \quad c_{1,2}b_1 + c_{2,2}b_2 \;=\; -c_{2,3}x_3.$$

Эта система всегда имеет единственное решение в силу положительной определенности ковариационных матриц. Из (2.5) и (2.1), (2.3) следует, что при фиксированном x_3 величины ξ_1, ξ_2 имеют двумерное нормальное распределение и коэффициент корреляции между ними задается при помощи (2.4) :

$$R_{12,(3)} \;=\; -\frac{c_{1,2}}{\sqrt{c_{1,1}c_{2,2}}}. \quad (2.6)$$

Этот коэффициент называют частным коэффициентом корреляции между ξ_1, ξ_2 при фиксированном ξ_3. Такой коэффициент всегда определяется формулой (2.6), даже если мы отказываемся от предположения нормальности рассматриваемых величин. Под $c_{i,j}$, как и выше, мы понимаем алгебраическое дополнение к месту (i, j) в ковариационной матрице V. Можно также распространить определение (2.6) на случай числа измерений, большего трех. При этом определяемый коэффициент обозначается $R_{12,(34\ldots p)}$ а в роли V выступает ковариационная матрица, имеющая в этом случае порядок p. Аналогично могут быть определены частные коэффициенты корреляции для любой пары выбранных координат при фиксированных остальных координатах.

Вычислив частные коэффициенты корреляции, мы считаем, что охарактеризовали "очищенную"связь между, например, ξ_1 и ξ_2, убрав из нее влияние оставшихся переменных. В том трехмерном варианте, который мы рассматривали выше, можно отметить следующую связь между частным и обычными коэффициентами корреляции:

$$R_{12,(3)} \;=\; \frac{\rho_{12} - \rho_{13}\rho_{23}}{\sqrt{(1 - \rho_{13}^2)(1 - \rho_{23}^2)}}.$$

Подобные связи через обычные коэффициенты корреляции могут быть получены и для остальных частных коэффициентов. Можно предложить следующую геометрическую интерпретацию этой теории, которая поясняет смысл полученных формул в случае, когда рассматриваемые коэффициенты оцениваются по результатам n наблюдений над изучаемыми величинами.

Будем представлять $\xi_i,\ i = 1, ..., p (p \geq 3)$ векторами $O\vec{Q}_i$ в $n-$мерном пространстве, имеющими длинами свои среднеквадратические отклонения. Тогда коэффициенты парной корреляции – это косинусы углов между векторами, а вычисление частных, "очищенных", корреляций соответствует нахождению косинусов углов между составляющими соответствующих векторов, которые ортогональны линейному подпространству, натянутому на фиксируемые векторы. Наличие упомянутых связей в этом контексте интерпретируем как то, что все углы между найденными компонентами могут быть вычислены, если мы знали все углы в первоначальной конструкции.

Нам понадобится т.н. остаток ξ_1 после фиксации остальных величин:

$$x_{1(2,...,p)} \ = \ \xi_1 \ - \ \mathbf{M}(\xi_1 / \ \xi_2, ..., \xi_p).$$

Ясно, что $\mathbf{M}x_{1(2,...,p)} \ = \ 0$. Обозначим дисперсию этого остатка через

$$\sigma^2_{1(2,...,p)} \ = \ \mathbf{M}x^2_{1(2,...,p)}.$$

Для случая многомерного нормального распределения понятия остатка и ошибки регрессии (см. главу "Регрессионный анализ")

$$\xi_{1(2,...,p)} \ = \ \xi_1 \ - \ \sum_{i=2}^{p} \beta_i \xi_i$$

совпадают, а значит и мы далее не будем их различать.

Коэффициенты β_i регрессии находятся из условия

$$\frac{\partial}{\partial \beta_i} \left(\sum_{j=1}^{n} \left(\xi_1^j \ - \ \sum_{i=2}^{p} \beta_i \xi_i^j \right)^2 \right) \ = \ 0,$$

позволяющего наилучшим образом аппроксимировать ξ_1 через линейные комбинации остальных величин. Поэтому

$$\sum_{j=1}^{n} \xi_1^j \left(\xi_1^j - \sum_{i=2}^{p} \beta_i \xi_i^j \right) \ = \ \sum_{j=1}^{n} \xi_1^j x^j_{1(2...p)} \ = \ 0.$$

Отсюда следует. что вектор остатков $x_{1(2,\ldots,p)}$ ортогонален любому из векторов $O\vec{Q}_i,\ i \geq 2$. Нетрудно понять, что на $R_{12(3\ldots p)}$ можно смотреть, как на косинус угла между $x_{1(3,\ldots,p)}$ и $x_{2(3,\ldots,p)}$.

Перейдем теперь к характеристике связей между одномерной величиной и многомерным случайным вектором. Вернемся к трехмерному случаю, определим $\vec{\eta} = (\xi_2, \xi_3)^t$ и будем считать, что $\vec{\eta}$ фиксирован.

Рассмотрим условную дисперсию

$$D(\vec{x}) = \mathbf{D}(\xi_1/\vec{\eta} = \vec{x})$$

и подставим в эту (неслучайную) функцию в качестве аргумента $\vec{\eta}$. Обозначим

$$\sigma^2_{1,(2,3)} = \mathbf{M}D(\vec{\eta}).$$

Эта величина описывает среднюю изменчивость ξ_1 при неизменном значении (ξ_2, ξ_3). Коэффициентом множественной корреляции между ξ_1 и ξ_2, ξ_3 называется число

$$R_{1,(2,3)} = \sqrt{1 - \frac{\sigma^2_{1,(23)}}{\mathbf{D}\xi_1}}.$$

Квадрат этой величины называют коэффициентом детерминации.

Действуя аналогичным образом, т.е. фиксируя некоторые подвекторы вектора произвольной размерности $\vec{\xi}$, рассчитывая математические ожидания оставшихся незафиксированными координат при выбранном условии (при этом в результате могут получаться векторные характеристики) и усредняя эти математические ожидания по фиксированным ранее значениям, мы можем получить коэффициенты множественной корреляции между любыми наборами координат.

Приведем формулы для вычисления множественного коэффициента корреляции через парные и частные коэффициенты корреляции координат нормального вектора. Пусть V, как и раньше, корреляционная матрица $\vec{\xi}$, $c_{i,j}$ - алгебраическое дополнение ее определителя к месту (i, j). Тогда множественный коэффициент корреляции между ξ_1 и $\xi_2, \ldots, \xi_n$ вычисляется так:

$$R^2_{1,(2,\ldots,n)} = 1 - \frac{|V|}{c_{1,1}}.$$

Можно также выписать его связь с частными коэффициентами корреляции $\rho(i,j)$:

$$R^2_{1,(2,\ldots,n)} \;=\; 1 \;-\; \prod_{j=2}^{n}(1-\rho^2_{1,j}).$$

Глава 3

Группировка и цензурирование

Задача формирования групп выборочных данных таким образом, чтобы сгруппированные данные могли предоставить практически тот же объем информации для принятия решения, что и выборка до группировки, решается исследователем в первую очередь. Целями группировки, как правило, служат снижение объемов информации, упрощение вычислений и придание наглядности данным. Некоторые статистические критерии изначально ориентированы на работу со сгруппированной выборкой. В определенных аспектах задача группировки очень близка задаче классификации, о которой подробнее речь пойдет ниже. Одновременно с задачей группировки исследователь решает и задачу цензурирования выборки, т.е. исключения из нее резко выпадающих данных, как правило, являющихся следствием грубых ошибок наблюдений. Естественно, желательно обеспечить отсутствие таких ошибок еще в процессе самих наблюдений, но сделать это удается не всегда. Простейшие методы решения упомянутых двух задач рассмотрены в этой главе.

3.1 Группировка в одномерном случае

В одномерном случае методы группировки (числовых) данных хорошо разработаны. Иногда группы уже подразумеваются в логике постановки задачи, иногда их, что называется, сразу видно. Напомним кратко, как поступать, если никаких априорных соображений не имеется.

Сначала определим число r групп-интервалов, которые мы хотим построить. Для начала можно воспользоваться эмпирической формулой

Стерджеса

$$r \;=\; [\log_2 n] + 1, \tag{3.1}$$

где n – объем выборки, [.] обозначена целая часть числа.

Далее определим максимальный $X_{(n)}$ и минимальный $X_{(1)}$ элементы выборки, ее размах $T \;=\; X_{(n)} - X_{(1)}$ и длину типичной группы-интервала $h \;=\; T/r$. Строим границы групп-интервалов $z_j,\ j = 0, ..., r$, полагая

$$z_0 \;=\; X_{(1)} - \varepsilon;\ \ z_j \;=\; z_{j-1} \;+\; h,\ j = 1, .., (r-1),\ \ z_r \;=\; X_{(n)} \;+ \varepsilon.$$

Здесь ε - произвольное малое число, не обязательно одинаковое в первой и последней формуле. В некоторых задачах $z_0,\ z_r$ принимаются равными $-\infty,\ \infty$ соответственно. После этого следует убедиться, что выполнено следующее условие:

Границы групп не должны совпадать с элементами выборки

Если это условие нарушается, то необходимо сдвинуть границы z_j на малое число ε вправо или влево.

Наконец, найдем количества n_j элементов выборки X, расположенных между z_{j-1} и $z_j,\ j = 1, ..., r$. Тут возникает самое главное условие

В каждом интервале должно быть от 3 до 19 элементов выборки

Если условие не выполнено, необходимо разбить "перенаселенные"интервалы на более мелкие и укрупнить "малообитаемые"интервалы за счет присоединения к соседним или просто за счет передвижения границ. Добившись выполнения последнего условия, полагаем $\Delta_j \;=\; (z_{j-1}, z_j),\ j = 1, ..., r$ (в роли z, r выступают их последние варианты, измененные в процессе борьбы за выполнение сформулированных условий), и группировка закончена.

В сгруппированной выборке принято все элементы X, попавшие в группу Δ_j, отождествлять с ее серединой и, тем самым, в нашей выборке появилось довольно большое число одинаковых элементов с повторностями n_j. За счет этого можно сэкономить на вычислениях, заменяя,

например, сложение одинаковых чисел умножением. Так,

$$\bar{X} \;=\; \frac{1}{n}\sum_{j=1}^{r} n_j \tilde{x}_j, \quad S^2 \;=\; \frac{1}{n}\sum_{j=1}^{r} n_j (\tilde{x}_j - \bar{X})^2,$$

где $\tilde{x}_j \;=\; (z_j \;+\; z_{j-1})/2$. Конечно, при этом совершаются определенные ошибки. Имеется специально разработанный аппарат поправок, называемых поправками Шеппарда, применимый в случае, когда все длины интервалов-групп одинаковы.

Отметим в заключение раздела, что группировка могла осуществляться произвольным способом, и только взятые в рамки условия должны быть выполненными всегда.

3.2 Одномерное цензурирование

В процессе осуществления группировки, описанной в предыдущем разделе, может возникнуть ситуация, когда одно из выборочных значений оказывается в своем интервале единственным и слишком удаленным от остальных (например, имеются "пустые"интервалы, отделяющие его от остальных значений). Это указывает на то, что "выпадающее"значение, скорее всего, является результатом грубой ошибки наблюдения и должно быть отброшено. Естественно, описанный метод является очень грубым. Ниже приведено несколько критериев, позволяющих производить более обоснованное цензурирование.

1. Вычислим среднее всей выборки $\bar{X}$ и среднее по выборке, из которой удален "подозрительный"элемент. Обозначим это среднее через $\bar{X}'$. Если сравнение двух этих средних, например, по критерию Стьюдента, не дает существенных различий, то изучаемое значение отбрасывать не следует.

2. Пусть x – "подозрительное"значение. Вычислим величину

$$t \;=\; \frac{|x \;-\; \bar{X}|}{S},$$

где S - выборочное среднеквадратическое отклонение, и на уровне доверия $1 - \alpha$ сравним с критическим значением

$$w_{n,\alpha} \;\approx\; u \cdot \sqrt{\frac{2(n-1)}{2n - 5 + u^2 + \frac{3+u^2+2u^4}{12n-30}}},$$

где u – квантиль стандартного нормального распределения уровня $1 - \frac{\alpha}{2n}$. Таблицы $w_{n,\alpha}$ можно найти, например, в [1, с. 199]. Если при этом $t \leq w_{n,\alpha}$, то отбрасывать подозрительное значение нет оснований.

3. Правило Томпсона. Это правило, в отличие от предыдущих, позволяет исключать сразу несколько значений. Рассчитаем

$$t_i = \frac{x_i - \bar{X}}{S},$$

и если

$$|t_i| > z_{n-2,\alpha} = \frac{t_{n-2,1-\alpha/2}\sqrt{n-1}}{\sqrt{n-2+t^2_{n-2,1-\alpha/2}}},$$

где $t_{n-2,1-\alpha/2}$ – квантиль распределения Стьюдента с $n-2$ степенями свободы, то x_i нужно отбросить с вероятностью $1-\alpha$. Таблицы $z_{n-2,\alpha}$ также можно найти в [1].

4. Можно также упорядочить выборку по возрастанию и воспользоваться одной из следующих статистик:

$$T_1 = \frac{X_{(n)} - X_{(n-1)}}{X_{(n)} - X_{(1)}},\ T_2 = \frac{X_{(n)} - X_{(1)}}{X_{(n)} - X_{(2)}},\ T_3 = \frac{X_{(n)} - X_{(n-2)}}{X_{(n)} - X_{(1)}}$$

в зависимости от того, хотим мы исключить наибольшее, наименьшее или два самых больших значения. Таблицы соответствующих критических значений можно найти в той же книге [1, с.201-202].

Отметим, что второй и третий алгоритмы применимы лишь в случае выборки из нормального распределения, первый алгоритм в случае отсутствия нормальности применим при больших объемах выборки, а последний устойчив к виду распределения и применим всегда.

3.3 Таблицы сопряженности

Общая задача группировки многомерных данных очень сложна, и мы будем далее неоднократно возвращаться к ней, исследуя ее с разных точек зрения и используя различные подходы. С простейшими алгоритмами объединения многомерных данных в группы мы сталкиваемся в задачах

регрессии и при проверке гипотез независимости и однородности для (одномерных) случайных величин. При этом возникают так называемые таблицы сопряженности. Опишем соответствующую процедуру.

Пусть результаты m наблюдений над случайным вектором $\vec{\xi}$ с координатами $\xi_1, ..., \xi_n$ собраны в матрице X размером $m \times n$ (значения координат наблюдаемого вектора заполняют строку этой матрицы). Будем смотреть на матрицу X как на набор n выборок объемом m каждая. Проведем группировку каждой из этих выборок, сообразуясь с алгоритмом, изложенным в разделе 3.1 . Отбрасывание "выпадающих"значений при этом не производится. Получим, что j-я выборка (столбец таблицы) разбита на r_j групп, $\Delta_i^j,\ i = 1, ..., r_j,\ j = 1, ..., n$. Определим теперь n-мерные группы в количестве $r_1 \cdot r_2 \cdot ... \cdot r_n$ следующим образом: значение $(x_{k,1}, ..., x_{k,n})$ попадают в группу, занумерованную цепочкой $i_1, ..., i_n$, если $x_{k,1} \in \Delta_{i_1}^1,\ x_{k,2} \in \Delta_{i_2}^2,\ ...\ x_{k,n} \in \Delta_{i_n}^n$. Проще говоря, группы представляют собой n-мерные прямоугольные параллелепипеды, и выборочное n-мерное значение попадает в соответствующую группу тогда, когда все координаты этого значения попадают в проекции параллелепипеда на оси координат.

Наконец, подсчитаем количество элементов выборки из распределения ξ, попавших в каждую из полученных групп и обозначим эти количества через $\nu_{i_1,...,i_n}$. Никаких изменений полученных групп в зависимости от полученных чисел в отличие от одномерного случая не производится. В дальнейшем условимся считать все $\nu_{i_1,...,i_n}$ значений наблюдаемого вектора, попавшие в эту группу, равными радиус-вектору ее геометрического центра. Группировка закончена.

По ее результатам заполняется n-входовая таблица. На ее месте, занумерованном цепочкой $i_1, ..., i_n$ стоит число $\nu_{i_1,...,i_n}$. Эта таблица и называется таблицей сопряженности.

Понятно, что группы-параллелепипеды удобны не во всех задачах, но могут успешно применяться в упомянутых выше. Коротко остановимся на задачах независимости и однородности (в одномерном случае).

3.3.1 Гипотеза независимости

Рассмотрим выборки из распределений двух случайных величин X, Y одного и того же объема n. Объединим их в двумерную выборку и произведем построение таблицы сопряженности как это было описано выше. В нашем случае это будет таблица $r_1 \times r_2$, в которой собраны числа

$\nu_{i,j},\ i = 1, ..., r_1,\ j = 1, ..., r_2$, где r_1, r_2 – количества интервалов, полученных при группировке X, Y соответственно. Пусть

$$n_i = \sum_{j=1}^{r_2} \nu_{i,j}, \quad m_j = \sum_{i=1}^{r_1} \nu_{i,j}.$$

Числа n_i – количества элементов X, попавших в i-й интервал при своей группировке, числа m_j имеют тот же смысл для Y. Мера отклонения наблюдаемой картины от предполагаемой при наличии независимости описывается статистикой

$$\chi^2 = n \left(\sum_{i=1}^{r_1} \sum_{j=1}^{r_2} \frac{\nu_{i,j}^2}{n_i m_j} - 1 \right).$$

После расчета значение χ^2 сравнивается с критической точкой распределения хи-квадрат с $(r_1 - 1)(r_2 - 1)$ степенями свободы. Если критическое значение не превзойдено - гипотезу о независимости можно принять.

Напомним, что понятие независимости в ряде практических случаев близко понятию некоррелированности. По крайней мере, если коэффициент корреляции существенно отличен от 0, то наблюдаемые величины, конечно же, зависимы. Сформулируем здесь соответствующий критерий. Вычислим выборочный коэффициент корреляции способом подстановки:

$$R = n \frac{\bar{XY} - \bar{X} \cdot \bar{Y}}{\sqrt{\sum_{i=1}^{n} (x_i - \bar{X})^2 \sum_{j=1}^{n} (y_j - \bar{Y})^2}}.$$

Здесь $\bar{XY}$ – среднее попарных произведений элементов выборок.

Положим

$$T = \frac{|R|\sqrt{n-2}}{\sqrt{1 - R^2}}$$

и сравним это число с двусторонней критической точкой распределения Стьюдента с $n-2$ степенями свободы. В случае, когда T оказалось больше критической точки, гипотезу о нулевом коэффициенте корреляции следует отвергнуть.

3.3.2 Гипотеза однородности

Здесь рассматриваются m выборок объемов $n_1, ..., n_m$ из распределений $\xi_1, ...\xi_m$. Высказываемая гипотеза состоит в том, что все случайные величины имеют одно и то же распределение. Введем вспомогательную

случайную величину η, связанную с рассмотренными ранее следующим образом. Рассмотрим объединенную выборку объема $n_1 + ... + n_m$ и выберем из нее произвольный элемент. Будем считать, что $\eta = j$ на этом элементе, если до объединения выборок он принадлежал j-й выборке, т.е. являлся значением ξ_j. Итак, мы рассматриваем двумерный случайный вектор (ζ, η), где вторая координата принимает натуральные значения $1, ..., m$. Гипотеза однородности заменяется далее на гипотезу независимости (ζ, η), которая проверяется изложенными выше методами. При этом производить группировку значений η нет необходимости, т.к. в качестве групп можно рассмотреть все ее значения.

Идея, используемая здесь, состоит в том, что если выборки действительно однородны, то, перемешав их, мы не сможем определить, из какой выборки было взято некоторое конкретное значение, т.е. номер выборки не зависит от величины ее элемента.

3.3.3 Поле корреляции

Наглядным вариантом таблиц сопряженности служат поля корреляции – рисунки, на которых значения наблюдаемых случайных векторов изображены точками в соответствующем масштабе. Конечно же, имеет смысл изображать поле корреляции только в случае, когда наблюдаемый вектор имеет размерность не более 3.

После того, как поле корреляции изображено, исследователь имеет возможность путем визуального осмотра установить наличие зависимости координат и, в случае ее присутствия, характер этой зависимости. В случае двумерного вектора зависимость можно определить по тенденции точек, изображающих выборочные значения, группироваться к некоторой кривой, которая и описывает характер зависимости координат. Отсутствие зависимости приводит к хаотичному расположению точек.

Полезным является также построение так называемых эмпирических линий регрессии - ломаных с узлами в геометрических центрах множеств точек, попавших в определенные "коридоры"которые, как правило, определяются границами групп, построенных по одной из координат. В случае двумерного вектора таких линий две: для построения первой рассматриваются множества точек, для которых первая координата попала в первый, второй, и т.д. интервал по x, а для второй - вторая координата последовательно фиксируется внутри первого, второго и т.д. интервалов по y.

3.4 Многомерная группировка данных - общие принципы

Как уже было упомянуто, общая задача группировки в многомерном случае не имеет однозначного решения. Для продвижения в этой задаче необходимо прежде всего задаться критерием, согласно которому следует признать близкими два выборочных многомерных значения. Разные критерии (или, точнее, различным образом задаваемые расстояния между точками) приводят к различным группировкам. Опять же, возможны разные постановки задачи. Например, если у нас уже имеется по крайней мере по нескольку выборочных точек, представляющих каждую из образуемых групп (классов, таксонов, кластеров), то мы имеем дело с задачей классификации при наличии обучающей выборки, которая решается методами дискриминантного анализа. Этим методам мы посвятим отдельную главу. Не следует также пренебрегать возможностью построения групп в форме параллелепипедов, как это было описано выше.

Если же заранее нет никаких соображений, какими должны быть группы и сколько их должно быть, то сначала следует выбрать расстояние, т.е. ту характеристику, по которой мы оцениваем близость точек в выборочном пространстве. После такого выбора мы получаем задачу, которую принято называть задачей кластерного анализа (или задачей классификации без обучения). Существует много самых разных алгоритмов решения этих задач, см. [2]. Мы рассмотрим только алгоритм, основанный на близости элементов многомерной выборки в смысле обычного (евклидова) расстояния.

Алгоритм построен на повторении определенных операций до наступления "стабилизации"наблюдаемой картины, т.е. имеет итеративный характер и известен под названием "ФОРЕЛЬ". Он строит группы, которые имеют форму одинаковых шаров.

Сначала выберем некоторое (небольшое) положительное число a, которое будет представлять собой радиус строящихся групп-шаров.

Шаг 0. Построим некоторое первоначальное разбиение всего множества выборочных точек на две группы. Вообще-то это разбиение можно выбирать произвольным образом, но для увеличения скорости сходимости алгоритма рекомендуется в качестве одной из групп взять одну или несколько близких точек, лежащих в окрестности геометрического центра облака выборочных точек, а все остальные точки пока отнести в

другую группу.

Шаг 1. Рассчитаем координаты геометрического центра точек, отнесенных в строящуюся группу как среднее арифметическое координат этих точек и отнесем в эту группу все те точки, которые расположены к определенному геометрическому центру ближе, чем на a.

Шаг 2. Если на предыдущем шаге строящаяся группа не изменилась, переходим к шагу 3, иначе к шагу 1.

Шаг 3. Очередная группа сформирована. Убираем из облака выборочных точек точки, составляющие эту группу. Если больше точек не осталось, то сформированы все группы, иначе переходим к шагу 0 с уменьшенным облаком точек.

Известно, что этот алгоритм сходится за конечное число шагов. Моментом, на которое следует обратить внимание, является правильный выбор радиуса шаров a. Для того, чтобы определить оптимальный радиус, можно прибегнуть к следующей процедуре: если после работы алгоритма "ФОРЕЛЬ"получилось приемлемое число групп (например, больше одной, но меньше числа всех точек), начнем уменьшать число a с определенным шагом и повторять работу алгоритма. Если некоторое время число групп не меняется, а с очередным уменьшением радиуса резко возрастает, то последнее (до уменьшения) значение a было оптимальным.

3.5 Многомерное цензурирование

Задача отбрасывания "посторонних"данных, являющихся грубыми ошибками наблюдения, в многомерном случае еще более сложна, чем задача группировки. В общей ситуации можно предложить простейший эмпирический метод, который, как и в одномерном случае, состоит в нижеследующем. Произведем группировку многомерных данных например, с помощью алгоритма "ФОРЕЛЬ", и если какое-то из выборочных значений оказалось в своей группе в единственном числе, причем его группа не примыкает достаточно близко ни к одной из остальных групп, то такое значение следует отбросить. Понятие "достаточно близко"здесь не формализуется, но понимать его можно как "сравнимо с радиусом групп".

Немного более точные рекомендации можно дать в случае наблюдения вектора, имеющего многомерное нормальное распределение. Как следует из доказанной выше теоремы 3 , основное количество значений многомерного нормального распределения со средним $\vec{a}$ и ковариацион-

ной матрицей V лежит внутри т.н. эллипсоида рассеивания, т.е. удовлетворяет неравенству

$$V^{-1}(\vec{x}-\vec{a})\cdot(\vec{x}-\vec{a}) \;\leq\; t_{1-\alpha}, \tag{3.2}$$

где $t_{1-\alpha}$ – квантиль распределения хи-квадрат с n степенями свободы уровня $1-\alpha$ для достаточно малого α. Поэтому для цензурирования многомерной нормальной выборки можно методом подстановки оценить $\vec{a}$, V и проверить выполнение (3.2) для каждого из выборочных значений. Те значения, для которых неравенство нарушается, должны быть отброшены.

Можно предложить и другие методы, например, являющиеся распространением методов раздела 3.2 на многомерный случай, но они также не будут давать более надежных результатов. Окончательный вывод: при сборе многомерных данных следует проявлять особую внимательность с целью избегнуть грубых ошибок уже на этом этапе.

Глава 4

Нечисловые данные и экспертные оценки

4.1 Вводные замечания

При проведении наблюдений чаще всего фиксируются числовые данные, т.е. численно измеряются некоторые характеристики наблюдаемой величины. Нам привычнее иметь дело именно с числами, а не какими-то описательными терминами, и числовые характеристики всегда несут в себе большую информацию. Например, мало сказать "автомобиль двигается с большой скоростью", надо выразиться определеннее – "скорость автомобиля 120 км/ч". Первый способ выражения имеет скорее эмоциональную окраску, при втором способе уже можно определенным образом обрабатывать сообщенную информацию и делать достаточно достоверные заключения, например, о длине тормозного пути.

Именно поэтому, а также в силу того, что разработанный аппарат математической статистики ориентирован на работу с числами, данные обычно пытаются представить именно в числовом виде, а если он первоначально не числовой, то строят специальные процедуры для перевода данных именно в эту форму.

Заметим, что абсолютно нечисловой формы данных не бывает. В любом случае имеются некоторые категории наблюдаемых величин, т.е. небольшое число классов (категорий), в один из которых можно отнести очередное наблюдение. Например, при наблюдении за характерами людей, можно предложить такие категории, как "спокойный", "неров-

ный", "вспыльчивый", "бешеный", при необходимости увеличивая этот список, затем приписать каждой из введенных категорий числовую метку (например, номер категории) и работать далее с числовыми данными.

Даже если разбиение на категории характеров, описанное в предыдущем абзаце вызывает затруднение исследователя, на множестве наблюдаемых объектов можно ввести некое отношение типа "лучше - хуже"и произвести сравнение, например: "У Кондрата характер хуже, чем у Пафнутия". После этого выделяются определенные характерные объекты, объявляющиеся границами категорий, и все остальные объекты размещаются между ними, согласуясь с выбранным отношением. Если разбиение на категории произведено, то условимся говорить, что мы имеем дело с категоризованными данными.

Нетрудно представить себе ситуацию, когда данные имеют смешанный характер. Представим себе больных, попадающих в травматологическое отделение больницы. Среди них есть больные с травмами определенных видов – рук, ног, позвоночника, черепной коробки и все возможные варианты их сочетаний (всего 15 вариантов). Каждый больной проводит в палате определенное время и при выходе из больницы получает инвалидность определенной группы (если наступило полное выздоровление, дадим ему нулевую группу). Итак, с каждым из больных мы связали три наблюдаемых переменных - вид травмы (15 категорий), время в больнице (числовая) и инвалидность по выписке (3 категории, возможно 4, если больной не выписывается). Первая и третья переменная – нечисловые.

Как мы видим, в работе с нечисловыми данными всегда присутствует некий субъективный элемент. Тем не менее, можно предложить и некие объективные процедуры, описанием которых мы дальше и займемся.

4.2 Шкалы сравнений

Одной из основных областей, где возникают нечисловые данные, является задача сравнения нескольких объектов между собой, причем исследователь (будем называть его "экспертом", поскольку в таких экспериментах чаще всего опрашивают сведущих людей), как правило, занимается тем, что оценивает "похожесть"предложенных объектов попарно. В этом случае эксперт заполняет данные по степени похожести, например объекта А на объект Б и объекта Б на объект В. То, что заполняется, называют шкалой сравнений.

Шкалы сравнений бывают двух видов. При заполнении графической шкалы эксперт должен поставить отметку тем правее, чем больше, по его мнению, различны рассматриваемые объекты:

Графическая шкала сравнения двух объектов

Похожи ——————x—————— Различны

Другой вид шкалы имеет вид заранее определенных категорий – степень похожести объектов изначально задается некоторым числом, имеющим, правда, не абсолютную, а относительную величину:

Категоризованная шкала сравнения двух объектов

Похожи Различны

1	2	3	4	5 x	6	7	8	9	10	11	12

Для обработки предпочтительнее второй вид шкал, т.к. для приведения графической шкалы в числовую форму необходимо производить измерение величины отрезка, отмеченного экспертам, что приводит к дополнительным затратам времени и является источником ошибок. С другой стороны, графический способ дает эксперту возможность более свободно выражать свое мнение относительно похожести объектов.

Естественно, что сколько экспертов, столько и мнений, поэтому для повышения надежности оценки степени сходства/различия нужно опросить как можно больше экспертов и в качестве окончательной оценки выбрать, например, среднюю арифметическую полученных оценок. Стоит также иметь ввиду, что иногда даже один эксперт по-разному оценивает степень похожести, если объекты на шкале расположены в другом порядке. Поэтому следует с некоторым интервалом попросить его оценить сначала сходство объекта А с объектом Б, а затем объекта Б с объектом А. В качестве оценки похожести этих объектов выбирают полусумму оценок.

Шкалы сравнений занимают промежуточное место между числовыми и нечисловыми данными и на их примере отчетливо видны проблемы, сопутствующие переходам между этими видами данных. Обычная сфера

применения шкал сравнений - многомерное шкалирование, которому мы посвятим отдельную главу.

Слегка затронем также вопрос о придании меток категориям, выделяемым при изучении нечисловых данных. Иногда из априорных соображений бывает ясно, что те или иные категоризованные переменные должны быть сильно связаны (или наоборот, практически независимы) между собой или с какими-то из числовых переменных. В этом случае задача выбора числовых меток ставится из условия максимизации (минимизации) соответствующих коэффициентов корреляции. Подробнее см. в главе, посвященной оцифровке нечисловых данных.

4.3 Экспертные оценки

Рассмотрим теперь вариант задачи об оценивании степени сходства. Пусть у нас имеется некоторое количество факторов (иногда называемых стимулами), которые предложено расположить в порядке убывания значимости или привлекательности. Располагая факторы в требуемом порядке, эксперт тем самым указывает для каждого из них порядковый номер, который мы будем называть рангом фактора, который ему присвоил эксперт. Так, самому важному фактору эксперт присваивает номер 1, а наименее важному – номер n, если всего предлагалось для ранжирования n факторов.

Особым образом оговаривается случай, когда эксперт не может или не хочет различать некоторые из предложенных факторов. В этом случае ему предлагается каждому из группы неотличимых с его точки зрения факторов присвоить ранг, равный среднему арифметическому тех рангов, которые должны были быть распределены среди факторов этой группы. Делается это для того, чтобы сумма всех рангов, присвоенных экспертом, была бы одной и той же независимо от его предпочтений.

После опроса нескольких экспертов мы изучим согласованность их мнений по данному вопросу, а также постараемся выделить группы экспертов с наиболее согласованными мнениями. А пока рассмотрим пример, заимствованный из [3].

Пяти студентам, обычно добирающимися до места учебы общественным транспортом, предложили ранжировать в порядке убывания важности следующие факторы: Ч – частота следования транспорта, О – оборудование салона (мягкие сиденья, кондиционер, музыка), Д – исправность

дверей и окон, К – настроение и доброжелательность кондуктора, С – освещение в салоне, Ц – стоимость проезда. Данные соответствующего опроса приведены в таблице:

Пять экспертов и семь факторов

Факторы \ Эксперты	1	2	3	4	5
Ч	1	2	2	1	3
З	3	1	2	5	2
О	5	3	6,5	5	5
Д	7	4	6,5	5	4
К	6	7	5	5	7
С	4	6	2	5	6
Ц	2	5	4	2	1

Как видим, первый, второй и пятый студенты полностью справились с задачей и каждому фактору присвоили свой ранг. Третий студент считает тремя самыми важными факторами частоту, заполненность и стоимость проезда, но внутри этой группы ранжирование произвести не захотел. Он же считает наименее важными, но равнозначными факторами оборудование салона и состояние дверей и окон. С точки зрения четвертого студента, который подошел к задаче наиболее безответственным образом, ни один из факторов, кроме Ч и З, не имеет особого значения, а стоимость проезда важнее, чем частота движения транспорта. Заметим, что уже по этим данным можно сделать определенные выводы о характерах этих студентов и поведении их на городских улицах.

4.4 Согласованные группы и коэффициенты ранговой корреляции

Для оценки степени согласованности мнений двух экспертов в задаче ранжирования некоторого количества факторов, поставленной в предыдущем разделе, используется коэффициент ранговой корреляции Спирмена, который вычисляется по формуле

$$R(X,Y) \;=\; 1 \;-\; \frac{6\sum_{i=1}^{n} d_i^2}{n^3 \;-\; n \;-\; (T_X \;+\; T_Y)/2}. \qquad (4.1)$$

Здесь:

- X, Y – номера или обозначения экспертов, согласованность мнений которых изучается;
- d_i – разность рангов, которые присвоили i-му факторы эксперты X и Y. Поскольку в формуле она возводится в квадрат, то безразлично, из чего что вычитать;
- n – число предлагаемых к ранжированию факторов;
- число T_X (T_Y) вычисляется для эксперта X (Y) следующим образом. Если некоторые из рангов, присвоенные экспертом, оказались равными, то будем говорить, что эти ранги образуют группу связанных рангов. Если у эксперта X нет групп связанных рангов, то для него $T_X = 0$. В противном случае полагаем

$$T_X = \sum_j (t_j^3 - t_j),$$

где сумма берется по всем группам связанных рангов у эксперта X, а t_j – количество факторов, образующих j-ю группу связанных рангов.

Коэффициент Спирмена с качественной точки зрения может так же характеризовать степень согласованности экспертов, как коэффициент корреляции характеризует степень связи случайных величин, см. таблицу в конце раздела 2.3 .

Рассмотрим пример со студентами и общественным транспортом. У первого, второго и пятого эксперта нет связанных групп рангов, поэтому $T_1 = T_2 = T_5 = 0$. У третьего есть две группы, состоящие из $t_1 = 2$ фактора, имеющих ранги по 6,5 и $t_2 = 3$ факторов, имеющих ранги по 2. Следовательно, $T_3 = (2^3 - 2) + (3^3 - 3) = 30$. У четвертого тоже одна группа связанных факторов, но их 5, значит $T_4 = 120$. Рассчитаем теперь, например, коэффициент ранговой корреляции между первым и вторым экспертами по формуле (4.1).

$$\sum_{i=1}^{7} d_i^2 = (1-2)^2 + (3-1)^2 + \ldots + (2-5)^2 = 29;$$

$$R(1,2) = 1 - \frac{6 \times 29}{7^3 - 7 - (0+0)/2} \approx 0,41.$$

Остальные коэффициенты Спирмена вычисляются аналогично. Заполним таблицу.

Коэффициенты Спирмена в задаче о транспорте

Эксперты	1	2	3	4	5
1	1,00	0,41	0,75	0,78	0,64
2	—	1,00	0,24	0,22	0,61
3	—	—	1,00	0,40	0,43
4	—	—	—	1,00	0,56
5	—	—	—	—	1,00

Нижняя часть таблицы не заполнена, т.к. $R(i,j) = R(j,i)$ для произвольной пары i, j.

Теперь займемся формированием групп экспертов с наиболее согласованными мнениями. Заметим, что в нашей таблице нет отрицательных коэффициентов, что означает, что среди экспертов нет противоположных (конфликтующих) мнений. Максимальное значение коэффициента – $R(1,4)$. Именно этих двух экспертов мы и объединим в группу. К ним мог быть подключен еще и третий эксперт, но у него с четвертым один из самых низких коэффициентов корреляции. После удаления соответствующих столбцов и строчек из таблицы, максимальная корреляция среди оставшихся наблюдается между вторым и пятым экспертами. Из них сформируем вторую группу. Третий эксперт, оставшись сам по себе, образует третью группу.

Следует подчеркнуть, что процесс формирования групп носит во многом субъективный характер и слабо формализуется. Ниже будут высказаны соображения объективного характера по поводу формирования групп с согласованными мнениями, но они связаны с большим объемом вычислений.

Следующий шаг - формирование групповых рангов, т.е. ранжирование изучаемых факторов по мнению группы в целом. Для их определения заполним две таблицы. Сначала определим коллективные ранги, которые для каждого фактора получаются суммированием рангов, присвоенных экспертами группы данному фактору. Затем, упорядочив коллективные ранги в порядке возрастания и проранжировав их в соответствии с алгоритмом предыдущего раздела, заполним вторую, итоговую таблицу.

Коллективные ранги $f^{(g)}$

группы \ факторы	Ч	З	О	Д	К	С	Ц
I	2	8	10	12	11	9	4
II	5	3	8	8	14	12	6
III	2	2	6,5	6,5	5	2	4

Групповые ранги

группы \ факторы	Ч	З	О	Д	К	С	Ц
I	1	3	5	7	6	4	2
II	2	1	4,5	4,5	7	6	3
III	2	2	6,5	6,5	5	2	4

В дальнейшем по той же формуле (4.1) можно вычислить коэффициенты ранговой корреляции между группами.

Для оценки степени согласования экспертов внутри группы рассчитывается так называемый коэффициент конкордации. Сначала для группы вычислим среднее L_g коллективных рангов (усреднение по факторам) и через S_g^2 обозначим среднее квадратичное отклонение коллективных рангов:

$$S_g^2 \;=\; \frac{1}{m_g}\sum_{i=1}^{n}(f_i^{(g)} - L_g)^2,$$

где m_g – число экспертов в группе, $f_i^{(g)}, i = 1, ..., n$ - коллективные ранги группы для каждого из факторов. Пусть T_g - сумма чисел T_i, определенных после формулы (4.1) по всем экспертам, входящим в группу. Тогда коэффициент конкордации W_g для группы определяется формулой

$$W_g \;=\; \frac{12S_g^2}{m_g(n^3 - n) \;-\; T_g}. \tag{4.2}$$

Величина коэффициента конкордации может быть только неотрицательной и не большей единицы. Увеличение его означает наличие большего согласования мнения экспертов в группе.

На основе введенного коэффициента можно построить тот алгоритм оптимального разбиения на группы, о котором было сказано выше. Если число групп заранее задано, рассматриваем все возможные разбиения на такое число групп и в каждой группе считаем коэффициент конкордации. Критерием оптимальности может служить, например, сумма чисел W_g по очередному разбиению. Оптимальным будет то, для которого эта

сумма максимальна. Впрочем, в зависимости от задачи, можно предложить и другие критерии.

Если же неизвестно и число групп, то следует рассмотреть все возможные разбиения на группы, включая выделение каждого из экспертов в отдельную группу и полный отказ от разбиения. Алгоритм остается тем же, только еще более возрастает количество требуемых вычислений.

Вернемся к нашему примеру. Расчеты показывают, что средние коллективные ранги по группам равны $L_I = L_{II} = 8$, $L_{III} = 4$. Средние квадратичные отклонения для групп $S_I^2 = 41$, $S_{II}^2 = 45$, $S_{III}^2 = 25,5$. Наконец, $T_I = T_1 + T_4 = 120$, $T_{II} = T_2 + T_5 = 0$, $T_{III} = 30$. Окончательно, $W_I = W_{II} = 0,80$, $W_{III} = 1$. Последний результат предугадать было нетрудно – мнение каждого эксперта максимально согласовано само с собой.

Глава 5

Доверительные и толерантные множества

5.1 Доверительные интервалы

Одна из задач, часто решаемых статистиками – задача оценивания неизвестного параметра распределения. Если коротко напомнить ее суть, то она состоит в следующем. Исследователю нет необходимости знать всю функцию распределения в целом, а достаточно знания некоторой (или некоторых) числовых характеристик, оцениванием которых по выборке он и занимается. Часто бывает известно, что наблюдаемая случайная величина имеет одно из распределений фиксированного априори параметрического семейства – например, нормально распределена, и тогда, оценивая неизвестные параметры, мы решаем все наши проблемы.

При попытке заменить неизвестный параметр значением какой-то статистики вероятность угадать истинное значение параметра за исключением небольшого количества вырожденных случаев оказывается равной нулю. Поэтому, чтобы повысить надежность решений, принято в ряде случаев не ограничиваться указанием одного числа, а указывать границы, в которые оцениваемый параметр попадет с достаточно большой вероятностью. Интервал, помещающийся в указанных границах, носит название доверительного.

Приведем точные формулировки. Пусть наблюдаемая случайная величина ξ имеет одно из распределений $\mathbf{P}_\theta$, $\theta \in \Theta$. Интервал (θ^-, θ^+), границы которого являются статистиками, т.е. зависят от выборки, на-

зывается доверительным интервалом для θ уровня доверия $1-\alpha$, если

$$\mathbf{P}_\theta(\theta^- \ < \ \theta \ < \ \theta^+) \geq 1-\alpha.$$

Иногда, когда в последней формуле стоит знак равенства, говорят о точном доверительном интервале. Если требуемое неравенство становится справедливым только при достаточно больших объемах выборки, то интервал (θ^-, θ^+) называют асимптотическим.

Широко известны формулы для построения доверительных интервалов для параметров одномерной нормальной совокупности:

1. если дисперсия σ^2 известна, то границы доверительного интервала для математического ожидания имеют вид

$$a^\pm \ = \ \bar{X} \ \pm \ \frac{\sigma t_{1-\alpha/2}}{\sqrt{n}},$$

где $t_{1-\alpha/2}$ - квантиль стандартного нормального распределения соответствующего уровня;

2. если дисперсия неизвестна, то формулы приобретают вид

$$a^\pm \ = \ \bar{X} \ \pm \ \frac{S\tau_{1-\alpha/2}}{\sqrt{n-1}},$$

где $\tau_{1-\alpha/2}$ квантиль распределения Стьюдента с $n-1$ степенью свободы, а S – выборочное среднее квадратическое отклонение;

3. если математическое ожидание a известно, то границы доверительного интервала для дисперсии записываются, как

$$(\sigma^2)^\pm \ = \ \frac{nQ^2}{t_\alpha^\mp},$$

где t_α^+ – квантиль распределения хи-квадрат с n степенями свободы уровня $1 \ - \ \alpha/2$, t_α^- – квантиль уровня $\alpha/2$ того же распределения, а

$$Q^2 \ = \ \frac{1}{n}\sum_{i=1}^{n}(x_i \ - \ a)^2.$$

4. если математическое ожидание неизвестно, то

$$(\sigma^2)^{\pm} \;=\; \frac{nS^2}{\tau_\alpha^{\mp}},$$

где, на этот раз, $\tau_\alpha^{\pm}$ – квантили распределения хи-квадрат с n степенями свободы.

Описанное выше построение точных доверительных интервалов основано на следующей теореме Фишера, доказательство которой дается в курсе математической статистики.

Теорема 4 (Фишер) *Пусть X – выборка из нормального распределения с параметрами a, σ^2 объема n. Тогда*

1. $\frac{\bar{X}-a}{\sigma}\sqrt{n}$ *имеет стандартное нормальное распределение;*
2. nS_a^2/σ^2, *где* $S_a^2 \;=\; \frac{1}{n}\sum_{i=1}^n (x_i-a)^2$ *имеет распределение хи-квадрат с n степенями свободы;*
3. nS^2/σ^2 *имеет распределение хи-квадрат с $n-1$ степенями свободы и не зависит от $\bar{X}$;*
4. $\frac{\bar{X}-a}{S}\sqrt{n-1}$ *имеет распределение Стьюдента с $n-1$ степенью свободы.*

В случае, когда распределение не является нормальным, первые две формулы используют для построения асимптотических доверительных интервалов для математического ожидания. Построение же доверительных интервалов для других параметров представляет из себя как правило, достаточно сложную задачу. Иногда ее удается решить, хотя бы в асимптотическом смысле, найдя для неизвестного параметра оценку, имеющую в пределе нормальное распределение.

Теорема 5 *Пусть θ^* - асимптотически нормальная оценка неизвестного параметра θ, имеющая асимптотически нормальное распределение с известным коэффициентом рассеивания σ^2, т.е.*

$$\mathbf{P}((\theta^*-\theta)\sqrt{n} \;<\; x) \;\longrightarrow\; \Phi\left(\frac{x}{\sigma}\right)$$

при произвольном x, тогда доверительный интервал для θ может быть в асимптотическом смысле построен по формулам

$$\left(\theta^* - \frac{\sigma t_\alpha}{\sqrt{n}};\ \theta^* + \frac{\sigma t_\alpha}{\sqrt{n}}\right),$$

где t_α - соответствующая квантиль стандартного нормального распределения.

5.2 Доверительные множества

Возможны разные обобщения понятия доверительного интервала на случай многомерного параметра. В общем случае, если параметр $\vec{\theta}$ имеет размерность m, то при фиксированном малом числе α подмножество A_α m-мерного пространства называют доверительным уровня $1 - \alpha$, если

$$\mathbf{P}(\vec{\theta} \in A_\alpha) \geq 1 - \alpha.$$

Задача построения доверительных множеств однозначно не решается и еще более сложна, чем задача построения доверительных интервалов. Но в случае нормальной совокупности нам известны совместные распределения выборочных средних и дисперсий, поэтому кое-что все же построить можно.

Далее разобраны отдельно случаи многомерного параметра для одномерного распределения и многомерной выборки.

5.2.1 Многомерный параметр

В случае выборки из (одномерного) нормального распределения с параметрами a, σ^2 нам известно, что случайная величина $\eta = \frac{\bar{X} - a}{\sigma}\sqrt{n}$ имеет стандартное нормальное распределение, а случайная величина $\chi = nS^2/\sigma^2$ – распределение хи-квадрат с $n - 1$ степенью свободы, причем эти величины независимы. Учтем теперь, что одновременное выполнение неравенств

$$c_1 < \frac{nS^2}{\sigma^2} < c_2, \quad b_1 < \frac{\bar{X} - a}{\sigma}\sqrt{n} < b_2$$

для произвольных чисел b_1, b_2, c_1, c_2 влечет выполнение неравенств

$$\frac{nS^2}{c_2} < \sigma^2 < \frac{nS^2}{c_1},\ \bar{X} - \frac{Sb_2}{\sqrt{c_1}} < a < \bar{X} - \frac{Sb_1}{\sqrt{c_2}},$$

а значит последние неравенства имеют, вообще говоря, большую, чем первые совместную вероятность. Зафиксируем малые числа ϵ, δ и положим $c_1 = \tau_{\epsilon/2},\ c_2 = \tau_{1-\epsilon/2}$ – квантили хи-квадрат распределения с $n-1$ степенями свободы, $b_1 = -b_2 = t_{1-\epsilon/2}$ – квантиль стандартного нормального распределения, так что

$$\mathbf{P}(c_1 < \chi < c_2) = 1-\epsilon, \quad \mathbf{P}(b_1 < \eta < b_2) = 1-\delta.$$

Тогда

$$\mathbf{P}\left(\frac{nS^2}{\tau_{1-\epsilon/2}} < \sigma^2 < \frac{nS^2}{\tau_{\epsilon/2}},\ \bar{X} - \frac{St_{1-\epsilon/2}}{\sqrt{\tau_{\epsilon/2}}} < a < \bar{X} + \frac{St_{1-\epsilon/2}}{\sqrt{\tau_{1-\epsilon/2}}}\right) \geq \geq (1-\delta)(1-\epsilon).$$

Осталось только подобрать числа δ, ϵ так, чтобы $(1-\delta)(1-\epsilon) = (1-\alpha)$ произвольным образом, например, положив

$$1-\delta = 1-\epsilon = \sqrt{1-\alpha},$$

и требуемое доверительное множество, имеющее прямоугольную форму, построено.

Для распределений, отличных от нормального, ограничимся рассмотрением алгоритма построения асимптотического доверительного множества в достаточно широких предположениях регулярности. Все доказательства и подробное изложение упомянутых условий можно найти в [4, с. 264 - 265].

Предположим, что нам известен аналитический вид плотности распределения наблюдаемой случайной величины или мы умеем вычислять функцию правдоподобия L, тогда эллипсоидальное доверительное множество для m-мерного параметра задается как

$$\{\vec{t}\,|\, L(X,\vec{t}) - L(X,\vec{\theta^*}) \geq -\tau_\alpha/2\}.$$

Здесь τ_α – квантиль хи-квадрат распределения с k степенями свободы, а $\vec{\theta^*}$ – оценка максимального правдоподобия (многомерного) параметра $\vec{\theta}$.

5.2.2 Многомерная выборка

Рассмотрим для примера построение доверительного множества для вектора $\vec{a}$ математических ожиданий k-мерного нормального распределения

с известной ковариационной матрицей V. Согласно теореме 3 $V^{-1}(\bar{X} - \vec{a}) \cdot (\bar{X} - \vec{a})$ имеет хи-квадрат распределение с k степенями свободы. Обозначая квантиль соответствующего распределения через $\tau_{1-\alpha}$, получаем доверительное множество, имеющие вид эллипсоида, задаваемого соотношением

$$\{\vec{t} \,|\, V^{-1}(\vec{t} - \bar{X}) \cdot (\vec{t} - \bar{X}) \ < \ \tau_{1-\epsilon}\}. \tag{5.1}$$

Можно также построить асимптотически доверительное множество, имеющее вид эллипсоида, отказавшись от предположения нормальности выборки, если для соответствующего k-мерного параметра $\vec{\theta}$ имеется асимптотически нормальная оценка $\vec{\theta^*}$, т.е. распределение $(\vec{\theta^*} - \vec{\theta})\sqrt{n}$ с ростом объема выборки n сходится к многомерному нормальному с нулевым вектором средних и некоторой ковариационной матрицей V, которую мы будем считать известной. Тогда для построения доверительного множества уровня доверия $1 - \alpha$ достаточно построить такой эллипсоид в k-мерном пространстве, который содержит внутри себя соответствующую долю всех значений нормального распределения. Согласно теореме 3 , требуемый эллипсоид имеет вид (5.1), только вместо $\bar{X}$ следует написать $\vec{\theta^*}$.

Построенные доверительные множества можно использовать для проверки некоторых гипотез, например, о равенстве значений некоторого параметра в случае наблюдения двух (и более) выборок. Пусть высказывается гипотеза о равенстве векторов математических ожиданий двух выборок. Построим два доверительных множества A_1, A_2 для математических ожиданий и изучим их пересечение. Если это пересечение пусто или имеет относительно небольшой размер, то гипотезу нужно отвергнуть.

Аналогичным образом можно и проверять гипотезы однородности, применяя для этого вместо доверительных множеств толерантные, построение которых описано в следующем разделе. Сравните предложенную процедуру с ранее описанной в 3.3.2 .

5.3 Толерантные множества

Множество, содержащее заданную долю всех значений наблюдаемой величины (или вектора), называется толерантным множеством заданного уровня. Конечно же, в одномерном случае рассматриваются толерантные интервалы. Как мы видим, задача построения толерантного множества

уровня $1-\alpha$ близка к задаче оценивания границ интервала, образованными квантилями уровней $\alpha/2$ и $1-\alpha/2$ наблюдаемого распределения. Опишем процедуру построения толерантных интервалов для нормального распределения.

Будем считать, что требуется построить толерантный интервал для стандартного нормального распределения. Положим

$$A(\lambda, S) \;=\; \frac{1}{\sqrt{2\pi}} \int\limits_{\bar{X}-\lambda S}^{\bar{X}+\lambda S} \exp\{-t^2/2\} dt.$$

Для заданного ε требуется найти такое λ, что

$$\mathbf{P}(A(\lambda, S) > \gamma) \;=\; 1-\varepsilon.$$

Зафиксируем $\bar{X}$. Единственный (в силу того, что по S функция $A(\lambda, S)$ монотонно возрастает) корень уравнения $A(\lambda, S) \;=\; \gamma$ обозначим $S(\gamma, \lambda)$. Введем также обозначение

$$r(\gamma) \;=\; \lambda S(\gamma, \lambda).$$

Тогда

$$\frac{1}{\sqrt{2\pi}} \int\limits_{\bar{X}-r(\gamma)}^{\bar{X}+r(\gamma)} \exp\{-t^2/2\} dt \;=\; \Phi(\bar{X}+r(\gamma)) \;-\; \Phi(\bar{X}-r(\gamma)) \;=\; \gamma, \quad (5.2)$$

что означает, что при заданных $\bar{X}$, γ число $r(\gamma)$ однозначно определено и не зависит от λ.

Кроме того, неравенство $A(\lambda, S) > \gamma$ эквивалентно неравенству $S > r(\gamma)/\lambda$. Отсюда

$$\mathbf{P}(A(\lambda, S) > \gamma \mid \bar{X}) \;=\; \mathbf{P}\left(nS^2 > \frac{nr^2(\gamma)}{\lambda^2} \mid \bar{X}\right).$$

При этом величина nS^2 по теореме Фишера имеет хи-квадрат распределение с $n-1$ степенями свободы и не зависит от $\bar{X}$, а значит, λ может быть найдено из условия

$$\mathbf{P}\left(\chi^2_{n-1} \;>\; \frac{nr^2(\gamma)}{\lambda^2}\right) \;=\; 1-\varepsilon.$$

Это рассуждение носит, конечно же, чисто теоретический характер. Теперь же укажем один из приближенных методов, дающий обычно достаточно точный результат, для определения безусловной вероятности $\mathbf{P}(A(\lambda, S) > \gamma)$. Разложим функцию

$$f(\bar{X}) \;=\; \mathbf{P}(A(\lambda, S) > \gamma \mid \bar{X})$$

по степеням $\bar{X}$. Поскольку она, очевидно, является четной функцией, то

$$f(\bar{X}) \;=\; f(0) \;+\; \frac{\bar{X}^2}{2!} f''(0) \;+\; o(\bar{X}^4). \tag{5.3}$$

Вычислим математическое ожидание по $\bar{X}$ от обеих частей последнего равенства:

$$\mathbf{P}(A(\lambda, S) > \gamma) \;=\; f(0) \;+ \frac{1}{2n} f''(0) \;+\; o(\frac{1}{n^2}).$$

Подставляя в (5.2) $\bar{X} \;=\; \frac{1}{\sqrt{n}}$, видим, что первые два слагаемых в последнем равенстве можно заменить на $\mathbf{P}(A(\lambda, S) > \gamma | \frac{1}{\sqrt{n}})$, откуда

$$\mathbf{P}(A(\lambda, S) > \gamma) \;\approx\; \mathbf{P}(A(\lambda, S) > \gamma | \frac{1}{\sqrt{n}}).$$

Подведем итог. Для построения толерантного интервала необходимо по заданному γ (уровню доверия) найти число r из (5.2), полагая $\bar{X} \;=\; \frac{1}{\sqrt{n}}$, затем по вероятности $1 - \varepsilon$ найти критическую точку t_ε распределения χ^2_{n-1} и вычислить

$$\lambda \;=\; r\sqrt{\frac{n}{t_\varepsilon}}.$$

Требуемый интервал будет иметь вид

$$(\bar{X} - \lambda S,\; \bar{X} + \lambda S).$$

В многомерном случае нами уже решена задача построения толерантного эллипсоида для нормального распределения с известной ковариационной матрицей, см (5.1).

Изложим один свободный от распределения алгоритм построения толерантного интервала в одномерном случае. За обоснованиями отсылаем читателя к [5, с. 695-697].

Введем обозначения. Через F обозначим (неизвестную) функцию распределения наблюдаемой случайной величины, через $X_{(k)}$ – k-ю порядковую статистику, т.е. k-й по величине выборочный элемент, если выборка упорядочена по возрастанию. Поставим задачу построения толерантного интервала, содержащего долю γ всех значений наблюдаемой случайной величины с вероятностью $1-\alpha$. Как уже отмечалось, естественно искать решение этой задачи в виде

$$\mathbf{P}(F(X_{(s)}) - F(X_{(r)}) \ \geq \ \gamma) \ = \ 1 - \alpha.$$

В [5] показано, что это соотношение эквивалентно следующему

$$\int\limits_{\gamma}^{1} \frac{x^{s-r-1}(1-x)^{n-s+r}}{B(s-r,\ n-s+r+1)} dx \ = \ 1-\alpha,$$

где

$$B(a,b) \ = \ \int\limits_{0}^{1} x^a (1-x)^b dx -$$

бета-функция. Если у исследователь имеет инструментарий для численного расчета неполной бета-функции, то последнее соотношение может быть записано в виде

$$1 \ - \ I_{\gamma}(s-r,\ n-s+r+1) \ = \ 1-\alpha.$$

На практике числа α, γ задаются или заранее известны, а r, s задаются симметрично, так что $s = n - r + 1$. Тогда

$$I_{\gamma}(n-2r+1,\ 2r) \ = \ \alpha.$$

Конечно, же, последнее уравнение относительно r должно решаться численно или при помощи таблиц. А такие существуют! См. [6, с. 284, объяснения на с.67] . Несколько ободряющим обстоятельством на трудном пути построения толерантных интервалов в этой ситуации может служить то, что построенные интервалы годятся для любых исходных распределений, и перед нами не стоит обычная проблема проверки гипотезы нормальности.

5.4 Засоренная и малая выборка

Проблемы борьбы с грубыми ошибками измерений были описаны выше. В этом разделе мы рассмотрим случай, когда после отбрасывания посторонних значений или по другим причинам (например, в случае дорогих или недоступных дополнительных наблюдений) в нашем распоряжении осталась выборка, объем которой недостаточен для принятия решения.

Задачи, связанные с малой выборкой, часто возникают на практике, и им посвящена специальная литература, см. например, [7] . Мы же упомянем здесь только один специальный метод, направленный на увеличение объема выборки – метод статистического моделирования bootstrap, принадлежащий В.Эфрону. Этот метод рекомендует принять имеющуюся выборку за генеральную совокупность и при помощи случайного отбора формировать из нее новые данные.

Глава 6

Регрессионный анализ

6.1 Постановка задачи

Пусть мы наблюдаем две случайные величины и пытаемся понять, зависимы ли они между собой. Если это одномерные случайные величины, то можно проверить гипотезу независимости, используя алгоритмы, описанные в 3.3.1. Если же мы имеем дело с векторами, то можно свести задачу к изучению зависимости их координат.

Предположим, что гипотеза независимости отвергнута, т.е. мы признаем наличие некоторых связей между наблюдаемыми величинами. Попробуем определить характер этой связи и выписать формулы, достаточно точно выражающие количественную сторону этой зависимости. То, что сейчас будет проделано, называется постановкой задачи линейной регрессии.

Наблюдаемый вектор $\vec{Z} = (Z_1, ..., Z_k)$ условимся считать неслучайным. Это предположение делается потому, что при проведении наших выкладок мы будем пользоваться уже имеющимися выборочными данными, которые изменяться не могут. Координаты вектора $\vec{Z}$ принято называть факторами. Величина X, вид зависимости которой от $\vec{Z}$ изучается, будет считаться одномерной и случайной. Случайность появится при расчете ее значений по заданному $\vec{Z}$ в виде случайной помехи. Ставится задача найти наилучшее представление в виде

$$X = \sum_{j=1}^{k} \beta_j Z_j + \epsilon, \tag{6.1}$$

где $\beta_j,\ j = 1, ..., k$ – неизвестные постоянные коэффициенты, называемые коэффициентами регрессии, ϵ - случайная помеха.

Как правило, делается предположение, что ϵ имеет нормальный характер, но мы такого предположения делать не будем. Договоримся только, что математическое ожидание помехи равно 0, а дисперсия σ^2 – она нам неизвестна, но остается неизменной в процессе проведения эксперимента. Назовем ее остаточной дисперсией.

Коэффициенты регрессии и остаточная дисперсия вместе называются параметрами регрессии. Ставится задача по выборочным данным (см. ниже) наилучшим образом оценить параметры регрессии.

Несмотря на кажущийся ограничительный характер постановки задачи, модель линейной регрессии может включать в себя и зависимости более высоких, чем первая, степеней, а также так называемые парные, тройные и т.д. взаимодействия. Для того, чтобы включить в модель (6.1), например, слагаемое $Z_1^2 Z_3$, добавим к вектору, наблюдаемому в нашем эксперименте, еще одну координату, связанную с первой и третьей выписанной формулой. Выборочные данные также дополним одной строкой, которая будет вычисляться по той же формуле.

Сходные действия (добавление в модель регрессии нелинейных слагаемых с одновременным преобразованием выборочных данных) рекомендуется предпринять, если после нахождения коэффициентов регрессии они оказываются слишком малы, или остаточная дисперсия оказывается слишком велика. О выборе подходящей модели регрессии и разнообразных методах оценки ее параметров, отличных от изложенных ниже, можно почитать в замечательной книге [8] , практически целиком посвященной этим вопросам.

Сформулируем требования к выборочным данным. Пусть проведено n независимых экспериментов, в каждом из которых замеряны значения $\vec{Z}$ и переменной X, которая в задачах регрессии называется выходом или откликом. Условимся, что значения случайной помехи ϵ при повторениях эксперимента были независимы друг от друга. Данные собраны в матрице

$$Z = \begin{pmatrix} z_{1,1} & z_{1,2} & \cdots & z_{1,n} \\ z_{2,1} & z_{2,2} & \cdots & z_{2,n} \\ \vdots & & & \vdots \\ z_{k,1} & z_{k,2} & \cdots & z_{k,n} \end{pmatrix},$$

в которой значения факторов $Z_1, ..., Z_k$ в каждом из n экспериментов

расположены по столбцам и в векторе $\vec{X} = (x_1, x_2, \ ... \ x_n)^t$, в котором размещаются соответствующие значения отклика. Знак t означает транспонирование, т.е. $\vec{X}$ на самом деле вектор-столбец.

Введем обозначения

$$\vec{\beta} = (\beta_1, ..., \beta_k)^t, \quad \vec{\epsilon} = (\epsilon_1, ..., \epsilon_n)^t.$$

Теперь мы можем записать задачу линейной регрессии в матричной форме: по заданным Z, $\vec{X}$ определить наилучший вектор коэффициентов регрессии $\vec{\beta}$ и остаточную дисперсию из соотношения

$$\vec{X} = Z^t\vec{\beta} + \vec{\epsilon}, \tag{6.2}$$

причем $\mathbf{cov}\vec{\epsilon} = \sigma^2 I$.

Осталось ввести критерий оптимальности выбираемого набора коэффициентов. В основе традиционно рассматривающегося критерия, немедленно приводящего и к соответствующему методу, лежит геометрическое представление, восходящее еще к К.Ф.Гауссу. Представим себе, что столбцы матрицы Z задают координаты точек в k-мерном пространстве, тогда уравнение (6.1) (без добавки ϵ) задает в этом пространстве гиперплоскость. Поставим задачу так провести эту гиперплоскость, чтобы она проходила как можно ближе к точкам, задаваемым матрицей Z. На языке формул, оптимальный $\vec{\beta}$ ищется из условия

$$\hat{\beta} \ : \ S(\hat{\beta}) = \min_{\vec{\beta}} S(\vec{\beta}), \tag{6.3}$$

где

$$S(\vec{\beta}) = (\vec{X} - Z^t\vec{\beta}) \cdot (\vec{X} - Z^t\vec{\beta}) = \sum_{i=1}^{n}(x_i - \sum_{j=1}^{k}\beta_j z_{j,i})^2.$$

Описанный метод называется методом наименьших квадратов, а $\hat{\beta}$, определенный (6.3), – оценкой коэффициентов по методу наименьших квадратов (ОМНК).

Методы нахождения ОМНК могут быть самыми различными. Например, можно методами математического анализа решить для функции k переменных $S(\vec{\beta})$ задачу на минимум. Традиционный метод наименьших квадратов, привлекающий только понятия линейной алгебры, будет рассмотрен в следующем разделе. А сейчас упомянем еще один метод, который получил название метода центра неопределенностей.

Суть его состоит в переходе от пространства наблюдений к пространству коэффициентов. Сначала зафиксируем некоторое число ε и рассмотрим систему неравенств относительно $\vec{\beta}$

$$-\varepsilon \ \leq \ x_i \ - \ \sum_{j=1}^{k} \beta_j z_{j,i} \ \leq \varepsilon, \ i = 1, ..., n. \tag{6.4}$$

Методами линейного программирования определим наименьшее возможное из чисел ε, при которых эта система имеет непустое множество решений, т.е. имеется хотя бы одно $\vec{\beta}$, удовлетворяющее (6.4). Затем определим это минимальное непустое множество и его геометрический центр $\vec{\beta}^*$. В литературе встречаются разные способы определения этого центра, например,в решения (6.4) вписывается эллипсоид и в качестве $\vec{\beta}^*$ берется пересечение его полуосей. Можно также представить себе, что наше множество решений заполнено однородной массой и тем или иным способом определить центр масс. Так или иначе, но найденное $\vec{\beta}^*$ объявляется оценкой коэффициентов регрессии по методу центра неопределенностей.

6.2 Нормальное уравнение регрессии

Оказывается, ОМНК, определенная в предыдущем разделе, всегда является решением некоторой системы линейных уравнений, выводом которой мы сейчас и займемся.

Лемма 4 *Для двух произвольных k-мерных векторов $\vec{\beta}, \vec{\gamma}$ справедливо соотношение*

$$S(\vec{\beta}) \ = \ S(\vec{\gamma}) \ + \ 2(\vec{Y} - A\vec{\gamma}) \cdot (\vec{\gamma} - \vec{\beta}) \ + \ 2(A(\vec{\gamma} - \vec{\beta})) \cdot (\vec{\gamma} - \vec{\beta}),$$

где $A \ = \ ZZ^t, \ \vec{Y} \ = \ Z\vec{X}$.

Доказательство. Проделаем следующие несложные выкладки:

$$\begin{gathered} S(\vec{\beta}) - S(\vec{\gamma}) \ = \ (\vec{X} - Z^t\vec{\beta}) \cdot (\vec{X} - Z^t\vec{\beta}) - (\vec{X} - Z^t\vec{\gamma}) \cdot (\vec{X} - Z^t\vec{\beta}) \ - \\ -((\vec{X} - Z^t\vec{\gamma}) \cdot (\vec{X} - Z^t\vec{\gamma}) - (\vec{X} - Z^t\vec{\gamma}) \cdot (\vec{X} - Z^t\vec{\beta})) \ = \\ = \ Z^t(\vec{\gamma} - \vec{\beta}) \cdot (\vec{X} - Z^t\vec{\beta}) \ - \ (\vec{X} - Z^t\vec{\gamma}) \cdot Z^t(\vec{\beta} - \vec{\gamma}). \end{gathered}$$

Если мы теперь учтем, что

$$(\vec{X} - Z^t\vec{\gamma}) \cdot Z^t(\vec{\gamma} - \vec{\beta}) \;=\; (Z\vec{X} - ZZ^t\vec{\gamma}) \cdot (\vec{\gamma} - \vec{\beta});$$
$$Z^t(\vec{\gamma} - \vec{\beta}) \cdot (\vec{X} - Z^t\vec{\beta}) \;=$$
$$ZZ^t(\vec{\gamma} - \vec{\beta}) \cdot (\vec{\gamma} - \vec{\beta}) \;+\; (Z\vec{X} - ZZ^t\vec{\gamma}) \cdot (\vec{\gamma} - \vec{\beta}),$$

то окончательно получим

$$S(\vec{\beta}) - S(\vec{\gamma}) \;=\; ZZ^t(\vec{\gamma} - \vec{\beta}) \cdot (\vec{\gamma} - \vec{\beta}) + 2(Z\vec{X} - ZZ^t\vec{\gamma}) \cdot (\vec{\gamma} - \vec{\beta}),$$

что и доказывает лемму.

Матрица A, фигурирующая в формулировке леммы, называется матрицей плана.

Лемма 5 *Матрица плана симметрична и неотрицательно определена. Она положительно определена, если строки матрицы Z линейно независимы,*

Доказательство. Так как $A = ZZ^t$, то $A^t = (Z^t)^tZ^t = ZZ^t = A$, что означает симметричность. Возьмем теперь произвольный k-мерный вектор $\vec{t}$. Тогда

$$A\vec{t} \cdot \vec{t} \;=\; ZZ^t\vec{t} \cdot \vec{t} \;=\; Z^t\vec{t} \cdot Z^t\vec{t} \;\geq\; 0,$$

а это означает неотрицательную определенность. Если для некоторого ненулевого вектора $\vec{t}$ в последнем неравенстве достигается равенство, то $Z^t\vec{t} = 0$, а значит, строки матрицы Z линейно зависимы с коэффициентами $t_1, ..., t_k$.

Уравнение

$$A\vec{\beta} \;=\; \vec{Y} \tag{6.5}$$

называется нормальным уравнением регрессии.

Теорема 6 *Любое решение нормального уравнения регрессии доставляет минимум функции $S(\vec{\beta})$, т.е. является ОМНК. Если матрица плана обратима, то $\hat{\beta} = A^{-1}\vec{Y}$ – несмещенная оценка $\vec{\beta}$, причем*

$$\mathbf{cov}\hat{\beta} \;=\; \sigma^2 A^{-1}.$$

Доказательство. Пусть $\hat{\beta}$ – решение уравнения (6.5). Тогда в силу леммы 4 и неотрицательной определенности A для произвольного $\vec{\beta}$ выполнено

$$S(\vec{\beta}) \;=\; S(\hat{\beta}) + A(\hat{\beta}-\vec{\beta})\cdot(\hat{\beta}-\vec{\beta}) \;\geq\; S(\hat{\beta}).$$

Тем самым доказано, что $\hat{\beta}$ –ОМНК. Нам известно, что

$$\vec{X} \;=\; Z^t\vec{\beta} \;+\; \vec{\epsilon},$$

откуда получаем

$$\vec{Y} \;=\; A\vec{\beta} \;+\; Z\vec{\epsilon}.$$

Сравнивая это соотношение с нормальным уравнением регрессии (6.5), получаем, что $A(\hat{\beta}-\vec{\beta}) \;=\; Z\vec{\epsilon}$, и если матрица плана невырождена, то

$$\hat{\beta} \;-\; \vec{\beta} \;=\; A^{-1}Z\vec{\epsilon}.$$

Вычислим от обеих частей математическое ожидание и учтем, что $\mathbf{M}\vec{\epsilon} = 0$. Получим $\mathbf{M}\hat{\beta} \;-\; \vec{\beta} \;=\; 0$, т.е. мы доказали несмещенность ОМНК. Наконец, в силу несмещенности,

$$\begin{gathered}\mathbf{cov}\hat{\beta} \;=\; \mathbf{M}(\hat{\beta}-\vec{\beta})(\hat{\beta}-\vec{\beta})^t \;=\\ =\; \mathbf{M}(A^{-1}Z\vec{\epsilon}\vec{\epsilon}^{\,t}Z^tA^{-1}) \;=\\ =\; A^{-1}Z\,\mathbf{cov}\vec{\epsilon}\,Z^tA^{-1} \;=\; \sigma^2A^{-1}ZZ^tA^{-1} \;=\; \sigma^2A^{-1}.\end{gathered}$$

Теорема доказана.

Теорема 7 *Статистика $S(\hat{\beta})/(n-k)$ является несмещенной оценкой остаточной дисперсии σ^2 в случае невырожденной матрицы плана.*

Доказательство. Пусть $\vec{\beta}$ – теоретический (неизвестный) вектор коэффициентов регрессии. Тогда

$$\mathbf{M}S(\vec{\beta}) \;=\; \mathbf{M}\vec{\epsilon}\cdot\vec{\epsilon} \;=\; \sum_{j=1}^{n}\mathbf{M}\epsilon_j^2 \;=\; n\sigma^2.$$

Далее, обозначая $\vec{h} \;=\; A(\hat{\beta}-\vec{\beta})$, расписывая скалярное произведение через координаты и используя вид ковариационной матрицы ОМНК из предыдущей теоремы, получаем

$$\mathbf{M}\vec{h}\cdot(\hat{\beta}-\vec{\beta}) \;=\; \sigma^2\mathbf{tr}AA^{-1} \;=\; k\sigma^2.$$

Здесь для матрицы B через $\mathbf{tr}B$ обозначен ее след, т.е. сумма диагональных элементов. Согласно лемме 4,

$$S(\vec{\beta}) \;=\; S(\hat{\beta}) \;+\; \vec{h}\cdot(\hat{\beta}-\vec{\beta}).$$

Вычисляя математические ожидания от обеих частей этого соотношения с учетом ранее сделанных замечаний, получим

$$n\sigma^2 \;=\; \mathbf{M}S(\hat{\beta}) \;+\; k\sigma^2,$$

что и завершает доказательство теоремы.

6.3 Задачи регрессии с ограничениями

Мы рассмотрели случай, когда любое из значений коэффициентов регрессии, полученное в результате наших исследований на оптимальность, нами принималось как допустимое. Но часто бывает так, что полученное значение $\hat{\beta}$ нас не может устроить в силу невозможности его реализации или каких-то иных априорных соображений. Таким образом, мы приходим к задаче регрессии, в которой на коэффициенты наложены некоторые ограничения.

Используя тот же метод, который применяется в задачах линейного программирования при приведении задачи к каноническому виду (имеется ввиду способ замены всех ограничений на равенства при помощи введения искусственных переменных), можно считать, что ограничения на коэффициенты имеют вид

$$f_1(\vec{\beta})=0,\;\; f_2(\vec{\beta})=0,\;\; \ldots\; f_m(\vec{\beta})=0,$$

где каждая из $f_i,\;\; i=1,\ldots,m$ – функция k переменных, принимающая действительные значения.

Теперь наша задача сводится к поиску условного экстремума функции $S(\vec{\beta})$ в выписанных ограничениях. В такой ситуации математический анализ рекомендует прибегнуть к так называемому методу Лагранжа. Введем функцию Лагранжа

$$L(\vec{\beta},\;\lambda_1,\ldots,\lambda_m) \;=\; S(\vec{\beta}) \;-\; \sum_{i=1}^{m}\lambda_i f_i(\vec{\beta}),$$

где $\lambda_1, ..., \lambda_m$ - искусственно введенные переменные и будем исследовать эту функцию на минимум. Найденные значения коэффициентов и будут решением нашей задачи с ограничениями.

В силу линейности нашей задачи особенно важным частным случаем является тот, когда ограничения на коэффициенты имеют линейный вид

$$Q\vec{\beta} \;=\; \vec{a}. \tag{6.6}$$

Здесь $\vec{a}$ – известный вектор размерности m, Q – $m \times k$-матрица, имеющая ранг m. Это предположение не нарушает общности, поскольку означает только, что ни одно из m линейных ограничений не является следствием остальных, и система ограничений непротиворечива.

Теорема 8 *Пусть ограничения имеют вид (6.6), $\hat{\beta}$ - ОМНК параметров регрессии в задаче без ограничений с теми же выборочными данными и матрица плана не вырождена. Тогда решение задачи с ограничениями имеет вид*

$$\hat{\beta_Q} \;=\; \hat{\beta} \;-\; A^{-1}Q^tD^{-1}(Q\hat{\beta} - \vec{a}),$$

где $D \;=\; QA^{-1}Q^t$ – квадратная матрица порядка m.

Доказательство. Прежде всего заметим, что

$$Q\hat{\beta_Q} \;=\; Q\hat{\beta} \;-\; DD^{-1}(Q\hat{\beta} - \vec{a}) \;=\; \vec{a}, \tag{6.7}$$

а так как для ОМНК справедливо (6.5), то

$$\vec{Y} \;-\; A\hat{\beta_Q} \;=\; Q^tD^{-1}(Q\hat{\beta} - \vec{a}).$$

Из (6.7) следует, что

$$Q(\hat{\beta_Q} - \vec{\beta}) \;=\; \vec{a} - Q\vec{\beta}$$

при произвольном $\vec{\beta}$. Осталось заметить, что в силу леммы 4,

$$S(\vec{\beta}) \;=\; S(\hat{\beta_Q}) + 2(\hat{\beta_Q} - \vec{\beta}) \cdot Q^tD^{-1}(Q\hat{\beta} - \vec{a}) + A(\hat{\beta_Q} - \vec{\beta}) \cdot (\hat{\beta_Q} - \vec{\beta}).$$

Если $\vec{\beta}$ удовлетворяет (6.6), то

$$(\hat{\beta_Q} - \vec{\beta}) \cdot Q^tD^{-1}(Q\hat{\beta} - \vec{a}) \;=\; Q(\hat{\beta_Q} - \vec{\beta}) \cdot D^{-1}(Q\hat{\beta} - \vec{a}) \;=$$
$$=\; (\vec{a} - Q\vec{\beta}) \cdot D^{-1}(Q\hat{\beta} - \vec{a}) \;=\; 0,$$

а следовательно,

$$S(\vec{\beta}) = S(\hat{\beta_Q}) \;+\; A(\hat{\beta_Q} - \vec{\beta}) \cdot (\hat{\beta_Q} - \vec{\beta}) \;\geq\; S(\hat{\beta_Q})$$

в силу свойств матрицы плана, что и завершает доказательство.

6.4 Оптимальный выбор матрицы плана

В задачах регрессии, а иногда и в других задачах, связанных с изучениями данных эксперимента, принято рассматривать две различных ситуации в зависимости от возможности ставить дополнительные эксперименты. Если серию экспериментов мы планируем сами, т.е. в состоянии задавать значения факторов в очередном эксперименте по собственному желанию, то говорят, что имеет место ситуация активного эксперимента. Если же такой возможности нет, и мы просто можем записывать, чему равны значения факторов, не в силах вмешаться в их изменение, то эксперимент называется пассивным. Близким (и в основном, тождественным) к ситуации пассивного эксперимента является случай так называемого архивного эксперимента: фактически эксперимент не ставится, а изучаются данные о проводившихся когда-то экспериментах.

Если мы имеем дело с активным экспериментом, то уместно поставить вопрос о том, какие значения следует придать факторам, чтобы оценки параметров регрессии получились бы возможно более точными. В принципе, постановка такого вопроса не лишена смысла и в случае пассивного эксперимента, просто возможностей выбора значений у нас здесь будет скорее всего меньше, ведь все, что мы в состоянии предпринять – это решить, включаем мы наблюдаемый эксперимент в наши данные или подождем до следующего. Будем всюду в этом разделе рассматривать только невырожденные матрицы плана.

Итак, задача поставлена. Естественным критерием надежности оценок коэффициентов являются дисперсии ОМНК $\hat{\beta}_j,\ j = 1, ..., k$ – чем меньше эти дисперсии, тем лучше оценки. Но, согласно теореме 6,

$$\mathbf{D}\hat{\beta}_j \ = \ \sigma^2 (A^{-1})_{j,j}, \quad j = 1, ..., k,$$

а значит, если мы значения всех факторов Z_i заменим на $hZ_i,\ i = 1, ..., k$, то, поскольку матрица $A^{-1} \ = \ (ZZ^t)^{-1}$ заменится на $h^{-2}A^{-1}$, то все дисперсии ОМНК уменьшатся в h^2 раз.

Это наблюдение приводит к пониманию того, что для корректной постановки задачи на минимизацию дисперсий необходимо наложить на строки матрицы Z некоторые ограничения. Обычно такие ограничения имеют вид

$$|Z_{(j)}|^2 \ = \ \sum_{i=1}^{n} z_{j,i}^2 \ = \ a_j^2,\ j = 1, ..., k. \tag{6.8}$$

Здесь $Z_{(j)} \ = \ (z_{j,1}, ..., z_{j,n})$ – j-я строка матрицы Z.

Теорема 9 *Если имеют место ограничения (6.8), то при любом выборе матрицы плана для ОМНК справедливы оценки*

$$\mathbf{D}\hat{\beta}_j \ \geq \ \frac{\sigma^2}{a_j^2}, \ j = 1, ..., k,$$

причем равенство во всех этих неравенствах одновременно достигается тогда и только тогда, когда строки матрицы Z ортогональны, т.е. при произвольных $i \neq j$ скалярное произведение $Z_{(i)} \cdot Z_{(j)} \ = \ 0$.

Доказательство. Заметим, что из определения матрицы плана и условий (6.8)

$$A \ = \ \begin{pmatrix} Z_{(1)} \cdot Z_{(1)} & Z_{(1)} \cdot Z_{(2)} & ... & Z_{(1)} \cdot Z_{(k)} \\ Z_{(2)} \cdot Z_{(1)} & Z_{(2)} \cdot Z_{(2)} & ... & Z_{(2)} \cdot Z_{(k)} \\ \vdots & \vdots & & \vdots \\ Z_{(k)} \cdot Z_{(1)} & Z_{(k)} \cdot Z_{(2)} & ... & Z_{(k)} \cdot Z_{(k)} \end{pmatrix} \ = \ \left(\begin{array}{c|c} a_1^2 & \vec{b}^t \\ \hline \vec{b} & F \end{array}\right),$$

где $\vec{b}^t \ = \ (Z_{(1)} \cdot Z_{(2)}, ..., Z_{(1)} \cdot Z_{(k)})$, а значит $\vec{b}$ – вектор-столбец размерности $k-1$, матрица F получена из матрицы плана вычеркиванием первой строки и первого столбца, а следовательно, является положительно определенной симметричной матрицей. Поскольку определитель

$$\left|\begin{array}{c|c} 1 & 0 \\ \hline -F^{-1}\vec{b} & I_{k-1} \end{array}\right| \ = \ 1,$$

то

$$|A| \ = \ \left|\left(\begin{array}{c|c} a_1^2 & \vec{b}^t \\ \hline \vec{b} & F \end{array}\right) \cdot \left(\begin{array}{c|c} 1 & 0 \\ \hline -F^{-1}\vec{b} & I_{k-1} \end{array}\right)\right| =$$

$$= \left|\begin{array}{c|c} a_1^2 \ - \ F^{-1}\vec{b} \cdot \vec{b} & \vec{b}^t \\ \hline 0 & F \end{array}\right| \ = \ (a_1^2 \ - \ F^{-1}\vec{b} \cdot \vec{b})|F|,$$

откуда

$$a_1^2 \ > \ F^{-1}\vec{b} \cdot \vec{b}.$$

Поскольку

$$(A^{-1})_{1,1} \ = \ \frac{|F|}{|A|} \ = \ \frac{1}{a_1^2 - F^{-1}\vec{b} \cdot \vec{b}}$$

и $F^{-1}\vec{b}\cdot\vec{b} > 0$, то

$$\mathbf{D}\hat{\beta}_1 = \sigma^2(A^{-1})_{1,1} \geq \sigma^2 a_1^{-2},$$

причем равенство достигается в том и только том случае, когда $\vec{b} = 0$, т.е. когда строка $Z_{(1)}$ ортогональна всем остальным. Итак, теорема доказана для $j = 1$. Для остальных j доказательство полностью аналогично.

Заметим, наконец, что в случае ортогональных строк Z

$$A = \begin{pmatrix} a_1^2 & \dots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \dots & a_k^2 \end{pmatrix},$$

а значит,

$$\hat{\beta}_j = \frac{Z_{(j)}\cdot\vec{X}}{a_j^2}, \quad j = 1, \dots, k. \tag{6.9}$$

6.5 Задача статистического прогноза

Рассмотрим задачу, похожую на задачу регрессии. Отличие этих задач будет состоять в том, что в рассматриваемой ниже задаче статистического прогноза случайность вмешивается в наш эксперимент в более общей форме - на этапе формирования факторов $Z_1, \dots, Z_k$.

Предположим, что случайный вектор $\vec{Z}$ размерности k доступен для наблюдения, а случайная величина X недоступна. Ставится задача угадать значение X по $\vec{Z}$. Любая функция ϕ, заданная на k-мерном пространстве, принимающая действительные значения и такая, что мы будем использовать $\phi(\vec{Z})$ вместо X, называется предиктором X по $\vec{Z}$. Другими словами, предиктор – это оценка X по $\vec{Z}$.

В ситуации, когда X не зависит от $\vec{Z}$ задача прогноза (оценки) X по $\vec{Z}$ лишена какого-либо смысла. Обычно считается что с теоретической точки зрения известно совместное распределение X и $\vec{Z}$, а значит и условные математические ожидания при фиксированном $\vec{Z}$. Какими же данными мы располагаем на практике?

Допустим, что в нашем распоряжении имеется достаточно обширный архив – сведения о том, какие значения принимали факторы и какие значения предсказываемая величина в предыдущих экспериментах. По этим данным разными методами (например, методом подстановки)

можно оценить любые характеристики совместного распределения: $\mathbf{M}X$, $\mathbf{cov}(X,\ Z_j)$, $j = 1, ..., k$ и т.п. В частности, если архив достаточно велик, можно выбрать из него сведения о разных значениях X при фиксированном наборе значений $\vec{Z} = \vec{Z}^{(0)} = (z_1^0, ..., z_k^0)^t$ и рассчитать

$$M^*(\vec{Z}^{(0)}) = \frac{1}{r}\sum_{j=1}^{r} X_{n_j},$$

где r – это число выборочных данных с набором $\vec{Z}^{(0)}$ в качестве $\vec{Z}$, а X_{n_j} – соответствующие значения предсказываемой величины. Если мы вернемся к теоретической точке зрения, то считаем, что функция

$$M(\vec{Z}^{(0)}) = \mathbf{M}(X \,/\, \vec{Z} = \vec{Z}^{(0)})$$

нам известна при каждом значении $\vec{Z}^{(0)}$, а $M^*(.)$ является ее оценкой. Эта функция называется функцией регрессии.

Говорят, что предиктор ϕ^* оптимален (в смысле среднего квадратического), если

$$\mathbf{M}(X - \phi^*(\vec{Z}))^2 = \min_{\phi} \mathbf{M}(X - \phi(\vec{Z}))^2.$$

Теорема 10 *Оптимальный предиктор всегда существует и имеет вид* $\phi^*(\vec{Z}) = M(\vec{Z})$, *т.е. получается подстановкой случайного вектора наблюдаемых величин в функцию регрессии.*

Как становится ясно из приведенных выше рассуждений, эта теорема имеет чисто теоретическое значение, ведь на практике знание функции регрессии – вещь весьма и весьма нечастая. Задача определения этой функции непроста даже для нормальных распределений. С другой стороны, справедливость утверждения теоремы совершенно очевидна с геометрической точки зрения – см. геометрическую интерпретацию в разделе 2.5. Поэтому эту теорему мы не будем доказывать.

Пусть нам заранее известно, что функция регрессии линейна, т.е. нашлись такие число β_0 и вектор $\vec{\beta} = (\beta_1, ..., \beta_k)^t$, что

$$M(\vec{Z}) = \beta_0 + \vec{\beta} \cdot \vec{Z}.$$

Если бы эти число и вектор нам удалось бы определить, то, согласно теореме, оптимальный предиктор имел бы вид

$$\phi^*(\vec{Z}) \;=\; \beta_0 \;+\; \vec{\beta}\cdot\vec{Z}, \tag{6.10}$$

и задача прогноза была бы успешно решена. Займемся оценкой параметров β.

Рассмотрим $S(\beta) \;=\; \mathbf{M}(X-\phi^*(\vec{Z}))^2$, где ϕ^* задан формулой (6.10). Тогда

$$\begin{gathered} S(\beta) \;=\; \mathbf{M}(\,(X-\mathbf{M}X) \;+\; b \;-\; \vec{\beta}\cdot(\vec{Z}-\mathbf{M}\vec{Z})\,)^2 \;= \\ =\; \mathbf{D}X \;+\; b^2 \;+\mathbf{M}(\vec{\beta}\cdot(\vec{Z}-\mathbf{M}\vec{Z}))^2 \;-\; 2\vec{\beta}\cdot\vec{a}, \end{gathered}$$

где $b \;=\; \mathbf{M}X-\beta_0-\vec{\beta}\cdot\mathbf{M}\vec{Z}$ – константа, k-мерный вектор $\vec{a}$ имеет в качестве своей i-й координаты $a_i \;=\; \mathbf{cov}(X,Z_i),\; i=1,...,k$. Так как $S(\beta)$ минимальна по β, то, очевидно, $b=0$, или

$$\beta_0 \;=\; \mathbf{M}X \;-\; \vec{\beta}\cdot\mathbf{M}\vec{Z}. \tag{6.11}$$

Введем в рассмотрение функцию

$$T(\vec{\beta}) \;=\; \mathbf{M}(\vec{\beta}\cdot(\vec{Z}-\mathbf{M}\vec{Z})\,)^2$$

и исследуем ее на минимум. Очевидно, что частная производная ее по переменной $\beta_j,\; j=1,...,k$ равна

$$\mathbf{M}\left(2\sum_{i=1}^{k}\beta_i(Z_i-\mathbf{M}Z_i)(Z_j-\mathbf{M}Z_j)\right)-2a_j \;=\; 2\sum_{i=1}^{k}\beta_i\mathbf{cov}(Z_i,Z_j)-2a_j.$$

Итак, приравнивая все эти производные к нулю, видим, что необходимо решить систему, которая в матричной записи имеет вид

$$B\vec{\beta} \;=\; \vec{a}, \tag{6.12}$$

где $B \;=\; \mathbf{cov}\vec{Z}$. Таким образом, если ковариационная матрица обратима, то из (6.11) и (6.12) следует, что

$$\beta_0 \;=\; \mathbf{M}X \;-\; B^{-1}\vec{a}\cdot\mathbf{M}\vec{Z}.$$

Решение этой системы действительно доставляет минимум (по крайней мере, нестрогий) функции S, поскольку матрица вторых производных B неотрицательно определена. Таким образом, нами доказана

Теорема 11 *Если* $B = \mathbf{cov}\vec{Z}$ *обратима и функция регрессии линейна, то оптимальный предиктор задается формулой*

$$\phi^*(\vec{Z}) = \mathbf{M}X + B^{-1}\vec{a} \cdot (\vec{Z} - \mathbf{M}\vec{Z}), \tag{6.13}$$

где $\vec{a}$ *–* k*-мерный вектор с координатами*

$$a_j = \mathbf{cov}(X, Z_j), \quad j = 1, ..., k.$$

Откажемся теперь от каких бы то ни было предположений о виде функции регрессии и подойдем к задаче с другой стороны. Будем искать наилучший предиктор среди тех, которые имеют линейный характер, т.е. такой ϕ^*, который при некотором наборе параметров β имеет вид (6.13) и оптимален в классе всех предикторов такого вида.

Таким образом, исходя из условия оптимальности, надо найти такой набор $\vec{\beta}$, что введенная выше функция $S(\vec{\beta})$ на нем достигает своего минимального значения. Но эта задача нами только что была решена, поэтому имеет место

Теорема 12 *Если матрица* $\mathbf{cov}\vec{Z}$ *обратима, то оптимальный линейный предиктор существует и единственен. При этом он задается формулой (6.13)*

Заметим, что, если B не является обратимой, то наилучших линейных предикторов бесконечно много: в качестве $\vec{\beta}$ можно взять любое из решений уравнения (6.12), а затем определить β_0 по (6.11). При этом, если ϕ_1, ϕ_2 – два предиктора, определенных таким образом, то

$$\mathbf{M}(X - \phi_1(\vec{Z}))^2 = \mathbf{M}(X - \phi_2(\vec{Z}))^2.$$

Глава 7

Дисперсионный анализ

7.1 Вводные замечания

Рассмотрим задачу выяснения наличия и оценки степени влияния некоторого фактора A на случайную величину X. Под фактором A условимся понимать величину, которая является нечисловой категоризованной или числовой, принимающей небольшое число различных значений. Категории или значения случайной величины A принято называть уровнями фактора.

Чтобы немного прояснить ситуацию, рассмотрим зависимость величины урожая от внесения в почву определенного вещества (удобрения). В простейшем случае имеется два уровня фактора – было удобрение внесено или нет. Но возможны и варианты: внесена двойная, тройная, полуторная доза удобрения, В этом случае число уровней фактора повышается.

Пусть нам заранее известна дисперсия величины X в случае, когда фактор A не действовал, обозначим ее через $\mathbf{D}_0X$. Теперь "включим в действие"фактор A и вычислим (или хотя бы оценим) дисперсию $\mathbf{D}X$ по полному набору данных. Если фактор A не оказывал влияния на изменчивость X, то $\mathbf{D}X$ не должна сильно отличаться от $\mathbf{D}_OX$. Если же $\mathbf{D}X$ значительно больше, чем $\mathbf{D}_0X$, то следует признать вклад фактора в изменчивость наблюдаемой случайной величины значительным. Вообще говоря,

$$\mathbf{D}X \;=\; \mathbf{D}_OX + \mathbf{D}_AX,$$

где через $\mathbf{D}_AX$ обозначана часть дисперсии, объясняемой влиянием фак-

тора A. Если же исследуемых факторов несколько, то

$$\mathbf{D}X \;=\; \mathbf{D}_O X \;+\; \mathbf{D}_A X \;+\; \mathbf{D}_B X \;+\; \mathbf{D}_{A,B} X \;+\; ...$$

Идея оценки степени влияния факторов основана на изучении доли той дисперсии, которая объясняется через изучаемый фактор в полной дисперсии. Она была предложена Р.Фишером в 1920 году.

Дисперсии X, рассчитанные в предположении, что каждый из факторов зафиксирован в каком-то из своих уровней и не меняется, называют частными дисперсиями.

Сформулируем здесь основные предположения, необходимые для применения описываемого далее инструментария:

1. наблюдаемая величина имеет нормальное распределение;
2. изучаемый фактор или факторы оказывают влияние на среднее значение изучаемой величины;
3. все частные дисперсии однородны, т.е. их различия незначимы.

Таким образом, любое исследование с применением дисперсионного анализа, претендующее на достоверность получаемых результатов, должно начинаться с проверки этих трех предположений. И если второе предположение проверяется, в основном, исходя из опыта экспериментатора (например понятно, что внесение удобрений влияет именно на среднюю величину урожая), то для проверки первого и третьего предположений статистик располагает достаточно разработанным аппаратом. Позволим себе напомнить хотя бы по одному способу проверки этих предположений (их, разумеется, гораздо больше).

7.1.1 Проверка гипотезы нормальности

Будем применять так называемый критерий Пирсона (критерий хи-квадрат). Сначала выполним группировку выборки в r групп $\Delta_j = (z_{j-1},\, z_j),\ j = 1, ..., r$. Как это сделать, было объяснено в разделе 3.1. Обозначим $n_j,\ j = 1, ..., r$ количества элементов выборки, попавших в j-ю группу и вычислим статистику

$$\chi^2 \;=\; \sum_{j=1}^{r} \frac{(n_j - np_j)^2}{np_j}, \tag{7.1}$$

где n – объем выборки, а числа p_j представляют собой вероятности попадания нормально распределенной случайной величины в соответствующий интервал, являющийся нашей группой. При этом предполагается, что параметры нормального распределения заменены их оценками максимального правдоподобия. Таким образом, если через Φ обозначена функция стандартного нормального распределения, то

$$p_j \;=\; \Phi\left(\frac{z_j - \bar{X}}{S}\right) \;-\; \Phi\left(\frac{z_{j-1} - \bar{X}}{S}\right). \tag{7.2}$$

Известна принадлежащая К.Пирсону и Р.Фишеру

Теорема 13 *Если выборка была произведена из нормального распределения, то статистика χ^2, вычисляемая по формулам (7.1) и (7.2) имеет распределение хи-квадрат с $r-3$ степенями свободы.*

Доказательства мы здесь не приводим. Напомним только, что если бы при вычислениях p_j мы использовали бы не оценки, а точно известные значения среднего и корня квадратного из дисперсии нормального распределения, то число степеней свободы хи-квадрат увеличилось бы до $r-1$, если же только один из параметров (любой) пришлось бы по ходу вычислений оценивать, то $r-2$.

Таким образом, проверку нормальности выборки осуществляем по следующей схеме. Сначала вычислим среднее $\bar{X}$ и выборочную дисперсию S^2. Затем произведем группировку и заполним следующую таблицу:

Проверка гипотезы нормальности

строка	содержание	способ вычисления
1	z_j	по выборке
2	$(z_j - \bar{X})/S$	по строке 1 и выборке
3	$\Phi\left(\frac{z_j-\bar{X}}{S}\right)$	по таблице $\Phi(x)$ и строке 2
4	p_j	по формуле (7.2) и строке 3
5	n_j	по выборке
6	$(n_j - np_j)^2$	по строкам 4,5
7	$(n_j - np_j)^2/np_j$	по строкам 4,6

Сумма последней строки и есть значение χ^2. После вычисления оно сравнивается с критической точкой распределения хи-квадрат с $r-3$ степенями свободы, и если расчетное значение меньше критического, то гипотезу о нормальности распределения можно принять.

7.1.2 Однородность дисперсий

Для проверки несущественности отличия r выборочных дисперсий (рассчитанных при фиксации фактора на определенных уровнях) после принятия предположения нормальности выборки, можно воспользоваться критерием Бартлетта, основанном на статистике

$$M \;=\; n \ln\left(\frac{1}{n}\sum_{i=1}^{r} n_i s_i^2\right) \;-\; \sum_{i=1}^{r} n_i \ln s_i^2.$$

Здесь s_i^2 – выборочные дисперсии в i-й группе наблюдений, n_i – количество наблюдений в этой группе , $i = 1, ..., r$, n – общее количество наблюдений, $n = \sum_{i=1}^{r} n_i$.

Теорема 14 *Если гипотеза о равенстве всех дисперсий верна, выборка нормальна и все n_i больше 3, то отношение*

$$M\left(1 \;+\; \frac{1}{3(r-1)}\left(\sum_{i=1}^{r}\frac{1}{n_i} - \frac{1}{n}\right)\right)^{-1}$$

имеет приближенно распределение хи-квадрат с $r-1$ степенью свободы.

Таким образом, для того, чтобы проверить гипотезу об однородности дисперсий, нужно рассчитать выписанное отношение и сравнить его с критической точкой распределения хи-квадрат. Если критическое значение не превзойдено, то отвергать гипотезу однородности нет оснований.

Если все числа n_i равны между собой, то можно воспользоваться более просто вычисляемым критерием Кокрена: вычислим отношение максимального из s_i^2 к их сумме

$$G \;=\; \frac{s_{max}^2}{\sum_{i=1}^{r} s_i^2}$$

и сравним с критическим значением G по специальной таблице. Эти таблицы можно найти в [1, с.156] и [6, с.242].

7.2 Однофакторный анализ. Распределение Фишера

Выделяются следующие разновидности дисперсионного анализа: по числу изучаемых факторов влияния (одно-, двух- , многофакторный), по числу уровней фактора (двух-, трехуровневый ...), по наличию и отсутствию параллельных испытаний (т.е. повторных испытаний при условии фиксации уровней всех факторов влияния). Различают также полный (имеются данные при всех наборах значений факторов) и дробный дисперсионный анализ. На самом деле для решения задач выявления степени влияния факторов различия между полным и дробным анализом несущественны – важно лишь, чтобы при фиксации одного из факторов на любом своем уровне нашлось хотя бы одно экспериментальное данное во всем массиве данных, в котором выбранный фактор фиксирован именно на этом уровне. Другими словами, забегая немного вперед, надо чтобы в каждом из рядов заполняемой таблицы было хотя бы одно значение.

Рассмотрим подробнее полный однофакторный анализ с параллельными испытаниями. Будем предполагать, что на каждом из m уровней фактора A поставлено одинаковое число опытов n по наблюдению случайной величины X. Ее значения, наблюденные в i-м опыте при фиксации фактора на j-м уровне обозначены $x_{i,j},\ i = 1, ..., n,\ j = 1, ..., m$. Данные соберем в таблицу:

Данные для однофакторного анализа

	Уровни		
Испытание	A_1	...	A_m
1	$x_{1,1}$	...	$x_{1,m}$
$\vdots$	$\vdots$	$\vdots$	$\vdots$
n	$x_{n,1}$	...	$x_{n,m}$
средние	$\bar{x_1}$	...	$\bar{x_m}$

Объем полного набора наблюдений здесь, таким образом, равен mn. Обозначим

$$\bar{X} \ = \ \frac{1}{m}\sum_{j=1}^{m}\bar{x_j} \ = \ \frac{1}{nm}\sum_{i=1}^{n}\sum_{j=1}^{m}x_{i,j}.$$

Нетрудно заметить, что если

$$Q = \sum_{j=1}^{m}\sum_{i=1}^{n}(x_{i,j} - \bar{X})^2, \quad Q_O = \sum_{j=1}^{m}\sum_{i=1}^{n}(x_{i,j} - \bar{x_j})^2,$$

$$Q_A = n\sum_{j=1}^{m}(\bar{x_j} - \bar{X})^2,$$

то

$$Q \;=\; Q_O \;+\; Q_A.$$

При этом Q интерпретируется как общая изменчивость X, Q_O как сумма изменчивостей внутри уровней, а Q_A – изменчивость между уровнями, т.е. при переходе от уровня к уровню.

После очевидных преобразований формулы приобретают вид

$$\begin{aligned} Q &= \textstyle\sum_{j=1}^{m}\sum_{i=1}^{n} x_{i,j}^2 \;-\; \frac{1}{mn}\left(\sum_{j=1}^{m}\sum_{i=1}^{n} x_{i,j}\right)^2; \\ Q_O &= \textstyle\sum_{j=1}^{m}\sum_{i=1}^{n} x_{i,j}^2 \;-\; \frac{1}{n}\sum_{j=1}^{m}(\sum_{i=1}^{n} x_{i,j})^2; \\ Q_A &= \textstyle\frac{1}{n}\sum_{j=1}^{m}(\sum_{i=1}^{n} x_{i,j})^2 \;-\; \frac{1}{nm}\left(\sum_{j=1}^{m}\sum_{i=1}^{n} x_{i,j}\right)^2. \end{aligned} \tag{7.3}$$

Определим

$$D_A \;=\; \frac{Q_A}{m-1}, \quad D_O \;=\; \frac{Q_O}{m(n-1)}, \quad F \;=\; \frac{D_A}{D_O}.$$

Теперь сравним рассчитанное по выборочным данным значение F с критической точкой распределения Фишера с $m-1, m(n-1)$ степенями свободы. Если критическое значение не превзойдено, то следует принять гипотезу об отсутствии значимого влияния фактора A на величину X.

Q_A иногда называют факторной вариативностью, Q_O – случайной вариативностью, Q – общей вариативностью. Принято считать, что отношение Q_A/Q дает долю общей вариативности, объясняемой через изменение фактора A.

Введем здесь понятие распределения Фишера, потому что в дисперсионном анализе оно встречается особенно часто. Пусть случайные величины ξ и η имеют распределения хи-квадрат соответственно с k и m степенями свободы и независимы. Тогда говорят, что величина $\zeta = \frac{m\xi}{k\eta}$ имеет распределение Фишера с n, m степенями свободы. Обозначается это распределение $F_{k,m}$ и иногда называется F-распределением. Докажем одно полезное свойство F-распределения.

Лемма 6 *Если* $F^*(k;m;\alpha)$ *– квантиль распределения* $F_{k,m}$ *уровня* α*, то*

$$F^*(m;k;1-\alpha) \;=\; \frac{1}{F^*(k;m;\alpha)}.$$

Доказательство. Пусть ζ имеет распределение Фишера с k, m степенями свободы. Тогда, по определению, $1/\zeta$ имеет распределение Фишера с m, k степенями свободы, и обе случайные величины положительны. Осталось лишь заметить, что для произвольного значения t справедливо

$$\mathbf{P}(\zeta < t) \;=\; \mathbf{P}(\frac{1}{\zeta} > \frac{1}{t}) \;=\; 1 \;-\; \mathbf{P}(\frac{1}{\zeta} < \frac{1}{t}),$$

что и завершает доказательство.

Эта лемма позволяет ограничиваться использованием таблицы $F_{k,m}$ только при $k < m$ (или наоборот). Такие таблицы содержатся в приложении.

Рассмотрим, наконец, следующий пример, приведенный в [2]. Сравнивается три различных метода преподавания. Результаты тестирования трех групп по 15 человек, обученных по разным методикам, приведены в таблице.

Три метода преподавания

Учащийся	Метод 1	Метод 2	Метод 3	
1	9	15	18	
2	11	16	14	
3	10	15	17	
4	12	10	9	
5	7	13	14	
6	11	14	17	
7	12	15	16	
8	10	7	15	
9	13	13	16	
10	11	15	8	
11	13	15	14	
12	11	14	10	
13	10	11	16	
14	12	15	15	
15	13	10	17	Всего
суммы S_j	165	198	216	$t = 579$
суммы квадратов S_j^2	1853	2706	3242	$u = 7801$
	27225	39204	46656	$v = 113085$

В этом примере $mn = 45$. Вычислим $z = (S_1^2 + S_2^2 + S_3^2)/15 = 7539,0$. Далее,

- Факторная вариативность $Q_A = z - \frac{t^2}{mn} = 89,2,\ D_A = 44,6.$
- Случайная вариативность $Q_O = u - z = 262,0,\ D_O = 5,95.$
- Общая вариативность $Q = u - \frac{t^2}{mn} = 351,2.$
- Значение критерия $F = 7,49.$

По таблице распределения Фишера с 2, 42 степенями свободы находим квантиль уровня 0,99. Она равна 5,18. Поскольку рассчитанное значение критерия больше, то следует признать существенное влияние метода преподавания на данное тестирование. Более того, определяя долю Q_A в Q, видим, что изменение методики преподавания дает примерно четверть $(Q_A/Q \approx 0,254)$ общей изменчивости.

7.3 Полный двухфакторный анализ с равными количествами параллельных испытаний

Основное соотношение при наличии двух факторов приобретает вид

$$\mathbf{D}X \;=\; D_AX \;+\; D_BX \;+\; D_{AB}X \;+\; D_OX,$$

где $D_AX, D_BX, D_{AB}X$ – части дисперсии наблюдаемой величины, объясняемые изменчивостью факторов A, B и совместной изменчивостью обоих факторов соответственно, D_O – остаточная часть, не объяснимая с точки зрения рассматриваемых факторов. При исследовании какой-либо практической задачи этим методом следует иметь ввиду, что выбираемые факторы должны быть практически независимы между собой, поскольку, если изменение фактора A приводит к (заранее понятному) изменению фактора B, то корректное решение задачи в предлагаемой форме невозможно.

В этой ситуации данные для проведения дисперсионного анализа располагаются в трехвходовой таблице: по одному ее измерению располагаются обозначения t уровней фактора A, по другому – g уровней фактора B, по третьему – n значений величины X, полученных при фиксации факторов A, B на определенных уровнях. С учетом этого условимся элементы таблицы обозначать $x_{ijk},\ i = 1, ..., t,\ j = 1, ..., g,\ k = 1, ..., n$. Как понятно из заголовка раздела, здесь мы ограничиваемся случаем, когда в каждой "клетке", возникающей при фиксировании уровней A, B содержится одно и то же число $n > 1$ наблюдений. Дело в том, что отказ от этого предположения ведет к значительному усложнению и без того непростых формул (см. ниже). Интересующихся общим случаем, логика которого в принципе ничем не отличается от рассматриваемого здесь более простого варианта отсылаем к [9], [10, главы 35-37].

Перейдем к формулам. Сохраняя смысл обозначений предыдущего раздела (но увеличивая количество этих обозначений), запишем

$$Q_A = \frac{1}{ng}\sum_{i=1}^{t}\left(\sum_{j=1}^{g}\sum_{k=1}^{n} x_{ijk}\right)^2 \;-\; \frac{1}{ngt}\left(\sum_{i=1}^{t}\sum_{j=1}^{g}\sum_{k=1}^{n} x_{ijk}\right)^2,$$

$$Q_B = \frac{1}{nt}\sum_{j=1}^{g}\left(\sum_{i=1}^{t}\sum_{k=1}^{n} x_{ijk}\right)^2 \;-\; \frac{1}{ngt}\left(\sum_{i=1}^{t}\sum_{j=1}^{g}\sum_{k=1}^{n} x_{ijk}\right)^2,$$

$$Q_{AB} = \frac{1}{n}\sum_{i=1}^{t}\sum_{j=1}^{g}\left(\sum_{k=1}^{n} x_{ijk}\right)^2 - \frac{1}{ng}\sum_{i=1}^{t}\left(\sum_{j=1}^{g}\sum_{k=1}^{n} x_{ijk}\right)^2 -$$

$$-\frac{1}{nt}\sum_{j=1}^{g}\left(\sum_{i=1}^{t}\sum_{k=1}^{n} x_{ijk}\right)^2 + \frac{1}{ngt}\left(\sum_{i=1}^{t}\sum_{j=1}^{g}\sum_{k=1}^{n} x_{ijk}\right)^2,$$

$$Q_O = \sum_{i=1}^{t}\sum_{j=1}^{g}\sum_{k=1}^{n} x_{ijk}^2 - \frac{1}{n}\sum_{i=1}^{t}\sum_{j=1}^{g}\left(\sum_{k=1}^{n} x_{ijk}\right)^2,$$

$$Q = \sum_{i=1}^{t}\sum_{j=1}^{g}\sum_{k=1}^{n} x_{ijk}^2 - \frac{1}{ngt}\left(\sum_{i=1}^{t}\sum_{j=1}^{g}\sum_{k=1}^{n} x_{ijk}\right)^2.$$

После этого рассчитаем оценки всех компонент, разделив соответствующие величины Q на количества степеней свободы:

$$D_A X = \frac{Q_A}{t-1}, \quad D_B X = \frac{Q_B}{g-1}, \quad D_{AB} X = \frac{Q_{AB}}{(t-1)(g-1)},$$

$$D_O X = \frac{Q_O}{(n-1)tg}, \; DX = \frac{Q}{tgn-1}.$$

Значения критерия для оценки наличия частного влияния каждого из факторов и их взаимодействия вычисляются по формулам

$$F_A = \frac{D_A X + D_{AB} X}{D_O X}, \; F_B = \frac{D_B X + D_{AB} X}{D_O X},$$

$$F_{AB} = \frac{D_{AB} X}{D_O X},$$

а затем сравниваются с критическими точками распределения Фишера. Числа степеней свободы при этом считаются так: первая равна сумме чисел свободы слагаемых числителя, вторая – число степеней свободы знаменателя ($(n-1)tg$). Например, для F_A получим распределение $F_{(t-1)g,(n-1)tg}$.

Приведем числовой пример, заимствованный из книги G.A.Ferguson. Statistical Analysis In Psychology and Education, McGraw-Hill Inc., 1966. Исследовалось время прохождения крысами лабиринта в зависимости от двух факторов – степени активности (A) и условий воспитания (B). Выделялись следующие уровни первого фактора – отличная активность

(A_1), средняя (A_2) и низкая (A_3). Для второго фактора имелось два уровня – свободные условия (B_1) и воспитание в клетке (B_2). При каждом сочетании уровней было поставлено $n = 8$ повторных испытаний. Данные экспериментов собраны в таблице.

Время прохождения лабиринта
в зависимости от активности A и воспитания B

	k	A_1	A_2	A_3		k	A_1	A_2	A_3
	1	26	41	36		1	51	39	42
	2	41	26	39		2	96	104	92
	3	28	19	59		3	97	130	156
	4	92	59	27		4	22	122	144
B_1	5	14	82	87	B_2	5	35	114	133
	6	16	86	99		6	36	92	124
	7	29	45	126		7	28	87	68
	8	31	37	104		8	76	64	142

Промежуточные данные вычислений собраны в следующей таблице

Суммы наблюдений в "клетках"

	A_1	A_2	A_3	суммы по строкам
B_1	277	395	577	1249
B_2	441	752	901	2094
суммы по столбцам	718	1147	1478	3343

По этой таблице нетрудно подготовить данные, необходимые для вычислений вариативностей Q:

$$\sum_{j=1}^{3}\sum_{i=1}^{2}\left(\sum_{k=1}^{8}x_{ijk}\right)^2 = 2137469,\ \sum_{j=1}^{3}\left(\sum_{i=1}^{2}\sum_{k=1}^{8}x_{ijk}\right)^2 = 4015617;$$

$$\sum_{i=1}^{2}\left(\sum_{j=1}^{3}\sum_{k=1}^{8}x_{ijk}\right)^2 = 5944837;\ \left(\sum_{j=1}^{3}\sum_{i=1}^{2}\sum_{k=1}^{8}x_{ijk}\right)^2 = 11175649.$$

Кроме того, по исходной таблице данных можно получить

$$\sum_{j=1}^{3}\sum_{i=1}^{2}\sum_{k=1}^{8}x_{ijk}^2 = 309851.$$

Заметим по ходу дела, что для того, чтобы избегнуть таких больших числовых значений, можно было перед началом вычислений из всех данных нашей таблицы вычесть одно и то же число – в нашем случае удобно, например, 80, – поскольку величина дисперсии не зависит от постоянного сдвига, а абсолютные величины вычисляемых характеристик при этом

станут более "удобоваримы". Однако эти вопросы нас не волнуют, потому что сегодня все промежуточные вычисления проделает компьютер, и нам не придется заботиться о таких "мелочах".

Остальные результаты приводятся ниже.

Результаты двухфакторного анализа
в примере с лабиринтом

вид вариативности	Q	ст.св.	оценка дисперсии D
по фактору A	14875,52	2	7437,76
по фактору B	18150,04	1	18150,04
взаимодействие AB	1332,04	2	666,02
остаточная (О)	42667,38	42	1015,89
общая	77024,98	47	1638,83

Наконец, рассчитаем значения критериев.

$$F_A = 7,98; \quad F_B = 18,52; \quad F_{A,B} = 0,66.$$

Отметим, что

$$F_A > F^*(4;42;0,99) = 3,83; \; F_B > F^*(3;42;0,99) = 4,31;$$

$$F_{A,B} < F^*(2;42;0,99) = 5,18,$$

где через $F^*(k;m;\alpha)$ обозначены квантили распределения Фишера с k,m степенями свободы уровня α.

Это означает, что по отдельности вклады факторов A,B в изменчивость времени прохождения лабиринта значительны (их можно оценить как соответственно $Q_A/Q \approx 0,19$ и $Q_B/Q \approx 0,24$ всей изменчивости этого времени), а вклад, объясняемый через взаимодействие факторов, составляет $Q_{AB}/Q \approx 0,02$ менее двух процентов, и мы можем не учитывать это влияние.

7.4 Некоторые рекомендации относительно общего случая

В заключение этой главы несколько слов о том, как производить расчеты в случае неравных чисел наблюдений в клетках и числа факторов, большего двух. Пусть сначала у нас имеется один фактор и m уровней

с количеством опытов $n_1, ..., n_m$ в клетках соответственно, $N = \sum_{j=1}^{m} n_j$. При этом формулы (7.3) заменятся на

$$\begin{aligned} Q &= \sum_{j=1}^{m}\sum_{i=1}^{n_j} x_{i,j}^2 - \frac{1}{N}\left(\sum_{j=1}^{m}\sum_{i=1}^{n_j} x_{i,j}\right)^2; \\ Q_O &= \sum_{j=1}^{m}\sum_{i=1}^{n_j} x_{i,j}^2 - \sum_{j=1}^{m}\frac{(\sum_{i=1}^{n_j} x_{i,j})^2}{n_j}; \\ Q_A &= \sum_{j=1}^{m}\frac{(\sum_{i=1}^{n_j} x_{i,j})^2}{n_j} - \frac{1}{N}\left(\sum_{j=1}^{m}\sum_{i=1}^{n_j} x_{i,j}\right)^2. \end{aligned} \tag{7.4}$$

Остальные расчеты производятся так же, как и в случае одинаковых чисел наблюдений в клетках. Единственное отличие - в вычислении числа степеней свободы (и, соответственно, нормировочного коэффициента - знаменателя для D_O). Он будет равен $N - m$.

Рассмотрим вкратце ситуацию с числом факторов, большим двух (при этом равного количества опытов в клетках не требуется). Заметим также, что случай двух факторов тоже может обрабатываться по излагаемой ниже методике, причем результаты совпадут с теми, что были изложены в предыдущем разделе.

Оценим влияние каждого из факторов в отдельности. Для этого все имеющиеся данные соберем в "укрупненные клетки", соответствующие изменению уровней этого фактора (как бы считая, что остальные факторы отсутствуют). В случае, например, двух факторов, клетки по которым расположены в квадратной таблице, укрупнению будут соответствовать клетки-столбцы, в которые объединены все данные, располагавшиеся ранее в этом столбце. После такого объединения используем формулы (7.4).

Для оценивания взаимодействий факторов оставим только клетки, отвечающие всевозможным сочетаниям уровней этих факторов. Например, если факторов было два, один из которых имел m, а второй r уровней, то мы как бы образуем новый фактор, описывающий взаимодействие двух факторов и имеющий mr уровней. После очевидного (если это необходимо) укрупнения клеток вновь прибегаем к формулам (7.4). При этом, конечно же, в роли фактора A выступает взаимодействие факторов AB (или ABC, $ADEY$, ...).

Глава 8

Общая проблема классификации и снижения размерности

8.1 Сущность задачи снижения размерности

Пусть при изучении n объектов у каждого из них измеряется большое количество p показателей. Если число p достаточно велико, то с ростом n возникает ряд проблем: объем информации очень велик, а нужно ли хранить ее всю? И как наглядно представить себе весь этот объем информации, чтобы извлечь из нее некую суть, необходимую для принятия решения? Тут-то и появляется желание уменьшить число p без нарушения существенной (для рассматриваемой задачи) структуры данных. Конечно, чтобы решить сформулированную задачу математически строго, необходимо многим употребленным сейчас нами понятиям придать точный смысл. Именно этим мы сейчас и займемся. Дальнейшее изложение этой главы следует, в основном, [2].

Итак, цели, которые может ставить перед собой задача снижения размерности, можно разбить на следующие группы:

1. большая наглядность полученных данных, возможность построения графиков и диаграмм в пространствах небольшого количества измерений;

2. лаконизм, обозримость и простота зависимостей после построения

математической модели, за счет участия в ней меньшего количества переменных;

3. резкое снижение объемов хранимой информации.

Конечно же, закрыть список на этом нельзя, но чаще всего решаются именно перечисленные задачи. Какими способами можно сократить размерность p задачи? Очевидно, за счет выбора значительно меньшего числа новых показателей q. Это могут быть как некоторые из уже имевшихся ранее показателей (какие-то из p можно в силу тех или иных причин удалить из рассмотрения без значительного ущерба) или новые показатели могут образовываться как комбинации старых. Возможны разные варианты требований к новым показателям, так или иначе обеспечивающих оптимальность их выбора, например:

- сохранение (в разных смыслах) наибольшей возможной доли информации , имевшейся в исходной выборке;
- взаимная независимость новых показателей (или, по крайней мере, их некоррелированность), что обеспечивает невозможность сокращения их количества без существенной потери информативности;
- наименьшее возможное искажение геометрической структуры данных при переходе от изображения исходной выборки облаком точек в p-мерном пространстве к такому же изображению в q-мерном пространстве ...

(список можно продолжать). В зависимости от выбранного варианта может быть построен критерий оптимальности – некоторая числовая характеристика качества снижения размерности – и поставлена математическая задача на оптимальное снижение размерности. Эти критерии оптимальности бывают внутренние, т.е. определяющиеся структурой исходных данных и строящиеся только по этим данным, и внешние, которые для своего построения привлекают соображения, лежащие за пределами самих наблюдений.

Имеются следующие основные типы предпосылок к тому, что задачу снижения размерности удастся эффективно решить:

1. сильная связь между исходными показателями, в результате которой информация, содержащаяся в них, дублируется;

2. слабая информативность некоторых показателей, которые состоят в основном из случайных помех. Это чаще всего выражается в том, что они мало изменяются при переходе от объекта к объекту, что позволяет исключить их из наших данных, даже повысив при этом количество полезной информации в оставшихся данных;

3. возможность объединения нескольких показателей в один, что бывает возможно, если на самом деле интересующее нас решение связано не с каждым показателем в отдельности, а с некоторым интегративным показателем.

Отметим, наконец, что задачу классификации, т.е. объединения данных в некоторые группы, с которой мы уже имели дело раньше под названием "задача группировки", можно рассматривать, как частный случай задачи снижения размерности. Дело в том, что индивидуальное (большое) разнообразие данных после успешного решения задачи классификации переходит в групповое разнообразие с некоторыми усредненными (одинаковыми) показателями данных в пределах одного класса. В роли p здесь выступает объем выборки, в роли $q \ll p$ – число классов (групп) после завершения классификации.

Перейдем к формальной постановке общей задачи с рассмотрением некоторых частных случаев.

8.2 Формальная постановка задачи снижения размерности. Частные случаи

Пусть $x^{(1)}, ..., x^{(p)}$ – наблюдаемые у каждого из n объектов показатели (случайные величины), $X = (x^{(1)}, ..., x^{(p)})$ – p-мерный вектор, $Z = Z(X)$ – q-мерная векторная функция,

$$Z(X) = (Z^{(1)}(X), ..., Z^{(q)}(X)),$$

Имеется функция $K_q(Z(X))$, принимающая неотрицательные значения – мера информативности или критерий оптимальности. Этот критерий определяется сущностью решаемой задачи. Задан также класс $\mathcal{F}$, которому должна принадлежать функция Z. Построить такую функцию $\hat{Z}$ из класса $\mathcal{F}$, такую, что

$$K_q(\hat{Z}(X)) = \max_{Z \in \mathcal{F}} K_q(Z(X)).$$

Тот или иной выбор критерия оптимальности и класса допустимых преобразований приводит к разного рода методам снижения размерностей. Далее следует краткий обзор наиболее важных из них. Подробно эти методы будут рассмотрены в следующих главах.

8.2.1 Метод главных компонент

Будем искать такие q линейных комбинаций исходных показателей, которые объясняют максимально возможную долю изменчивости (суммы дисперсий) p исходных показателей.

Здесь $\mathcal{F}$ – класс линейных преобразований вида

$$Z^{(j)} \;=\; \sum_{k=1}^{p} c_{j,k}(x^{(k)} \;-\; \bar{x}^{(k)}), \quad j = 1, ..., q,$$

причем на коэффициенты накладываются условия нормировки

$$\sum_{k=1}^{p} c_{j,k}^2 \;=\; 1, \quad j = 1, ..., q,$$

$$\sum_{k=1}^{p} c_{i,k}c_{j,k} \;=\; 0, \quad i, j = 1, ..., q,\ i \neq j.$$

Последнее условие, конечно же, означает, что столбцы матрицы коэффициентов C ортогональны и, как векторы, имеют единичные нормы.

Критерием оптимальности здесь будет

$$K_q(Z) \;=\; \frac{\sum_{j=1}^{q} \mathbf{D}Z^{(j)}}{\sum_{j=1}^{p} \mathbf{D}x^{(j)}}.$$

8.2.2 Экстремальная группировка признаков

Поставим задачу разбить исходные показатели на заранее заданное число групп $\Delta(1), ..., \Delta(q)$ и одновременно внутри каждой группы заменить p показателей одним (интегративным) показателем, являющимся линейной комбинацией $x^{(j)},\ j = 1, ..., p$ так, чтобы внутри одной группы показатели были коррелированы сильно, а между группами наблюдалась бы относительно слабая корреляция.

Здесь $\mathcal{F}$ – класс линейных преобразований, переводящих исходные показатели в Z, нормированных условием $\mathbf{D}Z = 1$, а критерий оптимальности зависит еще и от разбиения на группы $\mathcal{D}$:

$$K_q(Z, \mathcal{D}) = \sum_{j=1}^{q} \sum_{k \in \Delta(j)} \rho^2(x^{(k)}, Z^{(j)}),$$

где $\rho(.,.)$ – коэффициент корреляции, $\mathcal{D} = (\Delta(1), ..., \Delta(q))$.

Простейший пример этого метода – группировка выборки по принципу "геометрической близости"с последующей заменой элементов выборки на геометрические центры соответствующих групп (см. соответствующий раздел выше).

8.2.3 Многомерное шкалирование

Предположим, что результатами наблюдения служат не числовые характеристики объектов, а меры их близости между собой. С такой ситуацией мы сталкивались при изучении шкал сравнений и экспертных оценок. Будем считать, что в нашем распоряжении имеются характеристики $\rho_{i,j}$ попарной близости объектов, на которые мы будем смотреть, как на оценки теоретических расстояний между ними $d_{i,j}$ в p-мерном пространстве исходных показателей (слово "расстояние"здесь может пониматься в самых разных смыслах). Задача будет состоять в наглядном представлении результатов в пространстве относительно небольшого (обычно не более трех) числа измерений с наименьшим возможным искажением геометрической структуры исходных данных.

Таким образом, класс допустимых преобразований $\mathcal{F}$ здесь представляет собой всевозможные размещения образов исходных выборочных точек в q-мерном пространстве. Определим

$$d(Z) = \sum_{i=1}^{n} \sum_{j=1}^{n} \frac{|\hat{d}_{i,j}(Z) - \rho_{i,j}|^{\alpha}}{\rho_{i,j}^{\beta}},$$

где $\hat{d}_{i,j}$ – расстояние между образами объектов в пространстве малой размерности, а числа α, β подбираются исследователем, исходя из содержания задачи (можно взять, например, $\alpha = \beta = 1$). Для завершения формальной постановки задачи многомерного шкалирования осталось ввести критерий оптимальности

$$K_q(Z) = (1 + d(Z))^{-1}.$$

8.2.4 Отбор показателей для дискриминантного анализа

В задаче дискриминантного анализа (классификация с обучением) требуется отнести изучаемые объекты к одной из групп, которые априори заданы некоторым количеством своих представителей. Такими группами могут быть, например фирмы с устойчивой репутацией против фирм с сомнительной репутацией, прогрессивные регионы против регионов депрессивных, или, например, врачебные диагнозы, т.е. количество классов не обязательно равно двум. В качестве образцовых представителей групп, на которые производится разбиение, берутся такие объекты, про которые заведомо известно, к какой из групп они относятся (диагноз уже установлен, например, путем вскрытия).

Задача сокращения размерности здесь возникает в следующем контексте. У нас имеется большое количество данных о каждом из объектов, поступающих на наше изучение и не все они имеют одинаковое значение для успешного решения задачи дискриминации. Например, цвет пижамы, которую предпочитает носить бухгалтер фирмы, вряд ли важен для признания этой фирмы финансово благонадежной. Поэтому некоторые из показателей стоит просто отбросить, при этом оставшиеся показатели окажутся существенно более информативными в рамках решаемых задач.

Образцом для построения критерия оптимальности здесь будут те типичные представители классов, на которые мы ориентируемся. Таким образом, здесь мы впервые сталкиваемся с внешним критерием оптимальности, поскольку эти представители не относятся к выборочным данным.

Как видим, класс допустимых преобразований $\mathcal{F}$ состоит из всевозможных выборов подсистем системы исходных показателей и не допускает их комбинаций:

$$Z(X) \;=\; (x^{(i_1)}, ..., x^{(i_q)}). \tag{8.1}$$

Критерий оптимальности строится из условия максимизации различий распределений отобранных показателей в различных группах, рассчитанных на их типичных представителях (их называют обучающей выборкой):

$$I_q(Z) \;=\; \sum_{i=1}^{r} \sum_{j=1}^{r} \delta(\mathbf{P}_i(Z), \mathbf{P}_j(Z)),$$

где r – число групп, $\mathbf{P}_i(Z)$ – распределение подвектора, определяемого преобразованием Z на типичном представителе i-й группы, $\delta(.,.)$ – какое-нибудь расстояние между распределениями, определяющееся спецификой решаемой задачи.

В простейшем случае, когда классификация производится по величине средних, можно взять

$$\delta(\mathbf{P}_i(Z), \mathbf{P}_j(Z)) \;=\; \sum_{k=1}^{q} \left(\bar{x}^{(i_k)}(i) \;-\; \bar{x}^{(i_k)}(j)\right)^2,$$

где $\bar{x}^{(i_k)}(i)$ – среднее соответствующего показателя, вычисленное по всем типичным представителям i-й группы, причем номера показателей, входящих в последнюю формулу, определяются преобразованием Z (см. (8.1)).

8.2.5 Отбор показателей в модели регрессии

Здесь могут быть сказаны практически те же самые слова, что в предыдущем подразделе, только речь идет об уменьшении количества случайных величин, от которых изучается зависимость Y. Заметим, что критерием здесь служит степень коррелированности с выходом Y, а значит критерий оптимальности вновь можно считать внешним.

Глава 9

Дискриминантный анализ

В этой главе мы рассмотрим задачу дискриминантного анализа – задачу классификации выборочных значений на два или более класса, заданные своими типичными представителями, набор которых называется обучающей выборкой. В этом ключе задачу дискриминантного анализа иногда называют задачей классификации с учителем.

9.1 Постановка задачи. Применимость линейной модели

Мы рассматриваем задачу классификации. На практике такие задачи встречаются весьма часто, хотя решающий ее иногда и сам не подозревает, что он этим занимается. Владельцу банка, например, приходится разбивать клиентов, обращающихся к нему на категории "надежный"и "ненадежный"клиент, пользуясь при этом разного рода данными его о доходах и расходах, сотрудники кадровой службы на основе результатов некоторых тестов должны принять решение о профессиональной пригодности или непригодности претендента, наконец, налоговые службы должны по каждой конкретной налоговой декларации принять решение о ее детальной проверке (или частичной проверке, или не проверять вообще). На последнем примере отчетливо видно, как за счет грамотно принятого решения можно добиться экономии финансов в государственном масштабе, а порой и получить солидную прибавку к средствам.

Рассмотрим для начала наиболее часто встречающийся случай двух классов. Заданы наборы значений случайных векторов – $X_1^{(1)}, ..., X_{n_1}^{(1)}$ и

$X_1^{(2)}, ..., X_{n_2}^{(2)}$, про которые заведомо известно, что первый набор представляет собой векторы параметров объектов, относящихся к первому классу, а второй – ко второму. Эти два набора образуют обучающую выборку объема $n = n_1 + n_2$, по которой мы будем строить в дальнейшем алгоритм классификации, называемый прогностическим правилом. Это название связано с тем, что обычно считают, что факт принадлежности вновь поступающего в наше рассмотрение объекта одному из классов будет достоверно установлен лишь впоследствии, а наше правило представляет лишь некий прогноз дальнейшего развития событий.

Как правило, предполагается, что оба рассматриваемых класса – это совокупности значений случайных векторов, имеющих распределения, принадлежащие одному и тому же параметрическому семейству распределений $\mathbf{P}_\theta,\ \theta \in \Theta$, но с разными, вообще говоря, неизвестными заранее значениями параметра θ_1, θ_2. Если это так и известен аналитический вид $p(x, \theta)$ плотностей распределений $\mathbf{P}_\theta,\ \theta \in \Theta$, то можно предложить следующее правило, называемое подстановочным алгоритмом с независимой оценкой параметров: оцениваем по обучающей выборке значения параметров θ_1, θ_2 (например, методом максимального правдоподобия) и организуем для очередного наблюдения X следующую проверку: относим X к первому классу, если

$$\gamma(X) \ = \ \frac{p(X, \hat{\theta}_2)}{p(X, \hat{\theta}_1)} \ < \ c, \tag{9.1}$$

где константа подбирается так, чтобы ошибка классификации имела не слишком большую вероятность. В основе этого алгоритма лежат те же стандартные соображения, по которым на отношении правдоподобия строятся наиболее эффективные критерии проверки гипотез.

На практике чаще всего пытаются найти прогностическое правило, которое имеет вид некоторого линейного неравенства. Как мы увидим в следующем разделе, в важнейшем частном случае, когда оба класса описываются нормальными распределениями, подстановочный алгоритм дает именно такое правило, а пока обратимся к геометрическим соображениям.

Известно, что в многомерном пространстве линейное уравнение задает гиперплоскость, а линейное неравенство – одно из двух полупространств, на которые эта гиперплоскость разбивает все пространство. Таким образом, если элементы обучающей выборки имеют вид векторов

размерности p, и мы будем представлять их себе как точки в p-мерном пространстве, то поиск наилучшего линейного прогностического правила будет представлять собой поиск гиперплоскости, которая наилучшим образом осуществляет разделение двух множеств точек. Совершенно понятно, что гиперплоскость, которая оставляет всех типичных представителей первого класса с одной стороны, а второго класса с другой, существует отнюдь не всегда, поэтому прежде, чем заниматься поисками нашего правила, следует убедиться в том, существует ли такое правило с не слишком грубыми ошибками. В этом случае говорят о применимости линейной модели.

Проведем предварительный анализ данных. Это один из наиболее ответственных и наименее формализуемых моментов применения аппарата дискриминантного анализа. Помочь нам может то, что исследование методами дискриминантного анализа редко производится изолированно, поэтому исследователь имеет возможность использовать предыдущий опыт и методики сходных работ. Опишем здесь графический тест применимости модели Фишера – классы определяются двумя многомерными нормальными распределениями с общей ковариационной матрицей.

Для лучшего понимания базовой идеи теста следует осознать, что как и при обычной проверке гипотез в математической статистике, один из классов находится в более привилегированном положении – представители его встречаются чаще, соответствуют некоему стабильному состоянию, и вообще, мы больше склонны верить в то, что наблюдаемый X попадет именно в этот класс. К тому же, этот класс (пусть второй) почти наверняка описывается нормальным распределением. Элементы такого класса называют не-случаями. Название легко понять, если мы рассмотрим медицинскую интерпретацию нашей задачи – к не-случаям относятся люди, признанные здоровыми. Сразу понятно, что второй класс ("случаи") отличается гораздо большим разнообразием проявлений, но его представители встречаются относительно реже.

Приняв предложенную терминологию, нормализуем p-мерные векторы случаев оценками среднего и ковариационной матрицы не-случаев, т.е. рассмотрим

$$X^{(1)}_{i,nr} \;=\; B_2^{-1/2}(X^{(1)}_i - \bar{X}^{(2)}).$$

Нетрудно понять, что (двумерная) плоскость, сумма квадратов расстояний $X^{(1)}_{i,nr}, i = 1, ..., n_1$ до которой минимальна, натянута на два собствен-

ных вектора, соответствующих двум наибольшим собственным числам корреляционной матрицы V не-случаев (эта же матрица является ковариационной для $X_{nr}^{(1)}$). Это можно доказать строго, решая задачу на минимум методом наименьших квадратов.

Спроектируем каждый вектор на эту плоскость и рассмотрим полученный рисунок. Если модель Фишера применима, то разброс полученных проекций вокруг центра случаев должен соответствовать двумерному стандартному нормальному распределению.

Еще одним аргументом служит гистограмма распределения расстояний точек, соответствующих случаям, до построенной плоскости. Эта гистограмма должна соответствовать хи-квадрат распределению с p степенями свободы. Для уверенного принятия модели Фишера в качестве основной необходимо принятие положительных решений в обеих описанных процедурах.

После принятия модели Фишера, как и всегда с нормальными распределениями, главную роль начинают играть вектор средних $\vec{a}$ и ковариационная матрица V, поэтому особенную важность приобретает вопрос их оценки по обучающей выборке. И если средние по первому и по второму классу, как и их оценки в принятой модели просто обязаны быть различными и, таким образом, никаких отступлений от традиционных методов оценки средних нет необходимости рассматривать, то с ковариационной матрицей все гораздо сложнее.

Мы приняли предположение о том, что ковариационная матрица в обоих классах одна и та же, поэтому оценки ее обычными методами в первом и втором классе не должны сильно отличаться. Таким образом, не будут сильно отличаться и матрицы, полученные следующими методами:

1. в качестве оценки V выбирается оценка ее по одному из классов;

2. оцениваем ковариационную матрицу по первому классу, затем по второму и в качестве окончательной оценки выбираем их среднее арифметическое;

3. оцениваем V по всей обучающей выборке, игнорируя разбиение на классы.

Разные авторы используют разные методы. Мы возьмем на вооружение комбинацию двух последних, т.е., имея ввиду несмещенность, оценим V

по формуле

$$V \;=\; \frac{1}{n_1+n_2-2}\left(\hat{X}^{(1)t}\hat{X}^{(1)}+\hat{X}^{(2)t}\hat{X}^{(2)}\right). \tag{9.2}$$

Здесь через $\hat{X}^{(i)},\ i=1,2$ обозначены матрицы, у которых по строкам расположены координаты центрированных векторов

$$X_j^{(i)}-\bar{X}^{(i)},\ j=1,...,n_i,\ i=1,2.$$

Так, матрица $\hat{X}^{(1)}$ имеет n_1 строк и p столбцов, а матрица $\hat{X}^{(2)}$ – n_2 строк и p столбцов. В результате, конечно же, V будет квадратной порядка p. Условимся далее до конца этой главы считать, что та оценка ковариационной матрицы, которой мы пользуемся, невырождена.

9.2 Линейное прогностическое правило

Данный раздел посвящен выводу формулы для оптимального линейного правила в предположении, что оно существует. Сначала остановимся на модели Фишера. Напомним, что это означает, что оба класса описываются многомерными нормальными распределениями с одной и той же ковариационной матрицей.

Теорема 15 *В модели Фишера подстановочный алгоритм приводит к линейному прогностическому правилу.*

Доказательство теоремы почти очевидно. Обозначая V оценку общей ковариационной матрицы, а через $\vec{m}_1, \vec{m}_2$ – оценки средних соответствующих распределений, получим

$$p(X, V, \vec{m}_2)/p(X, V, \vec{m}_1) \; = \\ = \; \exp\left\{\tfrac{1}{2}(V^{-1}(X - \vec{m}_1) \cdot (X - \vec{m}_1) - V^{-1}(X - \vec{m}_2) \cdot (X - \vec{m}_2))\right\} \quad ,$$

и, после логарифмирования и раскрытия скобок, правило подстановочного алгоритма имеет вид

$$V^{-1}X \cdot X \; - \; 2V^{-1}X \cdot \vec{m}_1 \; - \; V^{-1}X \cdot X \; + \; 2V^{-1}X \cdot \vec{m}_2 \; < \; C,$$

или

$$V^{-1}X \cdot (\vec{m}_2 - \vec{m}_1) \; < \; C',$$

что очевидно имеет линейный характер.

Теперь просто будем предполагать, что мы хотим выписать формулу именно линейного прогностического правила, т.е. априори считать, что оптимальное правило линейно.

Введем в рассмотрение вектор $\vec{\beta}^t \; = \; (\beta_1, ..., \beta_q)$ неизвестных коэффициентов линейной прогностической функции, так что эта функция имеет вид

$$Y(X) \; = \; \vec{\beta} \cdot X \; + \; \beta_0,$$

а само правило относит X к первому классу, если $Y(X) > 0$, иначе ко второму. Пусть $\bar{Y}^{(1)}$ – среднее значение функции по всем X из обучающей выборки, отнесенным нашим правилом к первому классу, $\bar{Y}^{(2)}$ – ко второму (групповые средние). Величина

$$V(\vec{\beta}) \; = \; \left(\bar{Y}^{(1)} \; - \; \bar{Y}^{(2)}\right)^2$$

называется межгрупповой вариацией разделения. Чем она больше, тем дальше "разводит"средние предлагаемое правило. Но одного сильного различия групповых средних мало, хотелось бы еще, чтобы отдельные точки плотно группировались вокруг своего группового среднего, т,е. относительно небольшой была бы суммарная величина внутригрупповых вариаций

$$v(\vec{\beta}) \;=\; \sum_{j=1}^{2}\sum_{i=1}^{m_j}\left(Y(X_i^{Y,(j)}) - \bar{Y}^{(j)}\right)^2.$$

Здесь мы предполагаем, что наша функция Y произвела разбиение исходной обучающей выборки на $X_1^{Y,(1)}, ..., X_{m_1}^{Y,(1)}$ и $X_1^{Y,(2)},\ ...,\ X_{m_2}^{Y,(2)}$, $m_1 + m_2 = n$, которое не обязано совпадать с первоначальным, что приводит к ошибкам дискриминации. Этого избежать невозможно, но нас приободряет допускаемое существование оптимального линейного правила, к которому мы и придем.

На практике, к сожалению, увеличение межгрупповой вариации приводит к увеличению внутригрупповых вариаций, поэтому в качестве критерия оптимальности прогностического правила выберем величину

$$F(\vec{\beta}) \;=\; \frac{V(\vec{\beta})}{v(\vec{\beta})}.$$

Обозначим $\vec{M}1, \vec{M}2$ средние величины элементов обучающей выборки, отнесенных в новой классификации к первому, второму классу:

$$\vec{M}i \;=\; \frac{1}{m_i}\sum_{j=1}^{m_i} X_j^{Y,(i)}, \qquad i = 1, 2$$

и введем центрированные значения

$$\hat{X}_j^{(i)} \;=\; X_j^{Y,(i)} \;- \vec{M}i,\ j = 1, ..., m_i,\ i = 1, 2.$$

Тогда несложные выкладки показывают, что

$$\begin{aligned} v(\vec{\beta}) \;&=\; \textstyle\sum_{j=1}^{m_1}\left(\vec{\beta}\cdot\hat{X}_j^{(1)}\right)^2 \;+\; \sum_{j=1}^{m_2}\left(\vec{\beta}\cdot\hat{X}_j^{(2)}\right)^2 \qquad = \\ =\;& \textstyle\sum_{i=1}^{m_1}\sum_{j=1}^{m_1}\beta_i\beta_j\hat{X}_i^{(1)}\hat{X}_j^{(1)} \;+\; \sum_{i=1}^{m_2}\sum_{j=1}^{m_2}\beta_i\beta_j\hat{X}_i^{(2)}\hat{X}_j^{(2)} \;= \\ =\;& \vec{\beta}^t\left(\hat{X}^{(1)t}\hat{X}^{(1)} + \hat{X}^{(2)t}\hat{X}^{(2)}\right)\vec{\beta}. \end{aligned}$$

Припомнив теперь оценку (9.2) ковариационной матрицы, и сделанные выше предположения о практическом совпадении разных оценок этой матрицы, видим, что можно положить

$$v(\vec{\beta}) \ = \ (n_1 + n_2 - 2)V\vec{\beta} \cdot \vec{\beta}$$

(здесь мы, конечно же, учли, что $m_1 + m_2 \ = \ n_1 + n_2$.).

Оценим межгрупповую вариацию.

$$V(\vec{\beta}) \ = \ (\vec{\beta} \cdot \bar{X}^{(1)} - \vec{\beta} \cdot \bar{X}^{(2)})^2 \ = \ \vec{\beta}^t(\bar{X}^{(1)} - \bar{X}^{(2)}))(\bar{X}^{(1)} - \bar{X}^{(2)})^t\vec{\beta}.$$

Обозначая

$$\vec{a} \ = \ \bar{X}^{(1)} - \bar{X}^{(2)}, \quad L \ = \ \vec{a}\vec{a}^t,$$

получим, что критерий оптимальности запишется в виде

$$F(\vec{\beta}) \ = \ \frac{L\vec{\beta} \cdot \vec{\beta}}{V\vec{\beta} \cdot \vec{\beta}}.$$

Исследуем эту функцию на максимум методами математического анализа. Нам понадобится следующее утверждение.

Лемма 7 *Если матрица B симметрична, $f(\vec{\beta}) \ = \ B\vec{\beta} \cdot \vec{\beta}$, то вектор f'_β, координатами которого являются частные производные $\frac{\partial f}{\partial \beta_j}$, $j = 1, ..., p$, может быть записан в виде*

$$f'_\beta \ = \ 2B\vec{\beta}.$$

Доказательство. Запишем $f(\vec{\beta}) = \sum_{i,j} B_{i,j}\beta_i\beta_j$, откуда

$$\frac{\partial f}{\partial \beta_j} \ = \ 2B_{j,j}\beta_j + \sum_{i \neq j}(B_{i,j} + B_{j,i})\beta_i \ = \ 2(B\vec{\beta})_j$$

для произвольного j, что и завершает доказательство.

Заметим, что

$$L\vec{\beta} \cdot \vec{\beta} \ = \ (\vec{a} \cdot \vec{\beta})^2. \tag{9.3}$$

В силу леммы

$$F'_\beta \ = \ \frac{2L\vec{\beta}(V\vec{\beta} \cdot \vec{\beta}) - 2V\vec{\beta}(L\vec{\beta} \cdot \vec{\beta})}{(V\vec{\beta} \cdot \vec{\beta})^2}. \tag{9.4}$$

Пусть эта производная обращается в $\vec{0}$. Тогда из (9.3) имеем

$$\vec{a}(\vec{a}\cdot\vec{\beta})(V\vec{\beta}\cdot\vec{\beta}) \;=\; V\vec{\beta}(\vec{a}\cdot\vec{\beta})^2.$$

Следовательно, либо $\vec{a}\cdot\vec{\beta}=0$, но тогда $F(\vec{\beta})=0$, и мы имеем дело с минимумом нашего критерия, либо

$$V\vec{\beta} \;=\; b\vec{a} \tag{9.5}$$

для некоторой константы b. Таким образом, поиск коэффициентов $\vec{\beta}$, для которых критерий F принимает максимальное значение, может проводится лишь в множестве параметров, удовлетворяющих (9.5) для какого-нибудь b.

Отсюда получаем

$$\vec{\beta} \;=\; bV^{-1}(\bar{X}^{(1)}-\bar{X}^{(2)}).$$

Учитывая вид критерия оптимальности, видим, что при любом выборе b значение этого критерия остается постоянным, а значит, мы можем взять $b=1$ (иногда, впрочем, выбирают b так, чтобы $\beta_1=1$).

Теорема 16 *Если оптимальное прогностическое правило линейно, то оно имеет вид*

$$Y(X) \;=\; V^{-1}(\bar{X}^{(1)}-\bar{X}^{(2)})\cdot(X-\frac{1}{2}(\bar{X}^{(1)}+\bar{X}^{(2)})),$$

причем наблюдение X относится к первому классу тогда и только тогда, когда $Y(X)>0$.

Смысл утверждения этой теоремы понять нетрудно – наблюдение относится к тому из классов, к среднему из которых оно лежит ближе, т.е. "по ту же сторону"от полусуммы средних. Выражение "по ту же сторону"как раз и формализуется в утверждении теоремы.

9.3 Что делать, если модель Фишера не может быть принята?

В любом случае способ построения прогностического правила на основе отношения правдоподобия, описанный выше, работает. Просто в случае, когда распределения, описывающие классы, нормальны, да еще и

обладают одинаковыми ковариационными матрицами, формулы особенно просты и легко доказуемы. В случае, когда никаких предположений о виде плотностей просто нет, а значит и нет возможности выписать отношение правдоподобия точно, можно предложить некоторые способы оценки отношения правдоподобия по выборочным данным.

Один из способов: оценим $\gamma(X)$, заданное в (9.1), так

$$\hat{\gamma}(X) \;=\; \frac{n_1 \sum_{i=1}^{n_1} k(\|X - X_i^{(1)}\|^2/b)}{n_2 \sum_{i=1}^{n_2} k(\|X - X_i^{(2)}\|^2/b)}, \tag{9.6}$$

где b – малый параметр, задаваемый исследователем, а $k(t)$ – произвольная функция, удовлетворяющая условиям

$$k(t) \geq 0,\ k(t) = k(-t),\ \int\limits_{-\infty}^{\infty} k(t)dt = 1.$$

Чаще всего в качестве этой функции берут плотность стандартного нормального распределения. Дальнейшие рассуждения стандартны. Поскольку мы собираемся создать прогностическое правило на основе $\hat{\gamma}(X)$, то для всех элементов обучающей выборки, принадлежавших первому классу, вычислим $\hat{\gamma}(X)$, и то же проделаем со всеми элементами второго класса. Мы получим n действительных чисел. Если все эти числа, соответствующие элементам разных классов, лежат по разные стороны от некоего порогового значения c, то прогностическое правило будет иметь следующий вид: если $\hat{\gamma}(X)$ лежит по ту же сторону от c, что и $\hat{\gamma}(X_1^{(1)})$, то X относим к первому классу, иначе ко второму.

Как и все остальные алгоритмы дискриминации, этот можно сделать самообучающимся: при наборе дополнительной статистики двигать число c в ту или другую сторону.

Если же множества значений $\hat{\gamma}(X)$ по разным классам пересекаются, то можно провести разделение таким же образом, как уже было описано, с некоторым количеством ошибочно классифицирующихся элементов. В этом случае самообучение алгоритма не только желательно, но и обязательно, поскольку лучшей классификации, чем основанной на отношении правдоподобия, построить, вероятнее всего, нельзя,

Известно также очень простое эмпирическое прогностическое правило, которое носит название "метод k ближайших соседей". Исходя из содержания решаемой задачи, исследователь задает расстояние между

выборочными p-мерными точками (например, обычное евклидово расстояние) и выбирает, как правило, нечетное число k. После этого для поступающей на изучение точки X ищется k ближайших соседей из обучающей выборки с точки зрения выбранного расстояния. Если большинство из этих соседей окажутся элементами первого класса, то мы отнесем X к первому классу, иначе ко второму.

Отметим также следующее. Очевидно, что включение в прогностическое правило малоинформативных переменных не только усложняет вычисления, но может заметно ухудшить качество классификации. Действительно, каждый малоинформативный показатель несет в себе достаточно большую долю "шума", и включение его в алгоритм приводит к серьезному ухудшению отношения сигнал/шум. Но, к сожалению, если объем обучающей выборки небольшой, то легко принять информативный показатель за малоинформативный, отбросить его при первом же подходе к задаче, и тем самым, сильно ухудшить последующую обучаемость алгоритма. Вывод: при малых объемах обучающей выборки показателей лучше не отбрасывать.

Мы видим, что дискриминация изучаемых объектов производится, в конце концов, по одному числовому параметру, который в особенно важном частном случае линейного прогностического правила, получается линейной комбинацией координат, что можно воспринимать, как переход к новым координатам, одна из которых и служит дискриминационной функцией. Таким образом, поиск линейной дискриминационной функции можно представлять себе, как переход (при помощи поворота, поскольку постоянный множитель не изменяет качества дискриминации) к новой системе параметров, характеризующих наши объекты, и выбор в качестве соответствующего правила первого из полученных параметров. Эти соображения позволяют слегка по иному подойти к задаче дискриминации, и вначале попробовать слегка "повращать"векторы параметров с целью добиться большего разделения классов по новым координатам.

9.4 Один пример

Рассмотрим условный пример, иллюстрирующий решение задачи дискриминантного анализа. В качестве обучающей выборки возьмем по 5 типичных рок-групп, играющих музыку в стиле рэп (первый класс) и в стиле эйсид-хаус (второй класс). Для каждой из них обозначим через

x_1 количество концертных выступлений за последние 2 года, а через x_2 показатель финансовой успешности проекта, рассчитанный в тыс. условных единиц на участника группы. Данные обучающей выборки собраны в таблице:

Обучающая выборка в музыкальном примере

Первый класс			Второй класс		
N	x_1	x_2	N	x_1	x_2
1	97	94	1	39	38
2	58	39	2	32	64
3	73	45	3	44	53
4	70	80	4	36	33
5	96	46	5	60	78
средние	78,8	60,8		42,2	53,2

Подготовимся к оценке ковариационной матрицы по формуле (9.2):

$$\hat{X}^{(1)t}\hat{X}^{(1)} = \begin{pmatrix} 1170,8 & 725,8 \\ 725,8 & 2414,8 \end{pmatrix}, \; \hat{X}^{(2)t}\hat{X}^{(2)} = \begin{pmatrix} 472,8 & 725,1 \\ 725,1 & 1370,8 \end{pmatrix}.$$

Тогда

$$V = \begin{pmatrix} 205,4 & 181,4 \\ 181,4 & 473,2 \end{pmatrix}, \; V^{-1} = \begin{pmatrix} 0,0074 & -0,0028 \\ -0,0028 & 0,0032 \end{pmatrix}.$$

Выпишем также

$$\frac{1}{2}(\bar{X}^{(1)} + \bar{X}^{(2)})^t = (60,5; 57), \; (\bar{X}^{(1)} - \bar{X}^{(2)})^t = (36,6; 7,6).$$

Таким образом, линейное прогностическое правило имеет вид

$$Y = 0,248x_1 - 0,079x_2 - 10,5,$$

что приводит к следующим значениям:

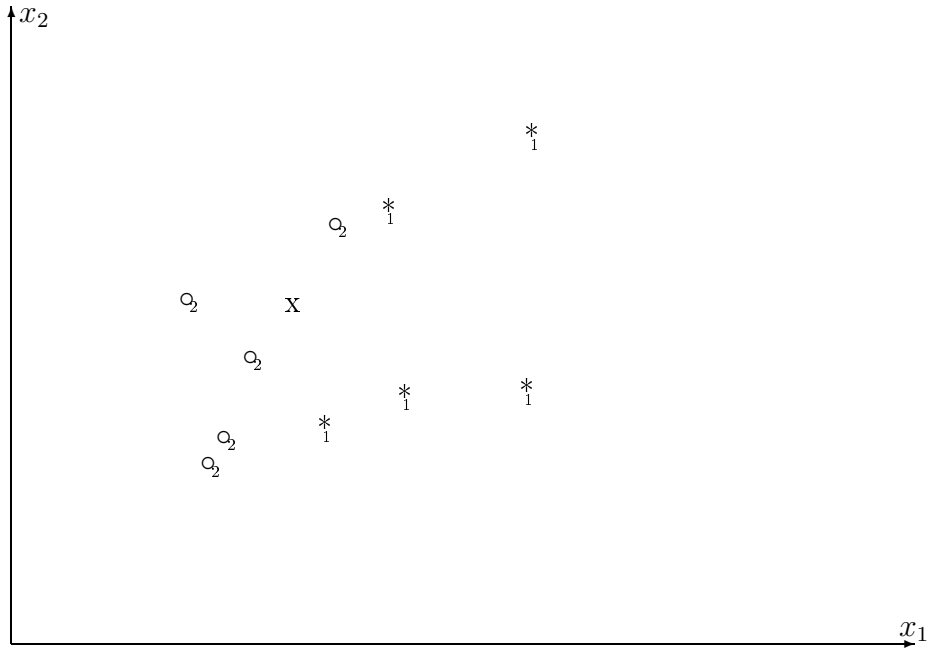
Значения линейного прогностического правила

Первый класс					Второй класс				
1	2	3	4	5	1	2	3	4	5
6,13	0,80	4,05	0,54	9,67	-3,83	-7,62	-3,77	-4,18	-1.78

Это вполне соответствует хорошей степени дискриминации. Возьмем теперь новый музыкальный коллектив с характеристиками $x_1 = 52$, $x_2 = 63$. Построенное правило дает $Y = -2,57$, а значит этот коллектив следует отнести ко второму классу.

Для сравнения произведем дискриминацию методом ближайших соседей, для чего изобразим точки, соответствующие обучающей выборке, на плоскости. Здесь единицами помечены точки первого класса, двойками – второго класса, а новый коллектив помечен символом x. Отчетливо видно, что три ближайших соседа нового коллектива принадлежат второму классу, а значит и его тоже следует отнести ко второму классу (скорее всего, ребята играют эйсид-хаус).

Метод ближайших соседей



Наконец, приведем значения, полученные при вычислении оценки отношения правдоподобия методом (9.6) при различных значениях b:

Различные дискриминационные функции
при оценке отношения правдоподобия

b	класс	значения $\hat{\gamma}$					интервал	x
1	1	1,00	1,00	1,00	1,00	1,00	1	0
	2	0,00	0,00	0,00	0,00	0,00	0	
10	1	0,99	0,77	0,99	0,63	0,99	(0,63;0,99)	0,12
	2	0,07	0,00	0,08	0,04	0,37	(0,00;0,37)	
15	1	0,98	0,56	0,86	0,59	0,99	(0,56;0,99)	0,27
	2	0,16	0,04	0,18	0,13	0,42	(0,04;0,42)	
50	1	0,63	0,47	0,52	0,53	0,60	(0,47;0,63)	0,46
	2	0,41	0,39	0,43	0,40	0,49	(0,39;0,49)	

В предпоследней колонке указаны границы, в которых изменяется $\hat{\gamma}(X)$ при изменении X в рамках представителей соответствующего класса. Мы видим, что с ростом b соответствующие интервалы начинают

сближаться и, в конце концов, становятся пересекающимися. Последняя колонка - значение оценки отношения правдоподобия на новом коллективе (он тот же, что и раньше). Видно, что это значение всегда находится в рамках интервала для второго класса, поэтому вопроса классификации его не возникает. Конечно же, b следует выбирать так, чтобы пересечения интервалов не возникало (по возможности).

9.5 Число классов, не меньшее трех

В случае, когда классов более двух, задача отнесения нового объекта к одному из этих классов усложняется. Для начала упомянем две следующие очевидные процедуры:

1. Будем применять описанную выше процедуру (любую из них) по отношению к произвольно выбранной паре классов, пусть, например, для первого и второго класса. Если новый объект классифицируется как объект второго класса, первый из дальнейшего рассмотрения исключаем. Завершение процесса обеспечивается уменьшением числа классов, остающихся в классификации. Можно также рассмотреть видоизменение этой процедуры, не зависящее от порядка рассмотрения пар – рассматриваем все пары и запоминаем на каждом шаге номер класса, к которому объект причисляется. Затем относим его к тому классу, который встречался чаще всего.

2. Для каждой пары из множества рассматриваемых классов построим линейное прогностическое правило. Оно задает в p-мерном пространстве гиперплоскость. После построения всех этих гиперплоскостей облако точек обучающей выборки, определяющее каждый из классов, окажется внутри некоторого p-мерного многогранника, гранями которого служат части построенных гиперплоскостей. Теперь мы отнесем новый объект к тому классу, в многогранник которого он попал.

Перейдем к более формальной постановке задачи. Пусть имеется k классов, определяемых плотностями распределений $p(.,j),\ j = 1,...,k$ и у нас имеется обучающая выборка объема $n = n_1 + ... + n_k$, в которой n_j элементов относятся к j-му классу, $j = 1,...,k$. Числа $\pi_j = n_j/n$ называют априорными вероятностями j-го класса. Пусть нам известна функция

$c(j|i)$, задающая цену ошибки при принятии объекта, относившегося к классу i за объект класса j. Очевидно, что

$$c(i|i) = 0, \quad c(i|j) > 0 \text{ при } i \neq j. \tag{9.7}$$

Задача состоит в том, чтобы разбить все пространство наблюдений на непересекающиеся множества $A_1, ..., A_k$, соответствующие принятию решения об отнесении объекта в тот или иной класс так, чтобы минимизировать потери.

Вероятность отнести объект к j-му классу, если он принадлежал на самом деле i-му, равна

$$P(j|i) \ = \ \int_{A_j} p(t,i)dt,$$

а значит, математическое ожидание понесенных потерь

$$Q \ = \ \sum_{i=1}^{k} \pi_i \sum_{j \neq i} c(j|i) P(j|i).$$

При фиксации значения вектора параметров нового объекта X получаем, что апостериорная вероятность отнести его в i-й класс равна

$$\pi_i(X) \ = \ \frac{\pi_i p(X,i)}{\sum_{j=1}^{k} \pi_j p(X,j)},$$

а ожидаемые потери при отнесении его в j-й класс составят

$$\sum_{i \neq j} c(j|i) \frac{\pi_i p(X,i)}{\sum_{s=1}^{k} \pi_s p(X,s)}.$$

Ясно, что эти потери будут минимальны, если будет минимален числитель результирующей дроби. Поэтому определим множество A_j как множество тех наборов показателей X, для которых минимальное значение функции

$$f(s) \ = \ \sum_{i \neq s} \pi_i p(X,i) c(s|i)$$

достигается при $s = j$.

Отметим, что если мы считали, что $c(s|i) = 1$ при всех $s \neq i$, то X будет согласно полученному правилу отнесено к тому из классов, для которого $\pi_j p(X,j)$ наибольшее. Описанное прогностическое правило называют байесовским.

9.6 Проверка качества дискриминации

Распространенной ошибкой неискушенного исследователя, строящего прогностические правило, является то, что в качестве процента верно классифицируемых данных сообщается процент данных обучающей выборки, которые были отклассифицированы правильно. Конечно же, такой процент всегда будет выше, чем процент правильных решений при последующей работе уже готового правила. Поэтому следует предусмотреть проверку – запасти так называемую экзаменующую выборку и по ней провести изучение построенного правила. Если объем первоначальной обучающей выборки велик, то это сделать нетрудно, но обычно бывает жаль тратить значительную часть априорной информации на проверку без использования ее при построении самой процедуры. Поэтому можно использовать "более хитрые"методы, особенно если первоначальный объем выборки невелик.

Во-первых, можно использовать следующий прием: один из элементов обучающей выборки исключается и строится прогностическое правило по оставшимся $n-1$ элементам и удаленный элемент используется в качестве экзаменующего. Затем удаленный элемент возвращается в выборку, удаляется второй элемент и процедура повторяется. Алгоритм заканчивает свою работу, когда удалению подвергнутся все элементы выборки по очереди.

Во-вторых, напомним уже упоминавшийся выше алгоритм В.Эфрона "bootstrap", заключающийся в том, что имеющаяся выборка объявляется генеральной совокупностью и экзаменующая выборка набирается из нее методом случайного отбора. Имеется и ряд других методов.

После расчета значений дискриминационной функции $\gamma(X)$ на элементах экзаменующей выборки, принадлежащих первому классу, обычно рассчитывают их среднюю величину $\bar{g}_{(1)}$. Пусть теперь $\bar{\gamma}^{(1)}$ – средняя величина значений дискриминационной функции на элементах первого класса первоначальной обучающей выборки. Построим теперь новую (исправленную) функцию

$$\gamma'(X) \;=\; \gamma(X) \;+\; (\bar{\gamma}^{(1)} - \bar{g}_{(1)}).$$

Ясно, что этот процесс при появлении новых экзаменующих выборок можно повторять многократно (и не обязательно с элементами первого класса). Тем самым мы обеспечим "самообучаемость"алгоритма дискриминации.

Настало, наконец, время сказать, что хотя объективных показателей качества проведенной дискриминации в практике используется много, на наш взгляд, одной из самых наглядных служит вероятность ошибочной классификации. Если мы припомним обозначения, введенные в предыдущем разделе и дополнительно обозначим через $m_j,\ j = 1, ..., k$ число элементов обучающей (или экзаменующей) выборки, относившихся к j-му классу, содержавшей m элементов и ошибочно отнесенных к какому-либо другому классу (кроме j-го), то вероятность ошибочной классификации q вычисляется по формуле

$$q \ = \ \sum_{j=1}^{k} \pi_j \frac{m_j}{m}.$$

Если же у нас имеются заданные штрафы за ошибочную классификацию (см. (9.7)), то можно рассмотреть величину средних потерь

$$Q \ = \ \sum_{j=1}^{k} \left(\sum_{i=1}^{k} c(i|j) \right) \pi_j \frac{m_j}{m},$$

которая также является характеристикой качества дискриминации.

По крайней мере беглого упоминания здесь заслуживает тот факт, что типичной ситуацией, в которой работает аппарат дискриминантного анализа, является постоянное увеличение количества изучаемых данных. Что при этом происходит, чтобы алгоритм не терял своих классифицирующих свойств в так называемой "классической"ситуации, когда растет только объем выборки, мы только что рассмотрели. Но иногда по мере увеличения объема выборки n растет также и число изучаемых параметров p. Как организовать работу исследователя в этих условиях? За ответом на этот вопрос отсылаем читателя к [2, с.88-96] и дальнейшим ссылкам на литературу там.

Глава 10

Группировка как метод снижения размерностей

Связь между группировкой и снижением размерности задачи уже обсуждалась в главе 8, а простейшие методы группировки – в 3. Здесь мы остановимся на нескольких методах, которые относятся к группировке не наблюдаемых объектов, как раньше, а тех признаков, которыми эти объекты обладают. Эти методы возможно применять в ситуации, когда задача не имеет четкой структуры, ситуации, которая типична при работе с данными медицины, биологии, экономики или психологии. Отметим прежде всего, что если задача неконкретна, границы методов расплывчаты и т.д., то имеет смысл сначала попытаться уточнить ее и лишь потом решать. Но иногда точное понимание приходит уже в процессе обработки данных, и это одна из причин, по которым описываемые ниже методы используются на практике.

10.1 Метод экстремальной группировки признаков

Пусть с каждым из изучаемых объектов связано p признаков, которые будем считать случайными величинами $X_1, ..., X_p$. Будем также считать, что задано некое натуральное число $q > 1$, $q \ll p$. Ставится задача разбить множество номеров $\{1, ..., p\}$ признаков на непересекающиеся подмножества $\mathcal{D} = \{\Delta(1), ..., \Delta(q)\}$ и построить для каждого из этих подмножеств такую линейную комбинацию признаков с номерами, по-

павшими в это подмножество, что выбранное разбиение и построенные комбинации максимизируют определенный критерий оптимальности.

Если это будет проделано, то весь набор исходных признаков с номерами, попавшими в $\Delta(j)$ заменяется (одной) соответствующей комбинацией, которую мы обозначим f_j, при $j = 1, ..., q$, и размерность задачи сокращается с p до q.

Чаще всего на линейные комбинации f_j накладывают дополнительное условие нормировки $\mathbf{D}f_j = 1,\ j = 1, .., q$, а критерии оптимальности связывают с коэффициентами корреляции признаков внутри выбранной группы $\Delta(j)$ и с формируемым фактором f_j.

Мы рассмотрим два различных варианта критериев оптимальности.

10.1.1 Критерий квадратов

Пусть

$$K_q(F, \mathcal{D}) \ = \sum_{j=1}^{q} \sum_{i \in \Delta(j)} \rho^2(X_i,\ f_j),$$

где через F обозначен набор всех q линейных комбинаций признаков f_j, нормированных условием единичной дисперсии. Это выражение и будет здесь являться критерием оптимальности.

Максимизация выписанного выражения как по $\mathcal{D}$, так и по выбору линейных комбинаций внутри строящихся групп отвечает интуитивному представлению о построении оптимальных групп признаков так, чтобы внутри каждой из групп признаки были коррелированы относительно сильно, а между группами относительно слабо. Действительно, все признаки, попадающие в i-ю группу сильно коррелируют с f_i, а значит, и между собой, в то же время при выборе разбиения $\mathcal{D}$ наилучшим образом очевидно, должно получится, что факторы $f_1, ..., f_q$ максимально различаются друг от друга.

Сначала предположим, что разбиение на группы уже задано. Тогда построения факторов из F следует производить отдельно для каждой группы, и для i-й группы задача будет заключаться в максимизации суммы

$$S_i \ = \sum_{j \in \Delta(i)} \rho^2(X_j,\ f_i)$$

по выбору коэффициентов в линейной комбинации, представляющей собой фактор f_i. Эта и похожие на нее задачи на максимум будут далее

встречаться не один раз, поэтому докажем здесь следующий полезный факт.

Теорема 17 *Пусть* $\mathbf{D}X_i = 1, i = 1, ..., p$, $\vec{\alpha} = (\alpha_1, ..., \alpha_p)^t$, $\vec{X} = (X_1, ..., X_p)^t$. *Тогда максимум функции*

$$S(\vec{\alpha}) = \sum_{i=1}^{p} \rho^2(X_i, \sum_{j=1}^{p} \alpha_j X_j)$$

при условии, что $\sum_{j=1}^{p} \alpha_j^2 = 1$ *равен квадрату максимального собственного числа ковариационной матрицы* $V = \mathbf{cov}\vec{X}$ *и достигается тогда, когда* $\vec{\alpha}$ *равен единичному собственному вектору этой матрицы, отвечающим этому собственному числу.*

Доказательство. Заметим, что наши условия нормировки приводят к следующему виду максимизируемой функции:

$$S(\vec{\alpha}) = \sum_{i=1}^{p} \mathbf{cov}^2(X_i, \sum_{j=1}^{p} \alpha_j X_j) = \sum_{i=1}^{p} \left(\sum_{j=1}^{p} \alpha_j V_{i,j} \right)^2,$$

а значит,

$$S(\vec{\alpha}) = \sum_{i=1}^{p} \sum_{j=1}^{p} \sum_{k=1}^{p} V_{i,j} V_{i,k} \alpha_j \alpha_k = \sum_{j=1}^{p} \sum_{k=1}^{p} \left(\sum_{i=1}^{p} V_{j,i} V_{i,k} \right) \alpha_j \alpha_k$$

(была учтена симметричность матрицы V). Используя определение произведения матриц, видим, что

$$S(\vec{\alpha}) = \sum_{j=1}^{p} \sum_{k=1}^{p} (V^2)_{j,k} \alpha_j \alpha_k = V^2\vec{\alpha} \cdot \vec{\alpha} = \|V\vec{\alpha}\|^2.$$

Вновь была использована симметричность V и свойство, позволяющее матрицу перебрасывать с аргумента на аргумент скалярного произведения, превращая ее в транспонированную.

Но из курса алгебры известно, что умножение вектора на матрицу представляет собой последовательное выполнение поворота (который не меняет длины вектора, в нашем случае равной 1) и растяжения, причем растяжение в чистом виде по определению совершается только вдоль направлений, отвечающих собственным векторам. Итак, величина $\|V\vec{\alpha}\|^2$,

которую мы исследуем на экстремум, может быть выбрана лишь из значений $\lambda_1^2,\ ...,\ \lambda_p^2$ – квадратов собственных чисел V. Поскольку ковариационная матрица V положительно определена, и собственные числа ее положительны, то теорема полностью доказана.

Итак, из доказанной теоремы следует, что при фиксации разбиения на группы оптимальный набор факторов определяется формулами

$$f_j \ = \ \frac{\sum_{i\in\Delta(j)} a_i^{(j)} X_i}{\sqrt{\sum_{i\in\Delta(j)}\sum_{k\in\Delta(j)} r_{i,k} a_i^{(j)} a_k^{(j)}}}, \quad j=1,...,q, \tag{10.1}$$

где $a_i^{(j)},\ i=1,...,p$ – координаты собственного вектора ковариационной матрицы признаков j-й группы, отвечающего максимальному ее собственному числу, $r_{i,k} \ = \ \rho(X_i,\ X_k)$. Появление квадратного корня в знаменателе связано с необходимостью добиться выполнения условия нормировки для $X_i,\ i=1,..,p$ в теореме 17.

Если мы теперь, наоборот, предположим, что факторы $f_j,\ j=1,...,q$ уже построены, то нетрудно построить разбиение $\mathcal{D}$, максимизирующее $K_q(F,\mathcal{D})$: положим

$$\Delta(j) \ = \ \{i \mid \rho^2(X_i,f_j) = \max_{k=1,...,q} \rho^2(X_i,f_k)\} \tag{10.2}$$

Понятно, что соотношения (10.1), (10.2) обязательно выполнены, если только на паре $F,\mathcal{D}$ достигается максимум функционала K_q, т.е. эти условия необходимы для максимума.

Для одновременной оптимизации $K_q(F,\mathcal{D})$ по обоим подбираемым переменным обычно используется следующий итерационный алгоритм, который не уменьшает значения критерия оптимальности на каждом своем шаге, а значит, обязательно закончит свою работу через конечное число шагов.

Шаг 0. Возьмем любое (первоначальное) разбиение $\mathcal{D}_0$. Построим систему факторов F_0 по формулам (10.1). К шагу 1.

Шаг 1. По $\mathcal{D}_{k-1}$ и (10.1) строим факторы $F_k \ = \ (f_1^{(k)},...,f_q^{(k)})$ и новое разбиение $\mathcal{D}_k$ с помощью очевидного видоизменения формул (10.2):

$$\Delta^{(k)}(j) = \{i \mid \rho^2(X_i,f_j^{(k)}) \ = \ \max_{s=1,...,q} \rho^2(X_i,f_s^{(k)})\},\ j=1,...,q.$$

Если при выполнении шага 1 получилось, что $\mathcal{D}_{k-1}=\mathcal{D}_k,\ F_k=F_{k-1}$, то оптимальное разбиение и оптимальная система факторов построена. Иначе шаг 1 повторить.

10.1.2 Критерий модулей

При втором подходе к понятию оптимальности экстремальной группировки признаков критерием служит функционал

$$K_q(F, \mathcal{D}) \;=\; \sum_{j=1}^{q} \sum_{i \in \Delta(j)} |\rho(X_i,\; f_i)|.$$

Из формулы видно, что максимизация этого функционала имеет тот же качественный смысл, что и только что рассмотренная задача. Исследуя этот функционал при помощи методов математического анализа и составляя функцию Лагранжа, исходя из условий

$$\mathbf{D} f_i \;=\; 1, \quad i = 1, ..., q,$$

можно показать, что в точке максимума каждый из факторов должен иметь вид

$$f_j \;=\; \frac{\sum_{i \in \Delta(j)} d_i X_i}{\sqrt{\sum_{i \in \Delta(j)} \sum_{k \in \Delta(j)} r_{i,k} d_i d_k}}, \quad j = 1, ..., q, \tag{10.3}$$

где коэффициенты $d_i,\; i = 1, ..., q$ принимают значения ± 1, и именно при этих значениях коэффициентов на оптимальном разбиении достигает своего максимума финкционал

$$J(d_1, ..., d_q,\; \mathcal{D}) \;=\; \sum_{j=1}^{q} \sqrt{\mathbf{D}\left(\sum_{i \in \Delta(j)} d_i X_i \right)}.$$

Значит, достаточно максимизировать функционал J. При фиксированном разбиении на группы $\mathcal{D}$ этот функционал достигает максимума, если внутри каждой группы коэффициенты $d_1, ..., d_q$ максимизируют величину соответствующей дисперсии, что означает связь этих коэффициентов со знаками коэффициентов корреляции $r_{i,j} \;=\; \rho(X_i, X_j)$. Для того, чтобы при фиксированном разбиении на группы построить коэффициенты оптимальным образом, нужно учесть соотношение

$$\mathbf{D}\left(\sum_{j \in \Delta} d_j X_j \right) \;=\; \sum_{j \in \Delta} d_j^2 \mathbf{D} X_j \;+\; 2 \sum_{j \in \Delta} \sum_{i \in \Delta, i \neq j} d_i d_j r_{i,j} \sqrt{\mathbf{D} X_i \mathbf{D} X_j}$$

и подобрать $d_i,\ i \in \Delta$ так, чтобы наибольшее возможное число знаков произведений $d_i d_j$ совпадало бы со знаком $r_{i,j}$.

Как и выше, используется итерационный алгоритм.

Шаг 0. Зададим произвольно начальное разбиение

$$\mathcal{D}_0 \ = \ \{\Delta_1^{(0)}, ..., \Delta_q^{(0)}\}$$

признаков на q групп и определим набор коэффициентов $\vec{d}_0$ описанным выше образом. Положим $k = 0, m = 1$. К шагу 1.

Шаг 1. Пусть $i = 1$. Запомним имеющиеся к этому моменту разбиение $\mathcal{D}_k$ и набор коэффициентов $\vec{d}_k$, и назовем эти числа m-состоянием разбиения $\mathcal{K}_m$. К шагу 2.

Шаг 2. Найдем такое $j*$, что $X_i \in \Delta^{(k-1)}(j*)$ из уже построенного набора $\mathcal{D}_{k-1}$ и. К шагу 3.

Шаг 3. Вычислим вспомогательные величины

$$\delta'_{i,j} \ = \ \sum_{s \in \Delta^{(k-1)}(j), s \neq i} d_s^{(k-1)} r_{s,i},\ j = 1, ..., q$$

$$\delta_{i,j} \ = \ \begin{cases} 1 \text{ при } \delta'_{i,j} > 0, \\ 0 \text{ при } \delta'_{i,j} = 0, \\ \text{-1 при } \delta'_{i,j} < 0. \end{cases} , \ j = 1, ..., q.$$

К шагу 4.

Шаг 4. Вычислим для всех $j = 1, ..., q$ разности

$$E_j(i) \ = \ \sqrt{\mathbf{D}\left(\sum_{(j),i,*} d_s^{(k-1)} X_s + \delta_{i,j} X_i\right)} \ - \ \sqrt{\mathbf{D}\left(\sum_{(j),i,*} d_s^{(k-1)} X_s\right)}^2.$$

Здесь $\sum_{(j),i,*}$ вычисляется по $s \in \Delta_j^{(k-1)}, s \neq i$. Выберем такой номер $j0$, что

$$E_{j0}(i) \ = \ \max_{j=1,...,q} E_j(i),$$

затем исключим признак X_i из группы $\Delta^{(k-1)}(j*)$ и добавим его в группу $\Delta^{(k-1)}(j0)$. Тем самым построено $\mathcal{D}_k$. К шагу 5.

Шаг 5. Определим новые коэффициенты по формулам

$$d_j^{(k)} \ = \ \begin{cases} d_j^{(k-1)} \text{ при } j \notin \Delta^{(k-1)}(j0), \\ \delta_{i,j} \quad \text{при } j \in \Delta^{(k-1)}(j0). \end{cases}$$

К шагу 6.

Шаг 6. Увеличим номер i. Если $i \leq p$, то к шагу 2, иначе к шагу 7.

Шаг 7. Рассмотрим $(m+1)$-состояние системы $\mathcal{K}_{m+1}$ (напомним, что это набор, состоящий из разбиения $\mathcal{D}_k$ и коэффициентов $\vec{d_k}$ перед проходом алгоритма по всем p признакам). Сравним $\mathcal{K}_{m+1}$ с $\mathcal{K}_m$. Если они совпадают, то конец работы алгоритма, иначе полагаем $m = m + 1$ и к шагу 1.

Тем самым, алгоритм заканчивает работу только тогда, когда при изучении всех признаков ни один не перешел в другую группу.

10.2 Метод корреляционных плеяд

Этот метод применяется тогда, когда среди p имеющихся у нас признаков нужно выделить группы признаков, наиболее тесно связанных между собой, причем количество этих групп ("плеяд") неизвестно. Тем самым, решаемая задача близка к экстремальной группировке признаков и две этих задачи могут решаться параллельно. Рассмотрим два варианта этого простого эмпирического метода.

Предположим, что по результатам наших наблюдений мы имеем оценку B корреляционной матрицы наших признаков (напомним, что это квадратная матрица порядка p, у которой (i,j)-м элементом является $r_{i,j}$ – коэффициент корреляции между X_i и X_j). Отметим на плоскости p точек, каждая из которых будет изображать один из признаков, соединим их каждую с каждой отрезком прямой и возле отрезка, соединяющего i-ю и j-ю точку подпишем число $r_{i,j}$. Этим числом мы определим вес соответствующей связи. Теперь зададим некое пороговое значение r и сотрем все отрезки, вес которых меньше, чем r. Если мы начнем постепенно увеличивать пороговое значение r, то рано или поздно наш рисунок разобъется на не связанные между собой части – "плеяды". Эти части и представляют собой искомые группы признаков.

Несмотря на свою простоту, описанный алгоритм эффективно работает на начальных этапах выяснения зависимостей и довольно часто применяется практиками.

Иногда наряду с выделением наиболее сильно коррелированных групп признаков ставится задача описания так называемого дерева зависимостей признаков. Для этого требуется указать для каждого из признаков те из остальных, которые наиболее сильно с ним связаны. Эту задачу ре-

шает так называемый алгоритм Крускала из теории графов. Его работу можно описать следующим образом:

Шаг 1. В корреляционной матрице B находим максимальный элемент $r_{i,j}$ из тех, что не расположены на главной диагонали (а на ней, напомним, находятся единицы) соединяем точки, изображающие X_i и X_j и приписываем нарисованному отрезку вес $r_{i,j}$. Полагаем $k = i,\ s = j$. К шагу 2.

Шаг 2. В k-й и s-й строчке B ищем максимальные элементы среди таких, номера которых ранее не встречались в нашем алгоритме. Если таких не удалось найти ни одного, то работа алгоритма закончена, иначе к шагу 3.

Шаг 3. Среди найденных элементов (возможно, он один) выбираем максимальный. Пусть это элемент $r_{a,b}$. Заметим, что один из индексов a, b обязательно совпадает с k или s, а другой нет. Рисуем отрезок, соединяющий точку, соответствующую X_a и X_b и придаем ему соответствующий вес. Производим переприсвоение значений – придаем тому из индексов k, s, который совпадал, например, с a, значение b (или наоборот), а второй оставляем неизменным. К шагу 2.

После того, как работа этого алгоритма закончена (а он обязательно закончится не позднее, чем через $p - 1$ проходов, так как повторение точек у нас запрещено), у нас будет нарисован граф без циклов, описывающий структуру коррелированности признаков. Согласно принятой в теории графов терминологии, этот граф будет деревом. Нетрудно понять, что среди всех деревьев, построенных на наших точках, это дерево будет "наиболее тяжелым", а в случае, когда совместное распределение признаков было нормальным, оно описывает структуру собственно зависимостей признаков.

Если теперь нужно выделить группы наиболее коррелированных признаков, то процедуру с увеличением порогового значения, описанную чуть выше, применим к построенному дереву. В практических задачах именно эта процедура доказала свою большую эффективность по сравнению с описанным в начале раздела методом. Это обусловлено многими причинами, например, тем, что обычно встречаются именно древообразные структуры зависимостей признаков, а не какие-либо иные.

Глава 11

Метод главных компонент

11.1 Постановка задачи

Выше немного уже говорилось о задаче выделения главных компонент. Суть ее в том, чтобы среди всех линейных комбинаций p признаков наблюдаемых объектов выделить гораздо меньшее число q таких, изменчивость которых в значительной степени описывает изменчивость всего первоначального набора признаков в целом. В дальнейшем можно использовать эти найденные комбинации (которые и называются главными компонентами) для классификации и других задач, связанных с изучавшимися объектами.

Простым бытовым примером, обычно приводимым в учебниках, служит процесс покупки готовой одежды, когда мы вполне обходимся (если примерка невозможна) двумя, в крайнем случае тремя показателями – размером, ростом и полнотой, тогда как опытный портной при получении заказа снимает с клиента до 11 различных размеров. Можно привести также пример, когда измеряемые по разным методикам коэффициенты экономической активности сводят путем их комбинирования к одному (интегративному) показателю. За последним примером стоит мысль о том, что, хотя каждый из предлагаемых коэффициентов по разному учитывает экономические факторы, но все они призваны объяснять одно и то же явление, и значит, это явление наилучшим образом должно описываться какой-то их комбинацией, являющейся как бы результатом "компромисса"между различными методиками – в споре, как говорят, рождается истина.

Перейдем к формальной постановке. Пусть, $\vec{X} = (X^{(1)}, ..., X^{(p)})^t$, как и выше, представляет собой вектор показателей, наблюдаемых у каждого из n объектов. Здесь мы будем предполагать, что среднее выборочное значение каждого из наблюдаемых показателей $X^{(j)}$, $j = 1, ..., p$ равно нулю. Это будет удобно при выписывании формул, а если для конкретных выборочных данных это не так, например

$$M_j = \frac{1}{n}\sum_{i=1}^{n} X_i^{(j)} \neq 0,$$

то можно центрировать наблюдаемый фактор, т.е. рассмотреть вместо него $X^{(j),\prime} = X^{(j)} - M_j$. Здесь мы сделаем небольшую оговорку. Конечно же, для того, чтобы добиться центрированности теоретического распределения $\vec{X}$, что понадобится нам для вывода формул главных компонент в следующем разделе, в качестве M_j должно фигурировать теоретическое математическое ожидание. Поэтому, имея ввиду задачи практики, условимся считать, что либо это математическое ожидание нам было известно, либо его оценка практически точно совпадает со значением. Если это не так, то точность всех формул следующего раздела настолько же велика, насколько достоверно упомянутое здесь совпадение.

Рассмотрим класс $\mathcal{F}$ всех линейных ортонормированных преобразований p-мерного вектора $\vec{X}$ в q-мерный вектор $Z(\vec{X})$, где $q \ll p$.

Под ортонормированным преобразованием здесь понимается умножение слева на матрицу A, имеющую q строк и p столбцов, элементы которой $A_{i,j}$, $i = 1, ..., q, j = 1, ..., p$ удовлетворяют условию ортогональности строк

$$\sum_{j=1}^{p} A_{i,j}A_{k,j} = 0, \text{ для произвольных } i \neq k.$$

и строки имеют единичные нормы:

$$\sum_{j=1}^{p} A_{i,j}^2 = 1.$$

Тогда $Z = A\vec{X}$.

Функционал

$$K_q(Z) = \frac{\sum_{j=1}^{q} \mathbf{D}Z^{(j)}}{\sum_{j=1}^{p} \mathbf{D}X^{(j)}}.$$

будем здесь считать критерием оптимальности. Таким образом, ортонормированное преобразование оказывается тем лучше, чем ближе сумма дисперсий координат его образа приближается с сумме дисперсий первоначальных показателей. Это означает, что изменчивостью q линейных комбинаций показателей объясняется изменчивость их самих в наибольшей степени.

Отметим, что величины дисперсий показателей, которыми определяется критерий оптимальности, существенно зависят от масштаба единиц измерения. Если первоначально показатели имеют одинаковую природу и сравнимый масштаб измерений, то изучение главных компонент можно продолжать в "естественных единицах"измерения. Если же это не так, то их можно привести в единый (безразмерный) масштаб единиц, нормируя среднеквадратическими отклонениями (см. пример ниже). При этом мы от ковариационной матрицы переходим к корреляционной. Описанные два подхода (через ковариационную или корреляционную матрицы показателей) некоторыми авторами рассматриваются как разумно альтернативные. При оценивании главных компонент через корреляционную матрицу результаты оказываются практически совпадающими с результатами факторного анализа (см. соответствующую главу).

Итак, если на некотором $Z_* = A_*\vec{X}$ выполнено

$$K_q(Z_*) \;=\; \max_{Z\in\mathcal{F}} K_q(Z),$$

то координатные отображения $z_*^{(1)}, ..., z_*^{(q)}$ преобразования Z_* называют q главными компонентами показателей $\vec{X}$.

Если постараться перевести это на "общечеловеческий язык без формул"(хотя у нас есть сильное сомнение насчет того, что именно такой язык в данной ситуации легче поддается пониманию), то первая главная компонента – это такая ортонормированная комбинация показателей, которая из всех таких комбинаций обладает самой большой дисперсией, т.е. при переходе от объекта к объекту меняется сильнее всего. Отметим, что в случае первой главной компоненты речь идет о сумме исходных показателей с коэффициентами, сумма квадратов которых равна единице, т.е. остается лишь второе из двух выписанных условий ортонормированности преобразования.

Если $k-1$ главных компонент уже определены, то k-й главной компонентой называется такая ортонормированная комбинация показателей, которая не коррелирована с найденными $k-1$ главными компонентами, и

среди таких обладает наибольшей возможной дисперсией. Следовательно, строя главные компоненты, мы шаг за шагом "забираем"из суммы дисперсий (изменчивости) всех показателей сначала самую большую из всех возможных долей, затем следующую по величине и т.д.

С некоторой степенью приблизительности можно предложить также следующую геометрическую интерпретацию (которая будет абсолютно адекватной в случае нормального распределения $\vec{X}$). Рассмотрим точки в p-мерном пространстве, отвечающие n наблюдаемым объектам. Множество этих точек по-разному вытянуто в разных направлениях. Тогда первая главная компонента отвечает направлению, вдоль которого это множество вытянуто наиболее сильно, а значения этой компоненты равны проекциям радиус-векторов выборочных точек на это направление. После того, как определены $k-1$ главная компонента, на них, как на векторы, натягивается $k-1$-мерное подпространство, из всех радиус-векторов вычитаются их проекции на это подпространство и оставшиеся векторы изучаются по той же схеме в $(p-k+1)$-мерном пространстве. k-я главная компонента будет направлением "наибольшей протяженности"в этом пространстве меньшего числа измерений.

11.2 Вычисление главных компонент и их числовые характеристики

Введем в рассмотрение векторы - строки искомой матрицы преобразования:

$$\vec{a}^{(i)} \;=\; (A_{i,1},...,A_{i,p})^t, \quad i=1,...,q.$$

Тогда задача поиска первой главной компоненты может быть поставлена следующим образом: найти такой вектор $\vec{a}_*^{(1)}$, что

$$\mathbf{D}(\vec{X}\cdot\vec{a}_*^{(1)}) \;=\; \max_{\vec{a}\,:\,\|\vec{a}\|=1} \mathbf{D}(\vec{X}\cdot\vec{a}). \tag{11.1}$$

Обозначая ковариационную матрицу $\vec{X}$ через V и вспоминая, что мы предположили центрированность всех координат векторов наблюдений, видим, что

$$\mathbf{D}\left(\vec{X}\cdot\vec{a}\right) \;=\; \mathbf{M}\left(\sum_{j=1}^{p} a_j X^{(j)}\right)^2 \;=\; V\vec{a}\cdot\vec{a},$$

что позволяет пересмотреть вид целевой функции в поставленной задаче. Условный экстремум будем искать методом неопределенных множителей. Функция Лагранжа для (11.1) имеет вид

$$L(\vec{a}, \lambda) \;=\; V\vec{a} \cdot \vec{a} \;-\; \lambda(\vec{a} \cdot \vec{a} - 1).$$

После расчета частных производных получаем систему

$$\begin{cases} (V \;-\; \lambda I)\vec{a} \;=\; \vec{0}, \\ \|\vec{a}\| \;=\; 1, \end{cases} \tag{11.2}$$

что означает, что искомый вектор является одним из собственных векторов ковариационной матрицы, имеющим единичную длину. Учитывая, что первое из уравнений в системе (11.2), умноженное справа на $\vec{a}$ дает

$$V\vec{a} \cdot \vec{a} - \lambda\vec{a} \cdot \vec{a} \;=\; 0,$$

или

$$V\vec{a} \cdot \vec{a} \;=\; \lambda,$$

получаем, что $\mathbf{D}z_*^{(1)} \;=\; \lambda$, где справа стоит одно (любое) собственное число матрицы V. Учитывая тот факт, что дисперсия первой главной компоненты должна быть максимальной, получаем, что величина этой дисперсии – максимальное собственное число ковариационной матрицы, а сама главная компонента определяется через единичный собственный вектор $\vec{a}_*^{(1)}$ этой матрицы соотношением

$$z_*^{(1)}(X) \;=\; \vec{a}_*^{(1)} \cdot \vec{X}.$$

После того, как первая главная компонента найдена, "нейтрализуем"ее значение с помощью ортогонализации выборочных данных. Точнее, пусть $\vec{a}_*^{(1)}$ – собственный вектор матрицы V, отвечающий ее максимальному собственному числу λ_1. Рассмотрим новый вектор выборочных показателей

$$\vec{X}' \;=\; \vec{X} \;-\; (\vec{X} \cdot \vec{a}_*^{(1)})\vec{a}_*^{(1)} \tag{11.3}$$

и изучим линейное преобразование V', которое p-мерному вектору $\vec{X}$, принадлежащему $p-1$-мерному подпространству $R^{(1)}$, ортогональному $\vec{a}_*^{(1)}$, ставит в соответствие p-мерный вектор $\vec{Y}$ по правилу

$$Y \;=\; V'\vec{X} \;=\; V\vec{X}$$

Определенное таким образом преобразование используем при поиске второй главной компоненты согласно приведенному выше определению.

Заметим, для произвольного $X \in R^{(1)}$ получается $Y \in R^{(1)}$, а значит, у введенного преобразования V' все собственные числа матрицы V, за исключением максимального, также являются собственными числами. Действительно, используя (11.3), получаем

$$V'(\mu X) \;=\; \mu\vec{X} \;-\; (\mu\vec{X}\cdot\vec{a}_*^{(1)})\vec{a}_*^{(1)} \;=\; \mu\vec{X}',$$

если только $\mu \neq \lambda_1$ – собственное число V. Поскольку максимальное собственное число V здесь уже не участвует, то мы получим, что максимальное собственное число сужения V' на $(p-1)$-мерное подпространство, ортогональное $\vec{a}_*^{(1)}$, равно второму по величине собственному числу ковариационной матрицы.

Совершенно понятно, что описанный только что процесс можно продолжать при переходе к третьему, четвертому и т.д. числу, а следовательно, нами доказана

Теорема 18 *Задача нахождения главных компонент совпадает с задачей на собственные числа и собственные векторы ковариационной матрицы V вектора наблюдаемых показателей. Если $\lambda_1, ..., \lambda_p$ – ее собственные числа, расположенные в порядке убывания, а единичные собственные векторы $\vec{a}_*^{(1)}, ..., \vec{a}_*^{(p)}$ отвечают этим числам, то главные компоненты вычисляются по формулам*

$$z_*^{(k)}(X) \;=\; \vec{a}_*^{(k)}\cdot\vec{X},$$

причем

$$\mathbf{D}z_*^{(k)}(X) \;=\; \lambda_k, \quad k = 1, ..., p.$$

Из этой теоремы вытекает способ определения количества q главных компонент, если это число не было известно заранее, а именно: решив задачу на собственные числа для матрицы V и расположив их в порядке убывания, как в теореме, будем вычислять дроби

$$s_1 = \frac{\lambda_1}{\sum_{j=1}^{p}\lambda_j},\; s_2 = \frac{\sum_{i=1}^{2}\lambda_i}{\sum_{j=1}^{p}\lambda_j}, ..., s_p = \frac{\sum_{i=1}^{p}\lambda_i}{\sum_{j=1}^{p}\lambda_j} = 1.$$

Выберем такой номер q, для которого в первый раз величина s_q станет настолько близка к единице, что эта близость нас устроит. Выбранный номер и будет искомым количеством главных компонент.

Предложенный алгоритм основан на том факте, что

$$\sum_{i=1}^{p} \mathbf{D}X_i \;=\; \sum_{i=1}^{p} V_{i,i} \;=\; \mathbf{tr}V,$$

а след матрицы при ортогональных преобразованиях (каким является преобразование $\vec{X}$ к набору всех p его главных компонент) не меняется, и на следующей теореме.

Теорема 19 *Математическое ожидание вектора p главных компонент равно нулю, а его ковариационная матрица диагональна с диагональными элементами* $\lambda_1, ..., \lambda_p$.

Доказательство. В силу сделанных в начале главы предположений,

$$\mathbf{M}Z_* \;=\; \mathbf{M}(A_*\vec{X}) \;=\; A_*\mathbf{M}\vec{X} \;=\; \vec{0},$$

а ковариационная матрица

$$B \;=\; \mathbf{cov}Z_* \;=\; \mathbf{M}Z_*Z_*^t \;=\; \mathbf{M}(A_*\vec{X}\vec{X}^tA_*^t) \;=\; A_*VA_*^t.$$

Но это означает, что B представляет собой матрицу преобразования V, записанную в базисе из собственных векторов, что и завершает доказательство теоремы.

11.3 Пример

Рассмотрим один числовой пример для иллюстрации алгоритма поиска главных компонент. Пусть у каждой из 10 производственных фирм, выпускающих однотипную продукцию, сняты 3 показателя. Назовем их "объем капиталовложений", "продажная цена"и "себестоимость единицы продукции". Эти показатели (по крайней мере первый в отличие от второго и третьего) измерялись в разных масштабах единиц. Обозначив перечисленные показатели в порядке перечисления X_1, X_2, X_3, приведем таблицу полученных в результате сбора информации числовых данных.

Данные по 10 фирмам
для выделения главных компонент

номер	X_1	X_2	X_3	номер	X_1	X_2	X_3
1	10,0	749,0	744,4	6	3,8	757,5	754,0
2	6,2	746,1	756,6	7	17,1	752,4	747,8
3	6,3	756,6	752,4	8	22,2	752,5	748,6
4	5,3	758,9	754,7	9	20,8	752,2	747,7
5	4,8	751,7	747,9	0	21,0	759,5	755,6

Для того, чтобы были выполнены условия центрированности данных, которые необходимы для того, чтобы применять формулы из предыдущего раздела, вычтем из всех значений средние $\bar{X}_1 = 11,75,\ \bar{X}_2 = 753,64,\ \bar{X}_3 = 749,47$, а для того, чтобы привести в соответствие масштаб единиц, нормируем получившиеся значения среднеквадратическими отклонениями $S_1 = 7,61,\ S_2 = 4,38,\ S_3 = 4,68$. После этого вычислим ковариационную матрицу. Она равна

$$V = \begin{pmatrix} 1 & 0,018 & -0,289 \\ 0,018 & 1 & 0,413 \\ -0,289 & 0,413 & 1 \end{pmatrix}.$$

Три ее собственных числа в порядке убывания составляют 1,496; 1,017; 0,487. Собственные векторы, имеющие единичную длину

$$\vec{a}_*^{(1)} = \begin{pmatrix} 0,396 \\ 0,579 \\ 0,713 \end{pmatrix},\ \vec{a}_*^{(2)} = \begin{pmatrix} 0,819 \\ -0,573 \\ 0,01 \end{pmatrix},\ \vec{a}_*^{(3)} = \begin{pmatrix} 0,419 \\ 0,578 \\ -0,70 \end{pmatrix}.$$

Это означает, что главные компоненты в этом случае при заданных значениях $\vec{X}$ должны вычисляться по формулам

$$\begin{aligned} z_*^{(1)}(X) &= 0,396X_1^* + 0,579X_2^* + 0,713X_3^*, \\ z_*^{(2)}(X) &= 0,819X_1^* - 0,573X_2^* + 0,01X_3^*, \\ z_*^{(3)}(X) &= 0,419X_1^* + 0,578X_2^* - 0,7X_3^*. \end{aligned}$$

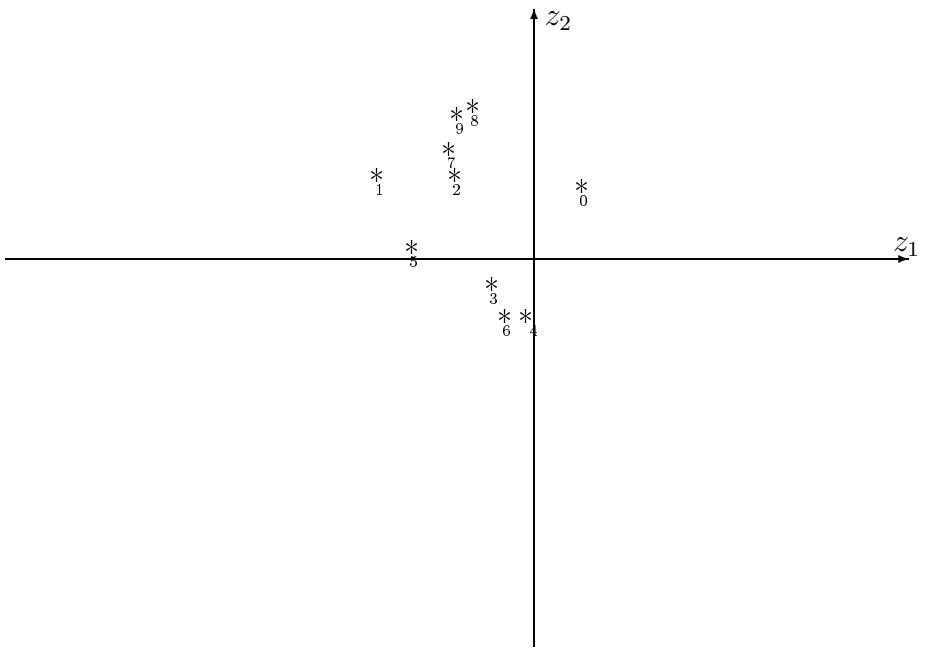
Здесь $X_i^*,\ i = 1,2,3$ – центрированные и нормированные первоначальные показатели. Из этих формул видно, что в первую и третью главную компоненту примерно равный вклад дают все использовавшиеся показатели, а во второй главной компоненте можно пренебречь влиянием третьего показателя.

При этом, привлекая только первую главную компоненту, мы с помощью нее объясним $S_1 = 0,499$ всей изменчивости $\vec{X}$, рассматривая две главных компоненты – $S_2 = 0,838$ изменчивости. Третья же компонента, как следует из наших формул, имеет весьма невысокую значимость. Мы видим, что по своей сути задача двумерная, и с минимальными искажениями мы можем оставить две главных компоненты. За счет этого мы можем, например, изобразить наши фирмы точками на плоскости, приписав каждой из них координаты, рассчитывающиеся по формулам первых двух главных компонент. Результат соответствующего пересчета приведен в таблице.

Две главные компоненты
в примере с 10 фирмами

номер	первая	вторая	номер	первая	вторая
1	-1,82	0,44	6	0,61	-1,35
2	-0,32	0,41	7	-0,43	0,73
3	0,35	-0,97	8	-0,01	1,27
4	0,99	-1,37	9	-0,28	1,15
5	-1,14	-0,50	0	2,54	0,24

Ниже точки, соответствующие фирмам, приведены в осях главных координат (начало координат смещено для придания большей наглядности). По рисунку можно делать наглядные выводы об отнесении фирм к той или иной группе по близости их друг к другу.



11.4 Самовоспроизводимость и малое геометрическое искажение

Будем в этом разделе, как и выше, предполагать, что исходные данные наблюдений центрированы и нормированы.

11.4.1 Самовоспроизводимость

Оказывается, главные компоненты наилучшим образом сохраняют информацию, содержавшуюся в исходных показателях и в том смысле, что по данным главным компонентам исходные признаки восстанавливаются как линейные комбинации наилучшим образом. Более точно, имеется ввиду следующее.

Пусть при заданном $q < p$ требуется построить такую линейную комбинацию $Z(A) = A\vec{X}$, переводящую p-мерный $\vec{X}$ в q-мерный $Z(A)$, что построенные методом наименьших квадратов векторы коэффициентов $\beta^{(j)},\ j = 1, ..., p$ позволяют осуществить наилучшее приближение исходного вектора $\vec{X}$ среди всех линейных комбинаций $Z(A)$ в том смысле,

что если

$$\Delta(A) \;=\; \sum_{j=1}^{p}\left(X_j \;-\; \sum_{i=1}^{q}\beta_i^{(j)}Z_i(A)\right)^2,$$

и требуется найти такую A_*, что

$$\Delta(A_*) \;=\; \min_A \Delta(A),$$

тогда A_* оказывается той матрицей, которая фигурирует в методе главных компонент, т.е. $Z_i(A_*),\ i=1,...,q$ – q первых главных компонент.

Покажем описанный процесс на примере из предыдущего пункта. Согласно обычной процедуре метода наименьших квадратов (см. теорему 6 и формулу (6.9)) в наших условиях ортогональности главных компонент,

$$\beta_j^{(i)} \;=\; \frac{\mathbf{cov}(X_i, Z_j(A_*))}{\mathbf{D}Z_j(A_*)}.$$

Отсюда (см. предыдущий раздел)

$$\begin{aligned}\beta_1^{(i)} \;&=\; 0,396\mathbf{cov}(X_i,X_1) \;+\; 0,579\mathbf{cov}(X_i,X_2) \;+\; 0,713\mathbf{cov}(X_i,X_3),\\ \beta_2^{(i)} \;&=\; 0,819\mathbf{cov}(X_i,X_1) \;-\; 0,573\mathbf{cov}(X_i,X_2) \;+\; 0,01\mathbf{cov}(X_i,X_3)\end{aligned}$$

Взяв значения ковариаций, входящих в эти соотношения из ковариационной матрицы $\vec{X}$ выше, получим

$$\begin{aligned}&\beta_1^{(1)} = 0,200 \qquad \beta_2^{(1)} = 0,806;\\ &\beta_1^{(2)} = 0,905 \qquad \beta_2^{(2)} = -0,561;\\ &\beta_1^{(3)} = 0,838 \qquad \beta_2^{(3)} = -0,463.\end{aligned}$$

Наконец, восстановим значения исходных показателей $\vec{X}$ по формулам

$$\begin{aligned}X_i' \;&=\; \beta_1^{(i)}Z_1 \;+\; \beta_2^{(i)}Z_2,\\ X_i \;&=\; \sqrt{\mathbf{D}X_i}X_i' + \bar{X}_i, \quad i=1,2,3.\end{aligned}$$

Результаты приведены в таблице. Заметим, что для восстановления исходных значений нам понадобилось вспомнить, чему равны средние и дисперсии исходных показателей. Отсюда следует, что главные компоненты в том виде, о котором мы говорим, несут информацию не о самих показателях, а только об их изменчивости (самих по себе и относительно друг друга). Впрочем, при внимательном отношении к излагаемому материалу, это можно было заметить и раньше.

Восстановленные по главным
компонентам исходные показатели

номер	X_1	X_2	X_3	номер	X_1	X_2	X_3
1	11,7	745,3	741,6	6	4,4	759,4	754,9
2	8,8	757,6	753,1	7	15,6	750,1	746,4
3	6,3	757,4	753,1	8	19,5	750,5	746,9
4	4,8	760,9	756,5	9	18,4	749,7	746,1
5	6,9	750,3	746,3	0	17,1	763,1	759,1

В заключение этого подраздела приведем значение величины ошибки восстановления $\Delta(A_*) = 4,15$. Если бы мы знали, что восстанавливаемые данные имеют нормальные распределения, все параметры которого точно известны, то с целью оценить качество восстановления мы должны бы были сравнить это число с критической точкой распределения хи-квадрат с 30 степенями свободы.

11.4.2 Геометрические свойства

При переходе к главным компонентам с геометрической точки зрения мы производим проектирование p-мерного пространства в q-мерное. Естественно, что при этой процедуре неизбежны искажения линейных размеров и углов. Оказывается, среди отображений упомянутых пространств, имеющих линейный характер, переход к главным компонентам является наилучшим. Просто перечислим наиболее простые свойства такого рода. Обозначим $R(q)$ подпространство исходного p-мерного пространства, порожденное первыми q главными компонентами, а $\mathcal{R}(q)$ – класс всех q-мерных подпространств исходного пространства показателей, порождаемых линейными комбинациями этих показателей.

1. Сумма квадратов расстояний от точек, изображающих наблюдения $\vec{X}_1, ..., \vec{X}_n$ до $R(q)$ наименьшая для всех подпространств класса $\mathcal{R}(q)$.

2. Сумма квадратов расстояний между всеми парами проекций выборочных точек в $R(q)$ по отношению к сумме квадратов расстояний между исходными точками $\vec{X}_1, ..., \vec{X}_n$ наименьшим образом искажена среди всех подпространств класса $\mathcal{R}(q)$.

3. Среди всех подпространств класса $\mathcal{R}(q)$ в пространстве $R(q)$ наименее искажаются расстояния проекций до их геометрического центра а также углы между отрезками прямых, соединяющих каждую из проекций с геометрическим центром множества проекций.

Итак, можно говорить о том, что среди всех линейным образом организованных проекций в пространство более низкой размерности, метод главных компонент наименьшим образом искажает геометрическую структуру облака выборочных точек. Более или менее точное представление о том, в каком смысле здесь говорится о геометрической структуре, можно получить, вернувшись к перечисленным выше свойствам. Но иногда бывает нужно при снижении размерности сохранить "пространственную изолированность"отдельных групп точек или оставить неизменными другие свойства, сохранение которых свойствами 1-3 не гарантируется. Более того, та оптимальность сохраняемых свойств, о которой мы все время говорим, касается лишь пространства $\mathcal{R}(q)$, связанного с линейными преобразованиями признаков.

Тогда имеет смысл обратиться к отображениям p-мерного пространства в q-мерное, не являющимися линейными, и к изучению их свойств. В этой ситуации Дж.Сэммоном предложен следующий критерий оптимальности искажения геометрической структуры, подобный стресс-критерию в задачах многомерного шкалирования, изучаемых в главе 14:

$$K_q(Z,a) \;=\; \frac{\sum_{i=1}^{n-1}\sum_{j=i+1}^{n}(\delta_{i,j}-d_{i,j})^2\delta_{i,j}^a}{\sum_{i=1}^{n-1}\sum_{j=i+1}^{n}\delta_{i,j}^a},$$

где $d_{i,j}$ – расстояния между i-й и j-й точками в исходном пространстве, $\delta_{i,j}$ – расстояния между их проекциями, число a выбирается исследователем в зависимости от содержания задачи. Рекомендуется выбирать $a<0$, чаще всего $a=-1$. Иногда для разных пар точек берут разные a. В любом случае вычислительная техника дает возможность попробовать разные варианты прежде, чем будет принято окончательное решение.

Построение наилучших в смысле критерия Сэммона нелинейных проекций организуется в виде итерационного процесса. В качестве начального приближения выбирают, как правило, проекцию на главные компоненты. За подробностями отсылаем читателя к [2, с.372-375] и [11].

Глава 12

Факторный анализ

12.1 Постановка задачи

Факторный анализ занимается определением относительно небольшого числа скрытых (латентных) факторов, изменчивостью которых объясняется изменчивость всех наблюдаемых показателей, связанных с каждым из подвергающихся изучению объектов. В этом смысле оцениваемые латентные факторы $f^{(1)}, ...f^{(q)}$ можно считать причинами, а наблюдаемые признаки $X^{(1)}, ..., X^{(p)}$, – следствиями. Результаты факторного анализа условимся считать успешными, если большое число признаков удалось достаточно точно объяснить малым количеством причин.

12.1.1 Связь с главными компонентами

Итак, факторный анализ направлен на снижение размерности рассматриваемой задачи. Если мы еще в явном виде сформулируем, что латентные (или, как их еще называют, общие) факторы ищутся в виде линейных комбинаций наблюдаемых признаков, то может создаться впечатление, что такую задачу мы уже решали. Например, сходные вопросы рассматривались в главах, посвященных регрессии и методу главных компонент. Но сейчас перед нами другая задача. Так, в задачах регрессии мы имели возможность наблюдать как значения тех факторов, от которых зависел выход, так и сам выход, а неизвестными, оцениваемыми параметрами являлись только коэффициенты этой зависимости (и, конечно же, остаточная дисперсия). Теперь же у нас к числу наблюдае-

мых относится лишь вектор признаков $\vec{X}$, тогда как латентные факторы сами относятся к числу оцениваемых.

Чтобы понять связь той задачи, которую мы намереваемся решать с задачей метода главных компонент, поставим задачу факторного анализа чуть аккуратнее. Пусть $\vec{X} = (X^{(1)}, ..., X^{(p)})^t$ p-мерный вектор наблюдаемых признаков, связанный с каждым из n изучаемых объектов. Требуется при заданном q (которое, как правило, значительно меньше p) найти такую матрицу A, содержащую q столбцов и p строк, называемую матрицей факторных нагрузок, и такой вектор $\vec{F} = (f^{(1)}, ..., f^{(q)})^t$ латентных факторов, являющихся линейными комбинациями исходных признаков, что

$$\vec{X} = A\vec{F} + \vec{\xi},$$

где $\vec{\xi}$ – вектор "случайных погрешностей", таким образом, что изменчивость $\vec{F}$ наилучшим образом объясняет изменчивость $\vec{X}$ среди всех линейных комбинаций $X^{(1)}, ..., X^{(p)}$. Оцениванию подлежит также матрица $V = \mathbf{cov}\vec{\xi}$.

Предположим теперь на минуту, что существует (возможно бесконечный) вектор "причин" $\vec{F}$ и (опять-таки, возможно, бесконечная) матрица A, такие, что

$$\vec{X} = A\vec{F},$$

т.е.

$$X^{(k)} = \sum_{j=1}^{\infty} A_{k,j} f^{(k)}, \quad k = 1, ..., p,$$

причем элементы матрицы A неизвестны. Исследователь собирается оценить q "причин" $f^{(1)}, ..., f^{(q)}$ и коэффициенты $A_{k,j},\ k = 1, .., p,\ j = 1, ..., q$ таким образом, чтобы, заменив полные (бесконечные) суммы в правой части последнего равенства их урезанными вариантами

$$X^{(k)}(q) = \sum_{j=1}^{q} A_{k,j} f^{(k)}, \quad k = 1, ..., p$$

так, чтобы полученная аппроксимация $\vec{X}$ была бы оптимальной в определенном смысле.

Оказывается, если мы в этой задаче в качестве критерия оптимальности возьмем минимальность отличия матрицы $B = \mathbf{cov}\vec{X}$ от ковариационной матрицы вектора $\vec{X}(q)$, то мы придем к задаче поиска главных компонент. Точнее, мы получим, что i-й столбец матрицы A равен $\sqrt{\lambda_i}\vec{a}_*^{(i)}$,

где λ_i – i-е собственное число матрицы B, а $\vec{a}_*^{(i)}$ – ее собственный вектор единичной длины. При этом также окажется, что

$$f^{(i)} = \frac{z_*^{(i)}}{\sqrt{\lambda_i}},\ i = 1, ..., q,$$

где $z_*^{(i)}$ – i-я главная компонента.

Если же главной задачей объявить наиболее полное объяснение корреляции между координатами вектора $\vec{X}$, то мы придем к задаче факторного анализа. Другими словами эту процедуру можно описать следующим образом. Первый из латентных факторов подбирается из того условия, чтобы, после его исключения из всех наблюдаемых признаков, коэффициенты корреляции между всеми парами признаков были бы минимальными. Затем процесс повторяем для признаков с исключенным влиянием уже построенного фактора (факторов).

Отсюда следует, что если остаточные дисперсии были невелики, то метод главных компонент и факторный анализ должны давать близкие результаты. Именно поэтому до недавнего времени метод главных компонент считался частью факторного анализа, да и в некоторых относительно недавно изданных учебниках (см. [13]) под названием "факторный анализ"скрывается именно метод главных компонент.

12.1.2 Однозначность решения

Для начала отметим, что принципиальная возможность решения задачи факторного анализа как разновидности общей задачи снижения размерности, связана, в основном, с высокой коррелированностью исходных признаков. В качестве латентных факторов и выступают те причины, которые обеспечивают названную коррелированность. Наиболее ясно сказанное иллюстрируется на примере так называемых тестов на определение уровня интеллекта. Ответы на вопросы теста (а точнее, качество этих ответов) в конце концов определяются способностями тестируемого. Сами способности при этом мы наблюдать не можем (да и вряд ли кто-нибудь в состоянии напрямую строго определить, что это такое), а наблюдаем только их следствия. И по ним требуется эти самые способности оценить, что психологи проделывают весьма успешно.

Приведенный пример в некотором смысле объясняет тот факт, что методы факторного анализа впервые были предложены и применены

именно в работах психологов. По некоторому размышлению становится понятно, что, руководствуясь сходными соображениями, можно методами факторного анализа придавать численные значения непосредственно не наблюдаемым и даже нечисловым признакам, таким, например, как "настроение", "общее состояние экономики", "климат"и т.д. Автору пособия удалось применить этот метод для выработки "общего показателя здоровья", который при определенной нормировке может интерпретироваться как вероятность перейти в разряд "больных"в ближайшее время.

Остановимся на вопросе возможности однозначного решения задачи факторного анализа. Во-первых, если мы уже нашли какое-то решение задачи $A, \vec{F}$, то, записывая

$$\vec{X} - \vec{\xi} \;=\; A\vec{F} \;=\; (AC)(C^{-1}\vec{F})$$

для произвольной обратимой матрицы C, мы обнаруживаем еще бесконечно много решений. Даже условие единичных дисперсий координат вектора $\vec{F}$, которое обычно накладывается, может быть обойдено, если мы возьмем в качестве C ортогональную матрицу. Поскольку умножение на ортогональные матрицы соответствует поворотам, то мы видим, что решение принципиально может быть определено лишь с точностью до поворотов.

Наряду с некоторыми неприятными сторонами этот факт несет в себе и полезную составляющую. Дело в том, что при решении практических задач часто возникают проблемы интерпретации найденных латентных факторов, приданию им некоего "качественного"смысла. С одной стороны, такая интерпретация предполагает первоначальное приведение наблюдаемых признаков $\vec{X}$ к одному масштабу единиц (иначе трудно понять смысл линейной комбинации, например, рублей и гектаров), а с другой, после того, как решение найдено, мы получаем возможность путем вращений слегка изменить содержание факторов без снижения качества решения и тем самым облегчить специалисту-практику их качественную интерпретацию.

Тем не менее, даже если мы смиримся с невозможностью совершенно точно решить поставленную задачу и будем считать решение однозначным, если оно единственно с точностью до поворотов, в нашей постановке остается еще много степеней свободы (число неизвестных значительно больше числа уравнений). Об этом подробнее – в следующем разделе.

Какие же цели ставит перед собой исследователь, решающий задачу факторного анализа?

- определить алгоритм оценки матрицы факторных нагрузок A и ковариационной матрицы V остаточной компоненты $\vec{\xi}$;

- построить оценки латентных факторов F для каждого из наблюдаемых объектов и предложить формулу для оценки значения этих факторов для объекта, вновь поступающего на изучение;

- проверить некоторые гипотезы, связанные с принимаемой моделью: адекватность ее по отношению к исследуемому явлению, гипотеза об истинном числе латентных факторов и т.д.;

- путем вращений дать возможность интерпретировать полученные оценки латентных факторов на понятном практику языке.

Возможны, конечно же, и другие цели. Большинство этих проблем будет рассмотрено ниже. В заключение этого раздела отметим, что все поставленные задачи могут быть решены только с большей или меньшей степенью приближенности, а некоторые (в частности, задачи интерпретации) вообще принципиально неформализуемы.

12.2 Математическая модель

Переходим к формулам. Пусть с каждым из n наблюдаемых объектов связан p-мерный вектор показателей $\vec{X} = (X^{(1)}, ..., X^{(p)})^t$, матрица $B = \mathbf{cov}\vec{X}$ считается известной (по крайней мере, как оценка). Предполагаем, что при заданном $q < p$ и некоторых (неизвестных) матрице A, содержащей p строк и q столбцов и q-мерном векторе $\vec{F} = (f^{(1)}, ..., f^{(q)})^t$ выполнено

$$\vec{X} = A\vec{F} + \vec{\xi},$$

где $\vec{\xi} = (\xi_1, ..., \xi_p)^t$ – случайный вектор, причем $V = \mathbf{cov}\vec{\xi}$ предполагается диагональной (т.е. координаты случайного вектора некоррелированны), а числа, стоящие по диагонали (также неизвестные) называются остаточными дисперсиями.

Матрица A называется матрицей факторных нагрузок, вектор $\vec{F}$ – вектором латентных (общих) факторов. По данным $\vec{X}$ требуется оценить факторные нагрузки, остаточные дисперсии и значения латентных факторов таким образом, чтобы изменчивость координат $\vec{F}$ наилучшим

образом объясняла бы изменчивость всего набора исходных показателей. Критерием качества, следовательно, служит функционал

$$K_q(A,V) \;=\; \sum_{i=1}^{p-1}\sum_{j=i+1}^{p}\frac{1}{V_{i,i}V_{j,j}}\left(B_{i,j}-\mathbf{cov}(\sum_{k=1}^{q}A_{i,k}f^{(k)},\;\sum_{s=1}^{q}A_{j,s}f^{(s)})\right)^2.$$

При этом на латентные факторы налагаются условия нормировки, которые, если мы обозначим набор значений факторов в k-м эксперименте через $\vec{f_k} \;=\; (f_k^{(1)},...,f_k^{(q)}),\; k=1,...,n$, могут быть записаны следующим образом:

$$\frac{1}{n}\sum_{k=1}^{n}\vec{f_k} \;=\; \vec{0},\;\; \frac{1}{n}\sum_{k=1}^{n}\|\vec{f_k}\|^2 \;=\; 1.$$

Эти условия можно, если считать латентные факторы случайными величинами, интерпретировать как

$$\mathbf{M}\vec{f}_j^{(k)}=0,\;\; \mathbf{D}\vec{f}_j^{(k)}=1,\;\; k=1,...,q,\;\; j=1,...,n.$$

Из сделанных предположений следует, что вектор $\vec{F}$ как случайный вектор не коррелирован с $\vec{\xi}$, следовательно,

$$B \;=\; AA^t+V. \tag{12.1}$$

Возможность подобрать в этом уравнении такие A,V (последняя, напомним, диагональна) совпадает с принципиальной возможностью решить задачу факторного анализа. Вероятно, это можно сделать не всегда. Но мы, чтобы не загромождать изложение, предположим, что, во-первых, решение этой системы существует, а во-вторых, оно единственно с точностью до поворотов.

Как мы уже говорили выше, в системе (12.1) неизвестных гораздо больше, чем уравнений. Поэтому обычно рассматриваются различные дополнительные условия на матрицы A,V, при которых система и решается. Ниже мы рассмотрим разные варианты таких условий и дадим некоторые рекомендации по решению системы (12.1) в каждой из ситуаций в отдельности. Сейчас же рассмотрим следующую общую схему итеративных алгоритмов для ее решения.

Шаг 1. Зададимся некоторым начальным приближением V_0, положим $\Psi_0 \;=\; B \;-\; V_0$ (это – приближение для матрицы AA^t). К шагу 2.

Шаг 2. С помощью специальной процедуры, по разному организованной для разных видов дополнительных условий (см. далее соответствующие подразделы) по Ψ_0 определим приближение A_0 матрицы A, положим $i = 0$. К шагу 3.

Шаг 3. Вычислим $\Psi_{i+1} = A_i A_i^t$ и по той же процедуре через нее определим A_{i+1}. К шагу 4.

Шаг 4. Положим $V_{i+1} = B - \Psi_{i+1}$. Сравним A_{i+1} с A_i, V_{i+1} с V_i . $i = i + 1$. Если у хотя бы одной из этих пар наблюдаются значительные отличия, то к шагу 3, иначе – конец алгоритма.

12.2.1 Условия на $A^t A$

Здесь мы рассмотрим два варианта возможных дополнительных условий, в которых решение общей задачи факторного анализа становится однозначным с точностью до поворотов. В первом из них решения системы (12.1) ищут только среди таких матриц, что $A^t V A$ диагональна, причем элементы на ее главной диагонали различны и расположены в порядке убывания, во втором те же предположения диагональности делаются относительно $A^t A$, т.е. ковариационная матрица случайной составляющей в условии не участвует.

Как правило, при таких условиях дополнительно предполагают нормальность распределения $\vec{X}$. Затем рассматривают функцию правдоподобия, где в качестве параметров фигурируют факторные нагрузки и остаточные дисперсии. После этого методами математического анализа (метод неопределенных множителей Лагранжа) ищут максимум функции правдоподобия, как правило, при помощи итеративного алгоритма, почти дословно повторяющего приведенный выше.

Учитывая наложенные условия, нетрудно показать (в предыдущих главах мы по крайней мере дважды решали подробно очень похожие задачи), что специальная процедура, упомянутая там, сводится к решению систем

$$\Psi \vec{a}_i = \lambda_i V \vec{a}_i, \quad i = 1, ..., q$$

относительно столбцов $\vec{a}_i$ матрицы A, где λ_i – i-й по величине корень уравнения

$$|\Psi - \lambda V| = 0$$

($|.|$ – определитель), т.е. решению проблемы на обобщенные собственные числа и обобщенные собственные векторы в случае первого варианта до-

полнительных условий. Если же условия были наложены во втором варианте, то мы получаем просто проблему собственных чисел и векторов – столбцы матрицы A будут собственными векторами матрицы Ψ, имеющими единичную длину.

Находя собственные векторы (или, соответственно обобщенные собственные векторы) очередного приближения Ψ_i и располагая их по столбцам, мы получим очередное приближение A_i к матрице факторных нагрузок.

12.2.2 Условия на матрицу нагрузок. Центроидный метод

Еще один вариант дополнительных условий накладывается на саму матрицу нагрузок A. Мы рассмотрим требование, заключающееся в том, что первый исходный признак $X^{(1)}$ должен выражаться только через первый латентный фактор $f^{(1)}$, второй – через комбинацию первого и второго факторов и т.д. Это означает, что матрица факторных нагрузок ищется в виде

$$A = \begin{pmatrix} A_{1,1} & 0 & 0 & ... & 0 \\ A_{2,1} & A_{2,2} & 0 & ... & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ A_{q,1} & A_{q,2} & A_{q,3} & .... & A_{q,q} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ A_{p,1} & A_{p,2} & A_{p,3} & ... & A_{p,q} \end{pmatrix}$$

Опишем суть специальной процедуры в этих предположениях. Основная идея состоит в том, что если $\vec{a}_i$ – i-й столбец матрицы факторных нагрузок, то для матрицы $\Psi = AA^t$ имеет место представление

$$\Psi = \sum_{i=1}^{q} \vec{a}_i \vec{a}_i^t.$$

Поэтому сначала, пользуясь установленным видом A, можно определить $\vec{a}_1$, для чего пользуемся соотношениями

$$A_{1,1}^2 = \Psi_{1,1},\ A_{1,j} = \frac{\Psi_{1,j}}{\sqrt{\Psi_{1,1}}},\ j = 2, ..., p,$$

далее переходим к матрице $\Psi^{(1)} = \Psi - \vec{a}_1\vec{a}_1^t$ и процесс повторяется.

Этот алгоритм можно улучшить, за счет выбора нормирующих коэффициентов, принимающих значения ± 1. Такая процедура носит название центроидного метода и допускает простое геометрическое истолкование.

Представим себе наши исходные признаки $X^{(1)}, ..., X^{(p)}$ в виде векторов в n-мерном пространстве, выходящих из начала координат, имеющих длины, равные $\sqrt{\mathbf{D}X^{(i)}}$, $i = 1, ..., p$ и расположенных друг относительно друга таким образом, что косинус угла между i-м и j-м признаками равен коэффициенту корреляции $r_{i,j}$ между ними. Конечно же, для практических задач вместо упомянутых характеристик берутся их оценки. Теперь изменим некоторые из направлений на противоположные так, чтобы как можно больше коэффициентов корреляции стали бы положительными. Это соответствует умножению векторов-признаков на коэффициенты вида ± 1. После этого большинство наших векторов собралось в "пучок". В качестве первого латентного фактора выберем нормированный "средний вектор"этого пучка, т.е. тот вектор единичной длины, который сонаправлен сумме всех векторов, составляющих полученный пучок. После определения первого фактора перейдем в пространство, ортогональное направлению его вектора (что означает исключение влияния этого фактора) и продолжим нашу процедуру уже в пространстве меньшей размерности. Так будет выделен второй фактор и т.д.

Изложим далее описанный с геометрической точки зрения метод на языке формул. Пусть исходные показатели нормированы при помощи своих среднеквадратических отклонений. Возьмем некоторое начальное приближение $V^{(0)}$ для матрицы остаточных дисперсий. Напомним, что согласно предположениям модели, она диагональна. Поэтому достаточно выбрать ее диагональные элементы. Обычно, если $r(i) = \min_{j \neq i} r_{i,j}$, то считают, что на i-м месте главной диагонали следует поместить $1 - r(i)$, хотя возможны и иные решения.

Возьмем $\Psi^{(0)} = B - V^{(0)}$, и положим $h_j = 1$, $j = 1, ..., p, \vec{h} = (h_1, ..., h_p)^t$. Пусть в соответствии с рассматриваемой процедурой

$$\vec{a}_1^{(0)} = \frac{\Psi^{(0)}\vec{h}}{\sqrt{\sum_{i,j=1}^{p} \Psi_{i,j}^{(0)} h_i h_j}}. \tag{12.2}$$

Далее вычислим $\Psi_1^{(0)} = \Psi^{(0)} - \vec{a}_1^{(0)}\vec{a}_1^{(0)t}$ и определим нулевое приближение

второго столбца матрицы факторных нагрузок

$$\vec{a}_2^{(0)} = \frac{\Psi_1^{(0)}\vec{h}^*}{\sqrt{\sum_{i,j=1}^{p}\Psi_{1,i,j}^{(0)}h_i^*h_j^*}}, \tag{12.3}$$

где вектор $\vec{h}^*$ состоит из ± 1 и подобран таким образом, что знаменатель последней дроби имеет наибольшее возможное значение.

Определив таким образом все нулевое приближение для матрицы нагрузок A, возвращаемся к исходному итеративному алгоритму, описанному в преамбуле.

Анализ приведенного алгоритма показывает, что для центроидного метода мы имеем дело практически с задачей поиска главных компонент, но координаты соответствующих векторов обязаны быть по модулю равными единице.

В заключение этого раздела отметим, что можно использовать более общий вариант дополнительных условий на матрицу нагрузок – см. [2, с.394], а также, конечно различные сочетания условий и предположений, описанных в двух последних подпунктах.

12.3 Способы оценивания значений латентных факторов

В предыдущем разделе при помощи некоторых итеративных процедур мы описали способы оценивания матрицы факторных нагрузок A и матрицы остаточных дисперсий V. При этом сами факторные нагрузки $A_{i,j}$ можно рассматривать, как коэффициенты корреляции между i-м исходным признаком $X^{(i)}$ и j-м латентным фактором $f^{(j)}$, $i = 1, ..., p$, $j = 1, ..., q$. Само по себе знание этих коэффициентов, конечно, полезно, но оно не решает одну из основных задач факторного анализа – придание численного значения латентным факторам для каждого из наблюдаемых объектов. Неплохо было бы также попытаться вывести формулы, которые по значениям показателей $\vec{X}$ позволяли бы восстанавливать значения латентных факторов для вновь поступающих на изучение объектов. Этими проблемами мы сейчас и займемся.

Ниже мы будем предполагать, что матрицы A и V уже оценены достаточно эффективным образом.

12.3.1 Метод Бартлетта

Перепишем модель факторного анализа в следующем скалярном виде

$$X^{(i)} \;=\; \sum_{j=1}^{q} f^{(j)} A_{i,j} \;+\xi_i \quad i = 1, ..., p$$

и рассмотрим эту запись как модель регрессии, где в качестве значений (уже известных) факторов выступают $A_{i,j},\ j = 1, ..., q$, а в качестве (многомерного) выхода – $\vec{X}$.

Делая обычные предположения о нормальности входящих в наше рассмотрение распределений, можно записать функцию правдоподобия. Если же нормальность отсутствует, то, приводя наши "выходы"к одному масштабу единиц измерения и к общему центру, т.е. нормируя их, мы все равно придем к следующему методу оценки значений латентных факторов как коэффициентов рассматриваемой регрессии: пусть

$$S(f^{(1)}, ..., f^{(q)}) \;=\; \sum_{i=1}^{p} \frac{1}{\sqrt{V_{i,i}}} \left(X^{(i)} - \sum_{j=1}^{q} f^{(j)} A_{i,j} \right)^2 .$$

Тогда значениями $\vec{F}$ будут те $(f^{(1)}, ..., f^{(q)})$, для которых введенная функция достигает своего минимума, т.е.

$$S(\vec{F}) \;=\; \min_{f^{(1)},...,f^{(q)}} S(f^{(1)}, ..., f^{(q)}).$$

Теми же методами, какими обычно работает метод наименьших квадратов, нетрудно найти

$$\vec{F} \;=\; (A^t V^{-1} A)^{-1} A^t V^{-1} \vec{X}.$$

При подстановке сюда последовательно значений векторов наблюдаемых показателей, мы получим оценки латентных факторов для каждого из наблюдаемых объектов. Заметим также, что если сделано предположение о нормальности, то выписанная оценка заодно является оценкой максимального правдоподобия.

12.3.2 Метод Томсона

А теперь предположим, что нам удалось решить поставленную задачу выразить латентные факторы как линейные комбинации исходных показателей. Тогда мы можем записать

$$\vec{F} \;=\; N\vec{X},$$

где N – неизвестная пока матрица, содержащая q строк и p столбцов, оценкой которой и предполагает заняться метод Томсона. Наилучшие значения для элементов этой матрицы можно искать методом наименьших квадратов, исходя из минимизации функции

$$S(N) \;=\; \sum_{k=1}^{n}\sum_{i=1}^{q}\left(f_k^{(i)} - \sum_{j=1}^{q} N_{i,j}X_k^{(j)}\right)^2.$$

Из главы "Регрессионный анализ"нам известно, что решение задачи на минимум функции $S(N)$ выражается в терминах ковариационной матрицы $\mathbf{cov}\vec{X}$ и ковариаций между исходными показателями и латентными факторами $\vec{F}$. Поэтому, хотя значения самих латентных факторов нам неизвестны, решить задачу мы все-таки можем, поскольку эти ковариации нам известны:

$$\mathbf{cov}\vec{X} \;=\; AA^t + V, \quad \mathbf{cov}\vec{F} \;=\; I,$$

$$\mathbf{cov}(X^{(i)},\, f^{(j)}) \;=\; A_{i,j},\;\; i = 1, ..., p,\;\; j = 1, ..., q.$$

Отсюда, вновь применяя известные формулы метода наименьших квадратов, получаем

$$\vec{F} \;=\; (I + A^tV^{-1}A)^{-1}A^tV^{-1}\vec{X}.$$

Это и есть искомая оценка. Сравнение оценок по методу Томсона и методу Бартлетта показывает, что если элементы матрицы $A^tV^{-1}A$ будут достаточно большими, то и оценки значений латентных факторов этими методами будут близки. А этот факт можно понимать так: если величина остаточных дисперсий мала по сравнению с факторными нагрузками, т.е. модель факторного анализа качественно дает близкую к реальности картину, то оба метода дают примерно одно и то же. Предоставляем читателю возможность самому интерпретировать ситуацию, когда оценки по этим двум методам дают существенно различные результаты.

12.4 Пример. Латентные факторы в задачах классификации

Рассмотрим работу описанного выше центроидного метода, а также методов Бартлетта и Томсона оценки значений латентных факторов на

примере следующих данных. В таблице собраны значения суммарной денежной массы США по данным Федерального резервного управления (заимствовано из [13]). В роли изучаемых объектов выступают годы, а в роли показателей, характеризующих эти объекты – денежные массы января, июня и сентября. Эти месяцы выбраны в достаточной степени произвольно и лишь для того, чтобы сделать работу алгоритма более простой. Полные данные приводятся в [13].

Данные о денежной массе США

годы	X_1 январь	X_2 июнь	X_3 сент.
1960	145,7	139,3	141,2
1961	145,2	142,1	143,9
1962	149,7	145,2	145,8
1963	152,5	149,0	151,2
средн.	148,3	143,9	145,5
$\sqrt{\mathbf{D}X_i}$	3,00	3,61	3,66

По этим данным считается матрица корреляций (или ковариационная матрица предварительно нормированных показателей)

$$B = \begin{pmatrix} 1 & 0,86 & 0,90 \\ 0,86 & 1 & 0,98 \\ 0,90 & 0,98 & 1 \end{pmatrix}.$$

Поставим задачу поиска двух латентных факторов, объясняющих, в основном, изменчивость этих показателей.

Первая итерация. Согласно приведенным выше рекомендациям, выберем начальное приближение матрицы остаточных дисперсий и "матрицы косинусов" Ψ:

$$V^{(0)} = \begin{pmatrix} 0,14 & 0 & 0 \\ 0 & 0,14 & 0 \\ 0 & 0 & 0,1 \end{pmatrix},$$

$$\Psi^{(0)} = B - V^{(0)} = \begin{pmatrix} 0,86 & 0,86 & 0,90 \\ 0,86 & 0,86 & 0,98 \\ 0,90 & 0,98 & 0,90 \end{pmatrix}.$$

По формуле (12.2) нетрудно найти нулевое приближение для первого столбца матрицы факторных нагрузок $\vec{a}_1^{(0)} = (0,92;\ 0,95;\ 0,98)^t$. Тогда

$$\vec{a}_1^{(0)}\vec{a}_1^{(0)t} = \begin{pmatrix} 0,85 & 0,87 & 0,90 \\ 0,87 & 0,90 & 0,93 \\ 0,90 & 0,93 & 0,96 \end{pmatrix}\ \Psi_1^{(0)} = \begin{pmatrix} 0,01 & -0,01 & 0 \\ -0,01 & -0,04 & 0,05 \\ 0 & 0,05 & -0,06 \end{pmatrix}.$$

Теперь – наиболее сложно формализуемый этап вычислений. Рассмотрим квадрат знаменателя формулы (12.3). Он имеет в нашем случае вид

$$z(\vec{h}) = 0,01h_1^2 - 0,02h_1h_2 - 0,04h_2^2 + 0,1h_2h_3 - 0,06h_3^2,$$

а значит, достигает своего наибольшего значения 0,03 на допустимых $\vec{h}$ при $h_1 = 1,\ h_2 = h_3 = -1$. Тогда при этих значениях по формуле (12.3) находим приближение для второго столбца $\vec{a}_2^{(0)} = (0,11;\ -0,11\ ;0,06)^t$. Итак, найдены нулевое приближение матрицы нагрузок и первые приближения $\Psi^{(1)}$ матрицы косинусов и $V^{(1)} = B - \Psi^{(1)}$ матрицы остаточных дисперсий:

$$A^{(0)} = \begin{pmatrix} 0,92 & 0,11 \\ 0,95 & -0,11 \\ 0,98 & 0,06 \end{pmatrix},$$

$$\Psi^{(1)} = \begin{pmatrix} 0,86 & 0,86 & 0,91 \\ 0,86 & 0,91 & 0,92 \\ 0,91 & 0,92 & 0,96 \end{pmatrix},\ V^{(1)} = \begin{pmatrix} 0,14 & 0 & -0,01 \\ 0 & 0,09 & 0,06 \\ -0,01 & 0,06 & 0,04 \end{pmatrix}.$$

Вторая итерация. Производя вновь вычисления по (12.2), видим, что значение квадрата знаменателя в наших условиях равно 2,85, а следовательно,

$$\vec{a}_1^{(1)} = \begin{pmatrix} 0,92 \\ 0,94 \\ 0,98 \end{pmatrix},\ \vec{a}_1^{(1)}\vec{a}_1^{(1)t} = \begin{pmatrix} 0,85 & 0,86 & 0,90 \\ 0,86 & 0,88 & 0,92 \\ 0,90 & 0,92 & 0,96 \end{pmatrix},$$

$$\Psi_1^{(1)} = \begin{pmatrix} 0,01 & 0 & 0,01 \\ 0 & 0,03 & 0 \\ 0,01 & 0 & 0 \end{pmatrix}.$$

Квадрат знаменателя (12.3):

$$z(\vec{h}) = 0,01h_1^2 + 0,02h_1h_3 + 0,03h_2^2,$$

и поэтому можно взять $h_1 = h_2 = h_3 = 1$. Тогда

$$\vec{a}_2^{(1)} = \begin{pmatrix} 0,08 \\ 0,12 \\ 0,04 \end{pmatrix} \vec{a}_2^{(1)}\vec{a}_2^{(1)t} = \begin{pmatrix} 0,85 & 0,87 & 0,90 \\ 0,87 & 0,90 & 0,93 \\ 0,90 & 0,93 & 0,96 \end{pmatrix} = \Psi^{(2)},$$

$$V^{(2)} = \begin{pmatrix} 0,15 & -0,01 & 0 \\ -0,01 & 0,10 & 0,05 \\ 0 & 0,05 & 0,04 \end{pmatrix}.$$

Сравнивая эти результаты с предыдущими, можно заметить некоторые признаки стабилизации, но приходится сделать еще две итерации, чтобы процесс действительно стабилизировался. Приведем их краткие результаты.

Третья и четвертая итерации (результаты практически идентичны). $\vec{a}_1^{(i)} = (0,92;\ 0,95;\ 0,98)^t$,

$$\Psi_1^{(i)} = \begin{pmatrix} 0,02 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{pmatrix},$$

откуда видно, что можно взять $h_1 = h_2 = h_3 = 1$, $\sqrt{z(\vec{h})} \approx 0,14$ и, следовательно, $\vec{a}_2^{(i)} = (0,14;\ 0;\ 0)^t$. Наконец,

$$\Psi^{(i+1)}) = \begin{pmatrix} 0,87 & 0,87 & 0,90 \\ 0,87 & 0,90 & 0,93 \\ 0,90 & 0,93 & 0,96 \end{pmatrix},\ V^{(i+1)}) = \begin{pmatrix} 0,13 & -0,01 & 0 \\ -0,01 & 0,10 & 0,05 \\ 0 & 0,05 & 0,04 \end{pmatrix}.$$

Итак, по результатам нашей процедуры можно принять

$$A = \begin{pmatrix} 0,92 & 0,14 \\ 0,95 & 0 \\ 0,98 & 0 \end{pmatrix},\ V = \begin{pmatrix} 0,13 & 0 & 0 \\ 0 & 0,10 & 0 \\ 0 & 0 & 0,04 \end{pmatrix}.$$

Учитывая смысл выписанных параметров, можно сделать вывод, что для достаточно полного описания изменчивости признаков достаточно одного фактора. Это же косвенно подтверждается тем, что в последнем из полученных приближений матрицы остаточных дисперсий элементы, соответствующие взаимодействию ошибок $V_{2,3}$ больше, чем диагональный элемент – скорее всего второй выделяемый нами фактор в основном

состоит из "шума", т.е. не относящихся к сути дела случайных помех. Окончательно,

$$\begin{cases} X'_1 = 0,92f^{(1)} + 0,14f^{(2)} + \xi_1, \ \mathbf{D}\xi_1 = 0,13, \\ X'_2 = 0,95f^{(1)} \ + \ \xi_2, \qquad \mathbf{D}\xi_2 = 0,10, \\ X'_3 = 0,98f^{(1)} \ + \ \xi_3, \qquad \mathbf{D}\xi_3 = 0,04. \end{cases}$$

Здесь штрихами обозначены нормированные признаки. Займемся теперь оцениванием значений латентных факторов. Начнем с метода Бартлетта. Вычисления показывают,

$$V^{-1} \ = \ \begin{pmatrix} 7,69 & 0 & 0 \\ 0 & 10 & 0 \\ 0 & 0 & 25 \end{pmatrix}, \ \Gamma = A^tV^{-1}A = \begin{pmatrix} 39,55 & 0,99 \\ 0,99 & 0,15 \end{pmatrix},$$

$$\Gamma^{-1} \ = \ \begin{pmatrix} 0,03 & -0,20 \\ -0,20 & 7,94 \end{pmatrix}, \ \Gamma^{-1}A^tV^{-1} = \begin{pmatrix} 0 & 0,28 & 0,73 \\ 7,16 & -1,90 & -4,90 \end{pmatrix},$$

откуда

$$f^{(1)} = 0,28X'_2 + 0,73X'_3, \ f^{(2)} = 7,16X'_1 - 1,9X'_2 - 4,9X'_3. \qquad (12.4)$$

Поскольку векторы латентных факторов по методам Томсона и Бартлетта связаны формулой

$$\vec{f}(toms) \ = \ (I + \Gamma)^{-1}\vec{f}(bart),$$

то для метода Томсона мы получаем

$$f^{(1)} = 0,14X'_1 + 0,23X'_2 + 0,61X'_3, \ f^{(2)} = 0,79X'_1 - 0,20X'_2 - 0,52X'_3. \quad (12.5)$$

Для получения формул, позволяющих вычислять значения латентных факторов для объектов, вновь поступающих на изучение, необходимо в (12.4) и (12.5) перейти к абсолютным (не нормированным) значениям показателей по формулам

$$X_i \ = \ \sqrt{\mathbf{D}X_i}X'_i + \bar{X}_i, i = 1, 2, 3.$$

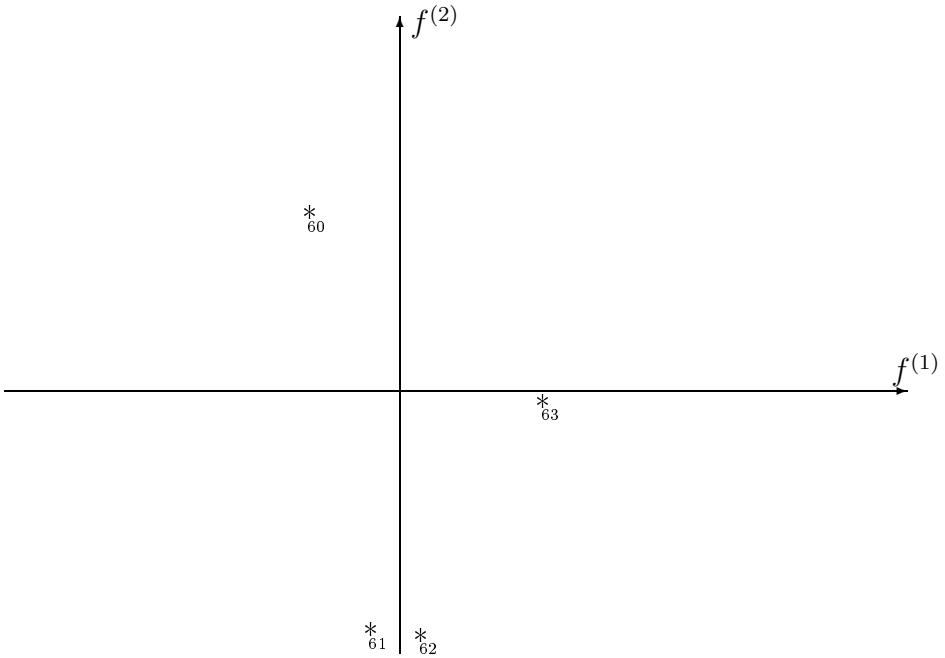
Приведем здесь только результат для (12.4):

$$f^{(1)} = 1,01X_2 + 2.67X_3 + 146,52,$$

$$f^{(2)} = 21.48X_1 - 6,86X_2 - 17.93X_3 + 78,57.$$

Результаты факторного анализа приведены в таблице.

Рис. 12.1: Точки-годы в латентных координатах по методу Бартлетта



Латентные факторы в задаче о денежной массе

метод				Бартлетт		Томсон	
годы	X'_1	X'_2	X'_3	$f^{(1)}$	$f^{(2)}$	$f^{(1)}$	$f^{(2)}$
1960	-0,86	-1,27	-1,18	-1,22	2,04	-1,14	0,20
1961	-1,02	-0,50	-0,44	-0,46	-4,23	-0,53	-0,47
1962	-0,47	0,36	0,07	0,15	-4,43	0,06	-0,48
1963	1,41	1,39	1,55	1,52	-0,14	1,47	0.01

Выделив латентные факторы, мы фактически понизили размерность задачи (в рассматриваемом примере с $p = 3$ до $q = 2$), и это дает нам возможность представить облако точек, задаваемое координатами $\vec{f}$ в q-мерном пространстве и, тем самым, провести классификацию изучаемых объектов по близости точек, их представляющих. Для рассматриваемого примера соответствующий рисунок имеет вид, приведенный ниже. Там изображены точки, получающиеся в координатах, рассчитанных по методу Бартлетта. По методу Томсона картина оказывается в принципе такой же.

На рисунке ясно видно, что 61 и 62 годы близки, а 60 и 63 либо следует отнести в отдельные группы, либо взять в одну (слишком мало для такого вывода находится в нашем изучении объектов – 4).

В заключение этого раздела приведем численные значения нормированных показателей, восстановленных по оцененным значениям латентных факторов (это поможет представить себе количество информации об исходных показателях, сохраняющееся в латентных факторах).

Восстановленные разными методами
исходные показатели в задаче о денежной массе

метод	Бартлетт			Томсон		
годы	X'_1	X'_2	X'_3	X'_1	X'_2	X'_3
60	-0,84	-1,08	-1,19	-1.02	-1.08	-1.12
61	-1,01	-0,44	-0,45	-0,55	-0,50	-0.52
62	-0,48	0,14	0,15	-0,01	0,06	0,06
63	1,38	1,44	1,49	1,35	1,40	1,44

Глава 13

Методы оцифровки нечисловых данных

Выше (в главе "Нечисловые данные") было рассмотрено несколько ситуаций, когда данные, с которыми сталкивается исследователь, имеют качественный категоризованный характер и между категориями требуется установить некоторое отношение близости, сходства. Если мы теперь захотим работать с такими данными одним из разработанных методов, то нам нужно присвоить каждой из категорий такую числовую метку, что система вводимых меток наилучшим возможным образом воплощает упомянутое выше свойство близости. Процесс присвоения таких меток носит название оцифровки. Именно проблемам оцифровки, которая, как становится понятно, связана еще и с тем набором статистических методов и инструментов, что будет в дальнейшем применяться к оцифрованным данным, и посвящена эта глава.

13.1 Случай двух переменных. Анализ соответствий

Объектом нашего первоначального изучения станет случай двух наблюдаемых категоризованных переменных. Обозначим эти переменные $X^{(1)}, X^{(2)}$ и пусть они имеют $m(1), m(2)$ категорий соответственно. Предположим, что в результате n наблюдений мы заполнили таблицу сопряженности $m(1) \times m(2)$ (обозначим ее M), у которой на месте (i, j) расположено количество $n_{i,j}$ экспериментов, в которых одновременно X_1 принимает зна-

чение своей i-й, а X_2 – своей j-й категории, $i = 1, ..., m(1),\ j = 1, ..., m(2)$. Конечно же,

$$\sum_{i=1}^{m(1)} \sum_{j=1}^{m(2)} n_{i,j} \ = \ n.$$

Поделив все элементы матрицы M на n, получим матрицу, элементы которой являются оценками вероятностей того, что двумерный случайный вектор $(X^{(1)}, X^{(2)})$ попадет в "прямоугольник", проекция которого на первую координату есть фиксированная категория $X^{(1)}$, а на вторую – некая категория $X^{(2)}$. Обозначим эту матрицу F.

Нам будет важно иметь оценки условных вероятностей того, что одна из переменных попала, например, в i-ю категорию, если мы знаем, что вторая попала в j-ю категорию. Это будут числа

$$p_{i,j} \ = \ \frac{n_{i,j}}{n(i,.)}, \text{ где } n(i,.) \ = \ \sum_{j=1}^{m(2)} n_{i,j}.$$

Вектор $\vec{p_i} \ = \ (p_{i,1}, ..., p_{i,m(2)})^t$ называется профилем i-й строки, где $i = 1, ...,\ m(1)$. Мы примем на вооружение интерпретацию i-й категории первой наблюдаемой величины, как точки в $m(2)$-мерном пространстве с координатами, равными соответствующему профилю.

Аналогично можно ввести понятие профилей для столбцов, а именно, если

$$n(.,j) \ = \ \sum_{i=1}^{m(1)} n_{i,j},$$

то профилем j-го столбца называется $m(1)$-мерный вектор $\vec{q_j}$ с координатами

$$q_{i,j} \ = \ \frac{n_{i,j}}{n(.,j)}, \quad i = 1, ..., m(1).$$

На этот раз мы интерпретируем категории второй переменной как точки в $m(1)$-мерном пространстве с профилями в качестве их координат. При работе с этими геометрическими интерпретациями мы можем в качестве базовой взять любую из них. Обычно стремятся взять ту, которая имеет меньшую размерность.

Введем также обозначения

$$f_{i,.} \ = \ \frac{n(i,.)}{n} \quad f_{.,j} \ = \ \frac{n(.,j)}{n}.$$

Эти числа представляют собой суммы строк и столбцов матрицы F соответственно.

13.1.1 Расстояние хи-квадрат

Теперь мы можем определить хи-квадрат расстояние между профилями. Так, хи-квадрат расстоянием между i-й и k-й категорией $X^{(1)}$, $i,k = 1,...,m(1)$ называется

$$\chi^2(i,k) \;=\; \sum_{j=1}^{m(2)} \frac{(p_{i,j} - p_{k,j})^2}{f_{.,j}}.$$

Аналогично, хи-квадрат расстояние между j-й и s-й категориями $X^{(2)}$, $j,s = 1,...,m(2)$ – это

$$\chi^2(j,s) \;=\; \sum_{i=1}^{m(1)} \frac{(q_{i,j} - q_{i,s})^2}{f_{i,.}}.$$

Для многих задач обработки данных, таких например, как изучение зависимостей или проблема снижения размерности задачи, поиск оптимальных меток производится именно по критериям, выраженных в терминах расстояния хи-квадрат.

Введенные расстояния особенно полезны благодаря следующему свойству инвариантности

Теорема 20 *Если две категории какой-либо из переменных обладают одинаковыми профилями, то объединение их в одну категорию не приведет к изменению хи-квадрат расстояний между категориями второй переменной.*

Доказательство. Без ограничения общности будем считать, что одинаковыми профилями обладают первая и вторая категории X_1. Это означает, что

$$\frac{n_{1,j}}{n(1,.)} \;=\; \frac{n_{2,j}}{n(2,.)}, \quad j = 1,...,m(2).$$

Пусть $n(2,.) = \alpha n(1,.)$. Тогда при $j = 1,...,m(2)$

$$n_{2,j} = \alpha n_{1,j}, \quad q_{2,j} = \frac{n_{2,j}}{n(.,j)} = \alpha q_{1,j}$$

и

$$f_{2,.} \;=\; \frac{n(2,.)}{n} \;=\; \alpha f_{1,.}.$$

Объединим теперь первую и вторую категории, и будем обозначать новую категорию вновь номером 1, но все относящиеся к ней обозначения пометим штрихами. Очевидно, что при объединении строк ни одна из рассматривавшихся характеристик столбцов не изменится. После объединения

$$f'_{1,.} = \frac{n'(1,.)}{n} = \frac{n(1,.)+n(2,.)}{n} = (1+\alpha)f_{1,.},$$
$$q'_{1,j} = (1+\alpha)q_{1,j}.$$

Если мы рассматриваем хи-квадрат расстояния между t-й и s-й категориями второй переменной, то и до, и после объединения категорий они будут состоять из одних и тех же слагаемых, лишь два слагаемых

$$\frac{(q_{1,t}-q_{1,s})^2}{f_{1,.}} + \frac{(q_{2,t}-q_{2,s})^2}{f_{2,.}} = Z$$

в сумме до объединения заменятся на одно после него. Это слагаемое имеет вид

$$\frac{(q'_{1,t}-q'_{1,s})^2}{f'_{1,.}} = \frac{((1+\alpha)q_{1,t}-(1+\alpha)q_{1,s})^2}{(1+\alpha)f_{1,.}} = \frac{(q_{1,t}-q_{1,s})^2}{f_{1,.}}(1+\alpha).$$

Но, в свою очередь,

$$Z = \frac{(q_{1,t}-q_{1,s})^2}{f_{1,.}} + \frac{(\alpha q_{1,t}-\alpha q_{1,s})^2}{\alpha f_{1,.}} = (1+\alpha)\frac{(q_{1,t}-q_{1,s})^2}{f_{1,.}},$$

и, тем самым, утверждение доказано. Доказательство теоремы завершает то, что если мы объединяем другие категории, то просто достаточно предварительно перенумеровать их так, чтобы они приобрели номера 1 и 2, а если мы объединяем категории второй переменной, то доказательство остается тем же с точностью до очевидных переобозначений.

Каждой из точек, соответствующих категориям $X^{(2)}$ припишем свой вес, равный $f_{.,j}$, $j = 1, ..., m(2)$, и аналогично точкам, интерпретирующим категории $X^{(1)}$, припишем веса $f_{i,.}$, $i = 1, ..., m(1)$. Введем также для удобства матричной записи диагональные матрицы

$$D^{(1)} = \begin{pmatrix} f_{1,.} & 0 & ... & 0 \\ 0 & f_{2,.} & ... & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & ... & f_{m(1),.} \end{pmatrix},$$

$$D^{(2)} = \begin{pmatrix} f_{.,1} & 0 & ... & 0 \\ 0 & f_{.,2} & ... & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & ... & f_{.,m(2)} \end{pmatrix},$$

тогда для расстояний хи-квадрат между векторами $\vec{a}$ и $\vec{b}$ в $m(i)$-мерном пространстве справедливы формулы

$$\chi^2(\vec{a},\vec{b}) = (D^{(i)})^{-1}(\vec{a}-\vec{b})\cdot(\vec{a}-\vec{b}), \quad i=1,2.$$

Более того, если мы введем очевидные поправки, положив

$$\vec{p}_i^* = (D^{(2)})^{-1/2}\vec{p}_i, \quad \vec{q}_j^* = (D^{(1)})^{-1/2}\vec{q}_j$$

при всех i,j, то расстояние хи-квадрат превратится в обычное (евклидово) расстояние в соответствующем многомерном пространстве.

С целью формализовать понятие близости категорий с учетом весов, определим матрицу рассеивания $T^{(1)}$ (для категорий X_1) – квадратную матрицу порядка $m(2)$:

$$T^{(1)} = \sum_{i=1}^{m(1)} f(i,.)\vec{p}_i^*\vec{p}_i^{*t},$$

или, в матричной форме,

$$T^{(1)} = (D^{(2)})^{-1/2}F^t(D^{(1)})^{-1}F(D^{(2)})^{-1/2}. \tag{13.1}$$

Полностью аналогично можно ввести матрицу рассеивания для категорий второй переменной:

$$T^{(2)} = (D^{(1)})^{-1/2}F^t(D^{(2)})^{-1}F(D^{(1)})^{-1/2}.$$

Эти матрицы являются аналогами корреляционных матриц для нечисловых переменных.

Будем искать такое направление $\vec{u}_1$, что проекции точек-строк (т.е. категорий первой переменной) на него разбросаны наиболее сильно. Но такая задача уже решалась нами, когда мы изучали метод главных компонент. Мы знаем, что эта задача эквивалентна нахождению собственного вектора матрицы $T^{(1)}$, отвечающего ее наибольшему собственному числу. Можно было подойти к поставленной задаче с другой стороны, и

искать направление наибольшего разброса для точек-столбцов, что приведет к задаче на собственные числа матрицы $T^{(2)}$.

Пусть

$$\Phi \;=\; (D^{(1)})^{-1/2} F (D^{(2)})^{-1/2},$$

тогда

$$T^{(1)} \;=\; \Phi^t \Phi, \quad T^{(2)} \;=\; \Phi \Phi^t,$$

что означает, что матрицы $T^{(1)}, T^{(2)}$ имеют один и тот же набор ненулевых собственных чисел, поэтому существенно различных картин при двух упомянутых выше подходах к нашей задаче мы не получим.

Обозначим найденные собственные векторы матриц $T^{(1)}, T^{(2)}$, отвечающие наибольшему их общему собственному числу λ, через $\vec{u}_1, \vec{v}_1$ соответственно. Тогда векторы

$$\vec{\psi}_1 \;=\; (D^{(1)})^{-1/2} \vec{v}_1, \;\; \vec{\phi}_1 \;=\; (D^{(2)})^{-1/2} \vec{u}_1$$

коллинеарны векторам, координаты которых суть проекции нормированных профилей строк (соответственно, столбцов) на вектор $\vec{u}_1$ (соответственно, $\vec{v}_1$).

Действительно, например, для строк вектор проекций нормированных профилей – это

$$(D^{(1)})^{-1} F (D^{(2)})^{-1/2} \vec{u}_1 \;=\; (D^{(1)})^{-1/2} \Phi \vec{u}_1 \;=\; \frac{1}{\sqrt{\lambda}} \vec{\psi}_1.$$

Этот факт показывает, что координаты введенных векторов $\vec{\psi}_1, \vec{\phi}_1$ можно использовать, как числовые метки для категорий $X^{(1)}$, $X^{(2)}$ соответственно. Заметим также, что процесс этот можно усложнить, находя второе, третье и т.п. собственные числа матриц $T^{(i)}, i = 1, 2$ – в этом случае каждой из категорий будет соответствовать двух-, трех- и т.д. мерная метка, что позволит изображать категории строчек и столбцов точками в пространствах все более высокой размерности (обычно ограничиваются плоскостью, т.е. находят два собственных вектора и каждой категории сопоставляют двумерную точку). При этом замечательным фактом является то, что и категории $X^{(1)}$, и категории $X^{(2)}$ изображаются в одном пространстве, что дает возможность по крайней мере визуально оценить близость категорий различных переменных друг к другу в смысле их взаимодействия.

Нелишне также будет напомнить, что одной из самых обычных задач, решаемых по таблице сопряженности, с которой мы, собственно, и начинали, является проверка гипотезы независимости.

13.1.2 Оцифровка для задач дискриминантного анализа

Как уже упоминалось, присвоение меток часто связано с решаемой далее задачей обработки данных. Например, дискриминантный анализ решает очень специфические задачи, основанные на принципиально (на первый взгляд) других мерах различия, чем введенное выше расстояние хи-квадрат. Разберем эту ситуацию.

Пусть у нас имеется один качественный (нечисловой) показатель X, а также обучающая выборка для проведения классификации на q классов. Предположим, что наблюдаемый показатель X может попадать в одну их m категорий. Мы поставим в соответствие каждой из категорий метку $c_i,\ i = 1, ..., m$ и обозначим вектор меток через $\vec{c} = (c_1, ..., c_m)^t$. Данные обучающей выборки представим в виде матрицы относительных частот F, имеющей q строк и m столбцов, элементы которой $f_{s,k}$ представляют собой частоты появления k-й категории в s-м классе на элементах обучающей выборки.

Ниже мы увидим, что присвоенные метки будут однозначно определяться этой матрицей F. Как и раньше, пусть

$$f_{s,\cdot} = \sum_{k=1}^{m} f_{s,k}, \quad f_{\cdot,k} = \sum_{s=1}^{q} f_{s,k}.$$

Мы интерпретируем $f_{s,\cdot}$ как (априорную) оценку вероятности очередному объекту попасть в s-й класс, а $f_{\cdot,k}$ – вероятности наугад выбранному объекту иметь k-ю категорию, $s = 1, .., q,\ k = 1, ..., m$.

Основным критерием для отнесения в данный класс для нас будет величина расстояний показателя вновь поступающего на изучение объекта до центров классов, рассчитанных по обучающей выборке, нормированная соответствующей дисперсией. Поэтому, чтобы классификация была более уверенной, надо, чтобы центры этих классов были бы максимально удалены друг от друга, что означает, что они должны иметь максимальный разброс относительно взвешенного среднего этих центров.

Формально. Введем центр $M(s,\vec{c})$ s-го класса, как средневзвешенную величину меток элементов обучающей выборки, принадлежащих этому классу:

$$M(s,\vec{c}) \;=\; \sum_{k=1}^{m} c_k \frac{f_{s,k}}{f_{s,\cdot}}, \quad s=1,..,q$$

и "центр центров"классов

$$M(\vec{c}) \;=\; \sum_{s=1}^{q} M(s,\vec{c}) f_{s,\cdot} \;=\; \sum_{s=1}^{q}\sum_{k=1}^{m} c_k f_{s,k} = \sum_{k=1}^{m} c_k f_{\cdot,k}.$$

Рассчитаем также дисперсии меток в каждом из классов

$$\sigma^2(s,\vec{c}) \;=\; \sum_{k=1}^{m} c_k^2 \frac{f_{s,k}}{f_{s,\cdot}} \;-\; M^2(s,\vec{c}), \;\; s=1,...,q$$

и полную (взвешенную) дисперсию

$$\sigma^1(\vec{c}) \;=\; \sum_{s=1}^{q} \sigma^2(s,\vec{c}) f_{s,\cdot} \;=\; \sum_{k=1}^{m} c_k^2 f_{\cdot,k} - \sum_{s=1}^{q} M^2(s,\vec{c}) f_{s,\cdot}.$$

Тогда нам нужно найти такой набор меток, чтобы величина, оценивающая среднюю взвешенную меру отклонения расстояний $M(s,\vec{c})$ от $M(\vec{c})$ в единицах дисперсии была бы максимальной:

$$Q(\vec{c}) \;=\; \sum_{s=1}^{q} \frac{(M(s,\vec{c}) - M(\vec{c}))^2}{\sigma^2(\vec{c})} f_{s,\cdot} \;\to\; \max_{\vec{c}}. \tag{13.2}$$

Здесь тоже наложим на метки условия нормировки

$$\sum_{k=1}^{m} c_k f_{\cdot,k} \;=\; 0, \quad \sum_{k=1}^{m} c_k^2 f_{\cdot,k} \;=\; 1.$$

Тогда (13.2) перепишется в виде

$$Q(\vec{c}) \;=\; \frac{T(\vec{c})}{1-T(\vec{c})},$$

где

$$T(\vec{c}) \;=\; \sum_{s=1}^{q} M^2(s,\vec{c}) f_{s,\cdot} \;=\; F(D^{(1)})^{-1} F^t \vec{c} \cdot \vec{c}.$$

Здесь и ниже мы вновь используем обозначения

$$D^{(1)} = \begin{pmatrix} f_{1,\cdot} & \dots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \dots & f_{q,\cdot} \end{pmatrix}, \; D^{(2)} = \begin{pmatrix} f_{\cdot,1} & \dots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \dots & f_{\cdot,m} \end{pmatrix},$$

полностью аналогичные введенным выше. Произведя очевидные преобразования, видим, что

$$Q(\vec{c}) \;=\; -1 \;+\; \frac{1}{1 - T(\vec{c})},$$

а это означает, что задачу (13.2) можно заменить на

$$T(\vec{c}) \;\to\; \max$$

в условиях нормировки.

Учтем теперь, что из условий нормировки в матричных обозначениях следует, что

$$\|(D^{(2)})^{-1/2}\vec{c}\| \;=\; 1.$$

Используя обозначение (13.1), запишем

$$T(\vec{c}) \;=\; F(D^{(1)})^{-1}F^t(D^{(2)})^{1/2}\left((D^{(2)})^{-1/2}\vec{c}\right)\cdot(D^{(2)})^{1/2}\left((D^{(2)})^{-1/2}\vec{c}\right) \;=$$

$$=\; T^{(1)}\vec{v}\cdot\vec{v} \;\to\; \max_{\vec{v}},$$

где

$$\vec{v} \;=\; (D^{(2)})^{1/2}\vec{c},$$

причем $\|\vec{v}\| = 1$. Тем самым, мы получили обычную задачу на максимизацию скалярного произведения, решением которой является собственный вектор матрицы $T^{(1)}$, отвечающий ее максимальному собственному числу. После того, как этот вектор $\vec{v}$ будет найден, метки категорий можно найти при помощи формулы

$$\vec{c} \;=\; (D^{(2)})^{-1/2}\vec{v}.$$

Итак, мы видим, что и при таком подходе алгоритм поиска меток совпадает с рассмотренным ранее.

13.2 Более двух переменных

Усложним рассматриваемую задачу. Пусть в нашем рассмотрении находится сразу p переменных $X_1, ..., X_p$, у которых имеется $m(1), ..., m(p)$ категорий соответственно. Обобщая предыдущие рассуждения, можно собрать все данные нашего эксперимента в p-входовую таблицу и организовать изучение и сравнение возникающих в этой таблице профилей. Но полной аналогии все равно не получится: в задаче анализа соответствий мы строили облако точек, каждой из которых приписан некоторый вес, значения которого определялись второй наблюдаемой переменной, после чего прибегали к различным нормировкам, приходя в конце концов к обычным расстояниям между точками и поиску направлений, вдоль которых наши точки наиболее разбросаны. Здесь же единый вес точек ввести не удастся, и при буквальном следовании описанной схеме придется каждой точке приписывать векторный вес, что создаст дополнительные трудности. Поэтому ниже описаны два различных подхода оцифровки категорий многих переменных, основанные на других принципах. Заметим только, что оба эти подхода могут применяться и для случая двух переменных, причем результаты их применения совпадают с результатами работы алгоритма анализа соответствий, описанного в предыдущем разделе, хотя доказывать это мы здесь не будем.

Сначала опишем форму записи исходных данных, которая будет использоваться для обоих подходов. Каждой из наблюдаемых переменных X_j поставим в соответствие вектор-строку $\vec{y}(j) = (y^1(j), ..., y^{m(j)}(j))$, где каждая из координат может принимать лишь значения 0 или 1, причем если наблюдаемая переменная X_j приняла значение из своей k-й категории, то $y^k(j) = 1$, а остальные $y^i(j) = 0$ при $i = 1, ..., m(j),\ i \neq k$. Теперь каждому из n поставленных экспериментов мы поставим в соответствие вектор-строку $\vec{Y}_s, s = 1, ..., n$ длины $m = \sum_{i=1}^{p} m(i)$, которая представляет собой объединение векторов $\vec{y}(j), j = 1, ..., p$, соответствующих этому эксперименту. Заметим, что в каждом из $\vec{Y}_s$ ровно p единиц, причем в каждом из блоков $\vec{y}(j)$ находится ровно одна единица.

Расположив векторы $\vec{Y}_s,\ s = 1, ..., n$ по строкам, мы получим матрицу Y, имеющую n строк и m столбцов, у которой строки соответствуют экспериментам, а столбцы – категориям наблюдаемых переменных. Эта матрица называется бинарной матрицей данных. Сумма всех ее элементов равна np.

13.2.1 Использование бинарной матрицы данных в качестве матрицы соответствий

Один из подходов, который мы рассматриваем, связан с таким приемом: будем считать, что матрица Y, построенная выше, представляет собой матрицу соответствий, к которой применим алгоритм анализа соответствий из предыдущего раздела. Тем самым, мы как бы объединяем p имевшихся переменных в одну с увеличенным числом категорий

$$m \;=\; \sum_{j=1}^{p} m(j)$$

и добавляем еще одну, искусственную переменную X_0, категориями которой являются номера по порядку проводимых экспериментов. Решив задачу оцифровки, мы присвоим числовые метки всем категориям всех p исходных переменных. Правда, при этом и номера экспериментов также окажутся оцифрованными, т.е. мы сможем изображать повторения экспериментов точками в том же пространстве, что и категории $X_1, ..., X_p$. Но,тем не менее, основная задача оказывается решенной.

Аналогом матрицы относительных частот из предыдущего раздела будет

$$F \;=\; \frac{1}{np} Y,$$

сумма всех элементов этой матрицы равна 1, сумма элементов каждой строки, содержащей, как мы выяснили, ровно p единиц, равна $1/n$. Для столбцов матрицы F удобно ввести двойную нумерацию. Мы будем говорить, что работаем со столбцом (k,t), если этот столбец отвечает за k-ю категорию t-й переменной, $k = 1, ..., m(t),\ t = 1, ..., p$.

Пусть теперь $n(k,t)$ – число экспериментов, в которых k-я переменная приобрела значение из t-й категории. Тогда сумма (k,t)-го столбца матрицы F равна $n(k,t)/np$. В обозначениях предыдущего раздела

$$f_{i,\cdot} = \frac{1}{n},\ i = 1, ..., n,\ \ f_{(k,t),\cdot} = \frac{n(k,t)}{np},\ k = 1, .., , m(t),\ t = 1, .., p,$$

Матрицы $D^{(1)}, D^{(2)}$ имеют перечисленные выше элементы в качестве диагональных, в частности $D^{(1)} = (1/n)I$, где I – единичная матрица порядка n.

Нетрудно понять, что хи-квадрат расстояние между k-й категорией переменной X_s и j-й категорией X_r будет задаваться формулой

$$\chi^2((s,k),\ (r,j)) = \begin{cases} 0 & \text{, если } s=r,\ k=j; \\ \frac{n}{n(s,k)} + \frac{n}{n(r,j)} & \text{, если } s=r,\ k\neq j; \\ \frac{n}{n(s,k)} + \frac{n}{n(r,j)} - \frac{2nn((s,k),(r,j))}{n(s,k)n(r,j)} & \text{, если } s\neq r. \end{cases}$$

Здесь $n((s,k),\ (r,j))$ – число экспериментов, в которых попадание X_s в k-ю, а X_r в j-ю категории произошло одновременно.

Обозначив $D\ = npD^{(2)}$ – матрицу, у которой по диагонали расположены $n(k,t)$, также, как в предыдущем пункте, получим

$$\Phi\ =\ (D^{(1)})^{-1/2}F(D^{(2)})^{-1/2} = \frac{1}{\sqrt{p}}YD^{-1/2},$$

тогда

$$T^{(1)}\ =\ \Phi^t\Phi = \frac{1}{p}D^{-1/2}Y^tYD^{-1/2},$$

$$T^{(2)}\ =\ \Phi\Phi^t = \frac{1}{p}YD^{-1}Y^t,$$

где первая из матриц имеет порядок m, а вторая n. У этих матриц будет общий набор ненулевых собственных чисел. Возьмем k первых из них $\mu_1,...,\mu_k$ и пусть $\vec{u}_1,...,\vec{u}_k$ – собственные векторы $T^{(1)}$, отвечающие этим собственным числам и имеющие единичные длины. Тогда k-мерные числовые метки категорий переменных будут являться наборами координат с одинаковыми номерами у k векторов

$$\vec{c}_j\ =\ \sqrt{\mu_j}(D^{(2)})^{-1/2}\vec{u}_j\ =\ \sqrt{\mu_j np}D^{-1/2}\vec{u}_j,\ j=1,...,k. \tag{13.3}$$

Иногда, в силу меньшей размерности задачи, удобней сначала присвоить числовые метки $\vec{z}_j$ строкам матрицы Y (повторениям эксперимента) по формулам

$$\vec{z}_j\ =\ \sqrt{\mu_j}(D^{(1)})^{-1/2}\vec{v}_j\ =\ \sqrt{\mu_j n}\vec{v}_j,$$

где $\vec{v}_j$ – собственный вектор $T^{(2)}$, отвечающий μ_j и имеющий единичную длину. Тогда метки категорий переменных вместо (13.3) рассчитываются по формулам

$$\vec{c}_j\ =\ \frac{1}{\sqrt{\mu_k}}D^{-1}Y^t\vec{z}_j,\ j=1,...,k. \tag{13.4}$$

13.2.2 Максимальные корреляции

Другой подход к присвоению числовых меток состоит в следующем. Поскольку мы поняли, что достаточно присвоить метки повторениям нашего эксперимента (или, если считать, что каждый из экспериментов состоит в изучении очередного объекта, у которого регистрируется p признаков, то метки присваиваются объектам, что выглядит не так непривычно), а затем воспользоваться (13.4) для придания меток категориям, то будем искать такой вектор меток повторений эксперимента $\vec{v} = (v_1, ..., v_n)$, что

$$\sum_{s=1}^{n} v_s = 0, \quad \frac{1}{n}\sum_{s=1}^{n} v_i^2 = 1. \tag{13.5}$$

Условие (13.5) позволит упростить вычисления, к тому же оно гарантирует невозможность присвоения повторениям эксперимента одинаковых меток. При этом будем рассматривать $v_s,\ s = 1, ..., n$ как значения, которые последовательно, при переходе от эксперимента к эксперименту, принимает некая случайная (одномерная) величина v.

Будем подбирать $\vec{v}$ так, чтобы его значения были наиболее сильным образом коррелированы с векторами бинарных случайных величин $\vec{y}(j),\ j = 1, ..., p$:

$$R(\vec{v}) \ = \ \sum_{j=1}^{p} R^2(v, \vec{y}(j)) \ \to \ \max,$$

где $R(v, \vec{y}(j))$ – коэффициент множественной корреляции между случайной величиной v и вектором $\vec{y}(j)$. Если мы обозначим блок матрицы Y, связанный с показателем X_j, $j = 1, .., p$ через Y_j и положим

$$P^{(j)} \ = \ Y_j(Y_j^t Y_j)^{-1} Y_j^t,$$

то

$$R^2(v, \vec{y}(j)) \ = \ \frac{1}{n} P^{(j)}\vec{v} \cdot \vec{v},\ \ j = 1, ..., p.$$

Поясним, что матрица Y_j имеет n строк и $m(j)$ столбцов, и по ее строкам расположены значения $\vec{y}(j)$, которые принимались им последовательно в каждом из экспериментов. Ясно, что матрица $Y_j^t Y_j$ является диагональной, причем по ее диагонали расположены числа $n_k^j,\ k = 1, ..., m(j)$

(суммы k-го столбца матрицы Y_j). Отсюда, как нетрудно понять, следует, что элементы матриц $P^{(j)}$ равны

$$P^{(j)}_{a,b} \;=\; \sum_{i=1}^{m(j)} \frac{y^i_a(j) y^i_b(j)}{n^j_i},\;\; a,b = 1,...,n,\;\; j = 1,..,p.$$

Здесь $y^i_a(j)$ – значение i-й координаты $\vec{y}(j)$, которое она принимает в a-м эксперименте, т.е.

$$y^i_a(j) = \begin{cases} 1, \text{ если в } a\text{-м эксперименте } X_j \text{ попал в } i\text{-ю категорию,} \\ 0, \text{ иначе.} \end{cases}$$

Таким образом, если $P \;=\; \sum_{j=1}^{p} P^{(j)}$, то

$$P_{a,b} \;=\; \sum_{i=1}^{m} \frac{Y_{a,i} Y_{b,i}}{n(.,i)},\;\; a,b = 1,...,n, \tag{13.6}$$

где $Y_{b,i}$ – элемент матрицы Y, $n(.,i)$ – сумма элементов i-го столбца этой матрицы. Здесь не употребляется введенная выше двойная нумерация столбцов, поскольку в формуле (13.6) они суммируются все, и это только сильно загромоздило бы ее.

Итак, во введенных обозначениях,

$$R(\vec{v}) \;=\; \frac{1}{n} P\vec{v} \cdot \vec{v} \;\to\; \max_v ,$$

что, как мы знаем, с учетом (13.5) эквивалентно нахождению единичного собственного вектора P, отвечающего ее максимальному собственному числу. Тем самым, задача решена.

В заключение этого подраздела давайте отметим, что из ранее введенных обозначений

$$pT^{(2)} \;=\; YD^{-1}Y^t \;=\; P.$$

Это равенство нетрудно проверить поэлементным сравнением матрицы $YD^{-1}Y^t$ с (13.6). Поскольку матрицы P и $T^{(2)}$ отличаются лишь числовым множителем, то наборы собственных чисел и собственных векторов у них совпадают, что означает, что найденное сейчас решение полностью совпадает с тем, что было дано в предыдущем подразделе.

13.3 О выборе размерности меток

Сделаем небольшое замечание о том, какой размерности должны быть метки, приписываемые объектам и категориям их характеристик (признаков). Имеется ввиду следующее: проще всего определить числовые метки, для чего при любом из рассмотренных выше подходах достаточно найти максимальное собственное число определенной матрицы и единичный собственный вектор, отвечающий этому собственному числу. Его координаты и будут искомыми числовыми метками для категорий признаков (или для объектов, если задача на поиск меток объектов оказалась решающейся легче, например, имеющей меньшую размерность). Оставшиеся невычисленными метки определяются по формуле типа (13.4).

Но одномерные метки иногда не позволяют уверенно произвести классификацию категорий и объектов, поскольку содержат весьма малую информацию, и большая часть исходной информации при таком грубом линейном приближении оказывается утерянной. Чтобы избежать этого, находят собственные векторы, отвечающие второму, третьему и т.д. собственным числам и из их координат формируют многомерные метки. Например, если найдены два собственных вектора, то двумерная метка i-й категории будет представлять собой пару, в которой на первом месте располагается i-я координата первого собственного вектора, а на втором – i-я координата второго. Ясно, что этот процесс можно продолжать. На какой же размерности можно остановиться, чтобы считать, что потеряно относительно немного информации, и в то же время не слишком усложнять вычисления?

Поскольку речь идет о поиске собственных векторов, отвечающих наибольшим собственным числам, то можно попытаться применить тот же прием, что сработал в методе главных компонент, а именно – изучить долю суммы уже использованных собственных чисел в полной их сумме. При этом сумма всех собственных чисел исследуемой матрицы нам известна – она равна ее следу, т.е. сумме диагональных элементов. В частности, для матриц $T^{(1)}$ и $T^{(2)}$ эта величина равна m/p. Заметим в скобках, что у той из этих двух матриц, которая имеет больший порядок, обязательно имеется некоторое количество нулевых собственных чисел, т.к. в силу доказанного выше, наборы их ненулевых собственных чисел должны совпадать.

К тому же, поскольку сумма элементов любого из профилей равна единице, у каждой из этих матриц имеется максимальное собственное

число, равное 1. Если мы начнем работать с этим собственным числом, то набор получившихся меток окажется тривиальным – все категории получат одинаковые метки. Поэтому фактически работают всегда со вторым по величине собственным числом. Очевидно, все остальные собственные числа не превосходят единицы, поэтому вклад, обеспечиваемый суммой первых q ненулевых собственных чисел в общую сумму собственных чисел, исключая неинформативную единицу, может быть оценен, как

$$\frac{1}{1-\frac{m}{p}}\sum_{j=1}^{q}\lambda_j \;\leq\; \frac{qp}{m-p}\;,$$

что при достаточно большом суммарном числе категорий m оказывается весьма малым. (Напомним, что размерность меток q не может быть большим числом – иначе теряется смысл решения.)

Таким образом, этот подход оказывается малопродуктивным. Чтобы выбрать размерность меток, обычно полагают достаточным рассмотреть все собственные числа, большие, чем среднее из всех положительных собственных чисел. Для того, чтобы снизить количество вычислений, оценим количество l ненулевых собственных чисел матриц $T^{(i)},\ i=1,2$.

Для этого заметим, что, поскольку для каждой из матриц $Y_j,\ j=1,..,p$, объединение которых дает матрицу Y, сумма каждого столбца равна 1, то ранг Y не может быть больше, чем $m-p+1$, т.к. в каждой из Y_j может быть не более $m(j)-1$ линейно независимых столбцов, $j=1,...,p$. Конечно же, в этом рассуждении мы считаем, что $n > m-p+1$, как чаще всего и бывает, иначе более точной оценкой сверху для ранга Y, а именно он и фактически и оценивается, будет число экспериментов n.

В силу изложенных соображений $l \leq m-p$, а значит для средней величины собственных чисел, строго заключенных между 0 и 1, получим оценку

$$\mu \;\geq \frac{\mathbf{tr}T^{(1)}-1}{m-p} \;=\; \frac{1}{p}.$$

Таким образом, последовательное вычисление собственных чисел и, следовательно, увеличение размерности приписываемых меток следует продолжать до тех пор, пока очередное собственное число не окажется меньше, чем $1/p$. Работать с меньшими собственными числами уже не имеет смысла.

13.4 Пример

Для иллюстрации разработанных методов рассмотрим один достаточно условный числовой пример. Пусть среди групп мелких, средних и крупных производственных фирм выделены категории убыточных и доходных. Всего изучалось 25 фирм, данные собраны в таблице (матрица N в ранее введенных обозначениях).

Таблица сопряженности
размера фирмы с доходностью

	мелкие	средние	крупные	всего
убыточные	5	3	3	$n_{1,\cdot} = 11$
доходные	5	3	6	$n_{2,\cdot} = 14$
всего	$n_{\cdot,1} = 10$	$n_{\cdot,2} = 6$	$n_{\cdot,3} = 9$	25

Матрица профилей имеет вид

$$P = \begin{pmatrix} \frac{5}{11} & \frac{3}{11} & \frac{3}{11} \\ \frac{5}{14} & \frac{3}{14} & \frac{6}{14} \end{pmatrix}.$$

Поскольку строк меньше, чем столбцов, то займемся присвоением меток строкам при помощи метода анализа соответствий. Нетрудно вычислить суммарные относительные частоты строк:

$$f_{1,\cdot} = \frac{11}{25}, \quad f_{2,\cdot} = \frac{14}{25}.$$

Для столбцов аналогичные характеристики понадобятся для построения матрицы $D^{(2)}$:

$$f_{\cdot,1} = \frac{10}{25}, \; f_{\cdot,2} = \frac{6}{25}, \; f_{\cdot,3} = \frac{9}{25}.$$

Теперь можно легко найти

$$(D^{(1)})^{-1/2} = \begin{pmatrix} 1,51 & 0 \\ 0 & 1,34 \end{pmatrix}, \; (D^{(2)})^{-1/2} = \begin{pmatrix} 1,58 & 0 & 0 \\ 0 & 2,04 & 0 \\ 0 & 0 & 1,67 \end{pmatrix}.$$

Отсюда получается, что

$$\Phi = (D^{(1)})^{-1/2} F (D^{(2)})^{-1/2} = \begin{pmatrix} 0,47 & 0,37 & 0,30 \\ 0,43 & 0,33 & 0,53 \end{pmatrix},$$

а

$$T^{(1)} = \Phi\Phi^t = \begin{pmatrix} 0,45 & 0,48 \\ 0,48 & 0,57 \end{pmatrix}.$$

Составляя характеристическое уравнение для $T^{(1)}$ – оно будет квадратным – и решая его, находим

$$\lambda_1 \approx 0,99, \quad \lambda_2 \approx 0,03.$$

Найдем также собственные векторы, отвечающие найденным собственным числам. Эти векторы

$$\vec{u}_1 = (0,73;\ 0,68)^t, \ \vec{u}_2 = (0,74;\ -0,66)^t.$$

Таким образом, мы имеем возможность приписать строкам двумерные метки: метка убыточных имеет вид (0,73;0,74), метка доходных – (0,74;-0,66).

Для вычисления меток столбцов используем связь между собственными векторами матриц $T^{(2)}$ и $T^{(1)}$:

$$\vec{v}_1 = \frac{1}{\sqrt{\lambda_1}}\Phi^t\vec{u_1} = \begin{pmatrix} 0,63 \\ 0,49 \\ 0,58 \end{pmatrix},$$

и, аналогично,

$$\vec{v}_2 = (0,35;\ 0,35;\ 0,75)^t.$$

Таким образом, двумерная метка для малых фирм (0,63;0,35), для средних – (0,49;0,35) и для крупных – (0,58;-0,75). Соответствующий рисунок приводится немного ниже. Из него видно, что большие фирмы близки к доходным, а малые близки к средним, и находятся ближе к убыточным, чем к доходным, хотя оба расстояния достаточно велики. Обратите внимание также на тот факт, что первые координаты столбцов и первые координаты строк почти равны между собой. Дело в том, что эти координаты построены по наибольшим собственным числам матриц T, которые равны 1, и метки по первым координатам должны быть просто одинаковыми (одна для всех строк и одна для всех столбцов), а те различия, которые у нас получились, вызваны очень неточным характером вычислений.

Чтобы ликвидировать эту оплошность, на рисунке показаны проекции точек на вторую координату (помечены теми же буквами).

Приведем также вкратце результаты второго способа, использующего бинарную матрицу данных. Почему обработка этим методом приведена не столь подробно, становится ясно при одном взгляде на матрицу Y, которая приведена на отдельной странице, ведь в ней 5 столбцов и 25 строчек! Матрицы Y_1 и Y_2 отделены в этой таблице двойной линией. При составлении Y для простоты предполагалось, что в верхнюю левую клетку нашей таблицы сопряженности попали первые 5 из изученных нами фирм и т.п. В принципе, это конечно же могло быть не так, но всегда можно добиться этого апостериорной перенумерацией объектов. Можно вычислить

$$Y^tY = \begin{pmatrix} 11 & 0 & 5 & 3 & 3 \\ 0 & 14 & 5 & 3 & 6 \\ 5 & 5 & 10 & 0 & 0 \\ 3 & 3 & 0 & 6 & 0 \\ 3 & 6 & 0 & 0 & 9 \end{pmatrix}.$$

В этой матрице просматриваются два диагональных блока, по диагонали которых расположены суммы всех остальных элементов соответствующей строки и еще два блока, каждый из которых воспроизводит первоначальную таблицу сопряженности. Оказывается, такой вид матрица Y^tY будет иметь всегда. Она носит название матрицы Берта. Далее, строя диагональную матрицу D, диагональные элементы которой совпадают с диагональными элементами матрицы Берта, вычислим

$$T^{(2)} = \frac{1}{2} D^{-1/2} Y^t Y D^{-1/2}.$$

Получим

$$T^{(2)} = \begin{pmatrix} 0,50 & 0 & 0,24 & 0,18 & 0,15 \\ 0 & 0,50 & 0,21 & 0,16 & 0,27 \\ 0,24 & 0,21 & 0,50 & 0 & 0 \\ 0,18 & 0,16 & 0 & 0,50 & 0 \\ 0,15 & 0,27 & 0 & 0 & 0,50 \end{pmatrix},$$

после чего при помощи математического пакета **Mathlab** вычислим собственные числа и собственные векторы этой матрицы.

Получим $\lambda_1 = 1,\ \lambda_2 = 0,58,\ \lambda_3 = 0,5, \lambda_4 = 0,47,\ \lambda_5 = 0$. При этом, как мы знаем, первое собственное число использовать не имеет смысла. Как было показано в предыдущем разделе, имеет смысл привлекать только собственные векторы, отвечающие собственным числам, не меньшим,

Таблица 13.1: Бинарная матрица в задаче о сопряженности доходности и размера фирм

1	0	1	0	0
1	0	1	0	0
1	0	1	0	0
1	0	1	0	0
1	0	1	0	0
1	0	0	1	0
1	0	0	1	0
1	0	0	1	0
1	0	0	0	1
1	0	0	0	1
1	0	0	0	1
0	1	1	0	0
0	1	1	0	0
0	1	1	0	0
0	1	1	0	0
0	1	1	0	0
0	1	0	1	0
0	1	0	1	0
0	1	0	1	0
0	1	0	0	1
0	1	0	0	1
0	1	0	0	1
0	1	0	0	1
0	1	0	0	1
0	1	0	0	1

чем $1/p = 0,5$ в нашем случае, поэтому используем второе и третье собственные числа. (Заметим в скобках, что для нашего случая среднее ненулевых собственных чисел, строго меньших единицы, равно 0,517, так что можно было обойтись одномерной меткой, не используя λ_3, что вполне соответствует результатам анализа соответствий выше.)

Два собственных вектора, соответствующие λ_2 и λ_3 и имеющие единичную длину, равны

$$\vec{c}_2 = \begin{pmatrix} 0,53 \\ -0,47 \\ 0,34 \\ 0,25 \\ 0,57 \end{pmatrix}, \quad \vec{c}_3 = \begin{pmatrix} 0 \\ 0 \\ -0,68 \\ -0,73 \\ -0.02 \end{pmatrix}.$$

Теперь мы можем выписать двумерные метки для всех 5 категорий:

Метки категорий объектов
в задаче о сопряженности размеров и доходности

категория	метка		обозначение
убыточные	0,53	0	у
доходные	-0,47	0	д
малые	0,34	-0,68	м
средние	0,25	-0,73	с
крупные	0,57	-0,02	к

Для определения меток объектов (повторений эксперимента) будем пользоваться формулами

$$\vec{z}_j = \frac{1}{p\sqrt{\lambda_j}} Y \vec{c}_j, \quad j = 2, 3,$$

которые дают 25 меток строк матрицы Y. К счастью, среди них много одинаковых, т.к. объекты, попавшие в одну клетку таблицы сопряженности, получают одинаковые метки. Эти метки приводятся в таблице.

Метки объектов (повторений эксперимента)
в задаче о сопряженности размеров и доходности

категория	метка		обозначение
малые убыточные	0,58	-0,48	му
средние убыточные	0,50	-0,52	су
крупные убыточные	0,71	-0,01	ку
малые доходные	-0,07	-0,48	мд
средние доходные	-0,15	-0,52	сд
крупные доходные	0,06	-0,02	кд

Соответствующие точки с приведенными обозначениями приведены на рисунке. Различия, полученные по отношению к методу анализа соответствий, легко объяснить тем, что во время того исследования мы использовали малоинформативное наибольшее собственное число, а второе (0,03) оказалось слишком малым, чтобы картинка оказалась похожей на правду – ведь мы знаем, что стоило работать лишь с собственными числами, не меньшими 0,5. Итак, рисунок, полученный сейчас, следует признать более удовлетворительно описывающим реальную картину.

13.5 Случай смешанных данных

Выше мы предполагали, что данные о всех объектах носят чисто качественный, нечисловой характер. Но в практических задачах нередко можно встретить сочетание характеристик (см. пример с травматологическим отделением больнице во вводной части главы, посвященной экспертным оценкам и прочим нечисловым данным). Поэтому здесь рассмотрен один из способов присвоения числовых меток категориям объектов, у которых некоторые из показателей числовые.

Предположим, что наши показатели пронумерованы так, что первые из них $X^{(1)}, ..., X^{(q)}$ – качественные (нечисловые) показатели, причем $X^{(i)}$ имеет $m(i)$ категорий, $i = 1, ..., q$. Будем писать $X_s^{(i)} \in (k)$, если в s-м эксперименте показатель $X^{(i)}$ принял значение из своей k-й категории. Остальные показатели – $X^{(q+1)}, ..., X^{(p)}$ являются числовыми.

Данные наблюдений собраны в таблицу из n строк и p столбцов, в каждой строке стоят данные наблюдений, полученные в очередном эксперименте над всеми показателями. В первых q столбцах расположены условные обозначения категорий, в которые попал соответствующий этому столбцу признак при проведении эксперимента, его номер совпадает с номером текущей строки, в следующих $p - q$ столбцах – значения, принимаемые числовыми показателями. Задача состоит в замене условных

обозначений категорий числовыми метками. Оказывается, для разных методов последующей обработки данных наилучшие возможные метки должны присваиваться по разному, но всегда они связаны с ковариационной матрицей показателей. При этом всегда удобно предполагать, что присваиваемые метки имеют нормированный характер – это упрощает записываемые формулы и гарантирует от присвоения разным категориям одинаковых меток. Поясним, что имеется ввиду.

Пусть k-й категории i-го признака присвоена метка $c^i_k,\ k = 1, ..., m(i)$, $i = 1, ..., q$. Условия нормировки имеют вид

$$\sum_{s=1}^{n} c^i_{r(s)} \ = \ 0, \ \ \frac{1}{n}\sum_{s=1}^{n}\left(c^i_{r(s)}\right)^2 \ = \ 1$$

при каждом $i = 1, ..., q$. Здесь $r(s)$ есть номер категории, который принял i-й признак в s-м эксперименте, т.е. определяется соотношением $X^{(i)}_s \in (r(s))$.

Если после присвоения числовых меток мы хотим заняться исследованием зависимостей между показателями или сокращением размерностей, то нужно подбирать числовые метки, максимизирующие величину

$$K^2 \ = \ \sum_{i=1}^{p-1}\sum_{j=i+1}^{p} \rho^2(X^{(i)},\, X^{(j)}),$$

Очевидно, что когда в выписанной сумме переменная суммирования i становится больше q, коэффициенты корреляции перестают зависеть от присвоенных меток, поэтому речь может идти только о максимизации

$$Q \ = \ \sum_{i=1}^{q-1}\sum_{j=i+1}^{q} \rho^2(X^{(i)},\, X^{(j)}) \ + \ \sum_{i=1}^{q}\sum_{j=q+1}^{p} \rho^2(X^{(i)},\, X^{(j)}).$$

Обозначим первую из двойных сумм через Q_1, а вторую через Q_2.

Пусть $\vec{c}^{\,i} \ = \ (c^i_1, ..., c^i_{m(i)})^t$ – вектор меток категорий i-го показателя, матрица $F(i,j)$, имеющая $m(i)$ строк и $m(j)$ столбцов – нормированная таблица сопряженности i-го и j-го показателя, т.е. на месте (k,s) этой матрицы располагается число экспериментов, в которых $X^{(i)} \in (k)$, $X^{(j)} \in (s)$ одновременно, деленное на общее число экспериментов:

$$F_{k,s}(i,j) \ = \ \frac{n(k,s)}{n}, \quad k = 1, ..., m(i), \ s = 1, ..., m(j).$$

Обозначим также через n_k^i число тех экспериментов, в которых $X^{(i)} \in (k)$, $k = 1, ..., m(i)$, $i = 1, ..., q$, и построим диагональные матрицы D_i, с диагональными элементами $n_1^i, ..., n_{m(i)}^i$, $i = 1, ..., q$. Вычислим при каждом наборе $j = (q+1), ..., p$ и $k = 1, ..., m(i)$, $i = 1, ..., q$

$$\bar{X}_k^{(j)}(i) \;=\; \frac{1}{n_k^i} \sum_{s: X_s^{(i)} \in (k)} X_s^{(j)} -$$

среднее значение числового показателя $X^{(j)}$ по тем экспериментам, в которых качественный показатель $X^{(i)}$ попал в свою k-ю категорию. Составим из рассчитанных средних величин вектор

$$\bar{X}^{(j)}(i) \;=\; (\bar{X}_1^{(j)}(i), ..., \bar{X}_{m(i)}^{(j)}(i)).$$

Тогда

$$Q_1 \;=\; \sum_{i=1}^{q-1} \sum_{j=i+1}^{q} \left(F(i,j)\vec{c}^{\,j} \cdot \vec{c}^{\,i} \right)^2,$$

$$Q_2 \;=\; \sum_{i=1}^{q} \sum_{j=q+1}^{p} \left(D_i \vec{c}^{\,i} \cdot \bar{X}^{(j)}(i) \right)^2.$$

Вычисляя частные производные Q по c_k^i, получаем систему уравнений для определения меток:

$$\sum_{i=1}^{q-1} \sum_{j=i+1}^{q} (F(i,j)\vec{c}^{\,j})_k F(i,j)\vec{c}^{\,j} \cdot \vec{c}^{\,i} + \sum_{i=1}^{q} \sum_{j=q+1}^{p} (D_i \bar{X}^{(j)}(i))_k D_i \vec{c}^{\,i} \bar{X}^{(j)}(i) \;=\; 0,$$

где k пробегает номера всех категорий всех качественных признаков (их общее количество равно $m(1) + ... + m(q)$).

Выписанная система решается при помощи итеративных процедур. Подробности можно прочитать в [11, глава 12].

Наконец, коротко рассмотрим задачу оцифровки для задач дискриминации. Пусть обучающая выборка содержит (возможно многомерные) данные о качественных признаках X и числовых признаках Y некоторого набора объектов, а также сведения о том, какому из имеющихся q классов принадлежал этот объект. Обозначим через m число категорий качественного признака.

Так же, как это было проделано в подпункте 13.1.2, вычислим для каждого из классов его центр $M(s, \vec{c})$ для качественного признака, а

также средние значения $\bar{Y}_s,\ s = 1, .., , q$ для числового признака. Будем предполагать, что и метки, и значения Y нормированы, а значит, "центр центров"в терминологии 13.1.2 имеет нулевые координаты.

Для каждого класса вычислим оценку ковариационной матрицы

$$\Sigma(s, \vec{c}) \ = \ \mathbf{cov}(X, Y, s),\ \ s = 1, ..., q$$

и средневзвешенную общую ковариационную матрицу

$$\Sigma(\vec{c}) \ = \ \sum_{s=1}^{q} \Sigma(s, \vec{c}) f_{s,.}.$$

Тогда метки $\vec{c}$ можно искать исходя из максимизации критерия

$$Q(\vec{c}) \ = \ \sum_{s=1}^{q} \| \Sigma^{-1}(\vec{c}) \left(M(s, \vec{c}); \ \bar{Y}_s \right) \|^2 f_{.,s}.$$

Рис. 13.1: Изображение категорий величины и доходности. Метод анализа соответствий.

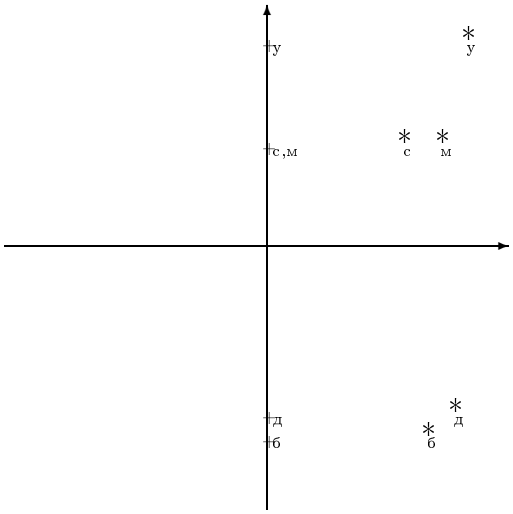
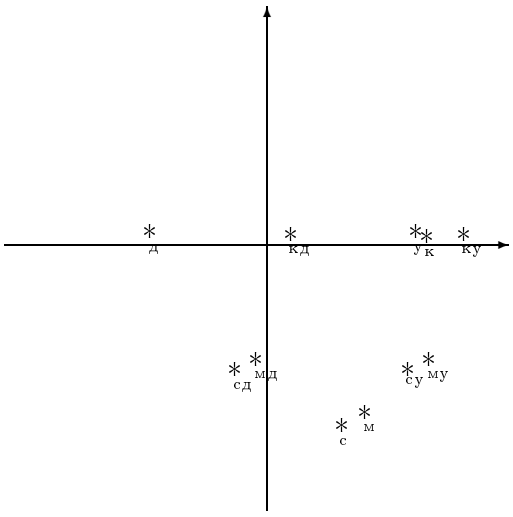


Рис. 13.2: Множественный анализ соответствий таблицы сопряженности в задаче о доходности и размере.



Глава 14

Многомерное шкалирование

В обработке данных, особенно в последние десятилетия, выделилось особое направление, которое правильнее всего было бы назвать анализом данных, понимая под этим скорее некоторый качественный процесс, чем определенные вычислительные процедуры. Выше мы уже сталкивались с проблемой анализа данных – например, визуального в задачах классификации. Задачи подобного рода решались нами в предыдущей главе, когда для осознания близости определенных качественных, а не числовых объектов, мы ассоциировали эти объекты с некоторыми числами или векторами, сводя тем самым задачу к ранее решенной.

Многомерное шкалирование ориентировано в рамках этого подхода в основном на придание наглядной структуры данным, полученным в результате некоторого эксперимента, т.е. решает задачу, близкую к оцифровке качественных данных. Но здесь, в отличие от ранее рассмотренных методов, в качестве исходных данных рассматривается матрица близостей определенных объектов или категорий одного объекта. Близости эти задаются в некоторой условной шкале – балльной относительно некоторого образца или порядковой, т.е. для изучаемых отношений "похожести"объектов задаются их ранги в порядке убывания этой "похожести". Затем при помощи определенных приемов эти близости переводятся в расстояния. Задача многомерного шкалирования считается успешно решенной, если удалось изобразить все данные точками в пространстве относительно небольшой размерности так, чтобы с точки зрения оцененных расстояний геометрическая структура экспериментальных данных подверглась бы минимальным возможным изменениям. Естественно, термин "геометрическая структура"нуждается в уточнении и может в принципе

быть формализован по разному.

14.1 Подготовка данных для многомерного шкалирования

Как уже было сказано, на входе любого из алгоритмов многомерного шкалирования должна фигурировать матрица различий или матрица сходств определенных объектов. Пусть, например у нас имеются различные объекты $X_1, ..., X_K$. Эти объекты могут быть самыми различными – заводы, университеты, автомобили, кандидаты на пост президента и т.п. Нам необходимо подготовить квадратную матрицу D порядка K, на месте (i, j) в которой будет стоять мера различия i-го и j-го объекта (или хотя бы его оценка). При этом мы хотим, чтобы в конце концов наша матрица была бы матрицей расстояний между объектами, т.е. для элементов матрицы D должен быть выполнен набор аксиом расстояния:

$$D_{i,j} \geq 0, \qquad D_{i,i} = 0, \qquad i, j = 1, ..., K; \tag{14.1}$$

$$D_{i,j} = D_{j,i}, \qquad i, j = 1, ..., K; \tag{14.2}$$

$$D_{i,j} + D_{j,k} \geq D_{i,k}, \quad i, j, k = 1, ..., K. \tag{14.3}$$

При этом довольно типичной является ситуация, когда матрица, оценивающая различия объектов, строится по результатам опросов экспертов, например, на основе шкал сравнений в графической или категоризованной форме. И если при таком подходе свойство (14.1) как правило, выполнено, то даже свойство (14.2), не говоря уже о (14.3), оказывается нарушенным. Чтобы понять, почему это происходит, достаточно представить себе ответ на вопросы "Похожи ли Москва и Петербург?"и "Похожи ли Петербург и Москва?"на геометрической шкале сравнений. Даже если эти вопросы задаются предельно добросовестному эксперту подряд (он имеет возможность сверить свои ответы), и он хочет, чтобы ответы совпадали, все равно добиться точного совпадения ответов практически невозможно. Поэтому, чтобы добиться выполнения условия (14.2) (симметричности), чаще всего полагают $D_{i,j}$ и $D_{j,i}$ равными полусумме соответствующих оценок различия между i-м и j-м и между j-м и i-м объектами:

$$D_{i,j} \;=\; D_{j,i} \;=\; \frac{1}{2}(\delta(i,j) + \delta(j,i)).$$

Если для $\delta(i,j),\ i,j = 1,...,K$ выполнены все свойства, кроме (14.3), то можно поступить следующим образом. Введем в рассмотрение

$$c \ = \ \max_{k,i,j}(\delta(k,j) - \delta(k,i) - \delta(i,j)).$$

После этого полагаем

$$D_{i,j} \ = \ \begin{cases} 0\ , & \text{если } i = j, \\ \delta(i,j) + c, & \text{иначе,} \end{cases} \ i,j = 1,...,K.$$

Лемма 8 *Если $\delta(i,j),\ i,j = 1,...,K$ удовлетворяли (14.1 – 14.2), то $D_{i,j}$ будут удовлетворять (14.1 – 14.3).*

Доказательство. Заметим сначала, что если вдруг оказалось, что $c \leq 0$, то это означает, что для произвольного набора индексов s,i,j выполнено

$$\delta_{s,j} \ - \ \delta_{s,i} \ - \ \delta_{i,j} \ \leq \ 0,$$

что совпадает с неравенством треугольника (14.3), а значит все наши выкладки будут излишни. Но, на самом деле, c всегда неотрицательно. Действительно, возьмем такой набор индексов, в котором $s = i$. Тогда, в силу того, что (14.1) выполнено, то $\delta_{s,s} = 0$ и

$$c \ \geq \ \delta_{s,j} - \delta_{s,s} - \delta_{s,j} \ = \ 0.$$

Поэтому $D_{i,j} \geq \delta_{i,j} \geq 0$ и (14.1) выполнено для введенных $D_{i,j}$.

Пусть теперь удалось найти такой набор индексов s,i,j, для которого неравенство треугольника нарушается, т.е.

$$D_{s,j} \ > \ D_{s,i} \ + \ D_{i,j},$$

что по определению означает, что

$$\delta_{s,j} \ > \ \delta_{s,i} \ + \ \delta_{i,j} \ + \ c,$$

или (для того же набора индексов), что

$$\delta_{s,j} - \delta_{s,i} - \delta_{i,j} \ > \ c,$$

что входит в противоречие с определением c. Лемма доказана.

Другой вариант построения матриц различий возникает тогда, когда каждый из объектов задается некоторым набором своих числовых характеристик, признаков, т.е. объекты заданы таблицей

Данные для построения матрицы различий

	Признаки		
Объекты	1	...	p
1	$v_{1,1}$	...	$v_{1,p}$
$\vdots$		$\vdots$	
K	$v_{K,1}$	...	$v_{K,p}$

Признаки $v_{i,j}$, сопутствующие i-му объекту могут иметь самое разное содержание, но именно по их значениям мы и должны судить о сходстве и различии объектов. При этом (по крайней мере, сейчас) мы предполагаем, что все имеющиеся у нас p признаков одинаково важны для формирования различий и сходств объектов. Тогда можно определить

$$D_{i,j} = \sqrt{\sum_{s=1}^{p}(v_{i,s} - v_{j,s})^2}, \quad i,j = 1,...,K.$$

При таком определении все требуемые свойства будут выполнены.

Иногда для придания большего "равноправия"признакам их центрируют и нормируют. Матрицу признаков будем называть стандартизованной по столбцам, если среднее значение элементов каждого столбца равно 0, а среднеквадратическое отклонение равно 1. Стандартизация по столбцам проводится обычным методом – достаточно от элементов j-го столбца отнять его среднее $\bar{v}_{.,j}$ и разделить на среднеквадратическое отклонение этого столбца $S_{.,j}$. В точности так же можно провести и стандартизацию матрицы признаков по строкам. Соответствующие характеристики обозначим $\bar{v}_{i,.}$ и $S_{i,.}$.

14.2 Модель Торгерсона

Итак, предположим, что построена матрица различий D, удовлетворяющая свойствам (14.1 – 14.3). В модели, предложенной классиком теории многомерного шкалирования У.С.Торгерсоном, принимается предположение о том, что возможно по этой матрице построить точки $x_1, ..., x_K$ в пространстве некоторого относительно небольшого числа измерений q таким образом, чтобы

$$D_{i,j} = \sqrt{\sum_{s=1}^{q}(x_{i,s} - x_{j,s})^2}, \ i,j = 1,...,K. \tag{14.4}$$

Здесь $x_{i,j}$ – j-я координата точки x_i. В модели предполагается также, что для каждого $s = 1, ..., q$ среднее значение соответствующих координат строящихся точек равно 0:

$$\sum_{i=1}^{K} x_{i,s} \;=\; 0.$$

Для работы алгоритма Торгерсона, который позволит нам по матрице различий определить конкретные числовые значения координат, предварительно необходимо перейти от матрицы различий к новой вспомогательной матрице Δ^* с двойным центрированием, т.е. организовать некое преобразование, в результате которого у построенной Δ^* окажется, что среднее значение элементов каждой строки и каждого столбца равно нулю. Предлагается следующая формула для элементов строящейся матрицы:

$$\delta^*_{i,j} \;=\; -\frac{1}{2}(D^2_{i,j} - D^2_{i,\cdot} - D^2_{\cdot,j} + D^2_{\cdot,\cdot}),\;\; i,j = 1, ..., K. \qquad (14.5)$$

Здесь $D^2_{i,\cdot}, D^2_{\cdot,j}, D^2_{\cdot,\cdot}$ – средние квадратов i-й строки, j-го столбца и всех элементов матрицы D соответственно, т.е.

$$D^2_{i,\cdot} = \frac{1}{K}\sum_{j=1}^{K} D^2_{i,j},\;\; D^2_{\cdot,j} = \frac{1}{K}\sum_{i=1}^{K} D^2_{i,j},$$

$$D^2_{\cdot,\cdot} \;=\; \frac{1}{K^2}\sum_{i=1}^{K}\sum_{j=1}^{K} D_{i,j}.$$

Лемма 9 *Матрица Δ^*, элементы которой определены в (14.5), является матрицей с двойным центрированием.*

Доказательство этой леммы практически очевидно. Вычислим, например, среднее по ее i-й строке.

$$\frac{1}{K}\sum_{j=1}^{K}\delta^*_{i,j} \;=\; -\frac{1}{2}\left(D^2_{i,\cdot} - \frac{1}{K}KD^2_{i,\cdot} - \frac{1}{K}\sum_{j=1}^{K} D^2_{\cdot,j} + D^2_{\cdot,\cdot}\right).$$

Для того, чтобы закончить доказательство, достаточно лишь отметить, что

$$\frac{1}{K}\sum_{j=1}^{K} D^2_{\cdot,j} \;=\; D^2_{\cdot,\cdot}$$

и понять, что рассуждение, проводимое по столбцам, дословно повторяет только что проведенное.

Теорема 21 *Если элементы матрицы D удовлетворяют (14.4) и элементы матрицы Δ^* построены по формулам (14.5), то*

$$\delta^*_{i,j} \;=\; \sum_{s=1}^{q} x_{i,s}x_{j,s}, \quad i,j = 1,...,K,$$

или, в матричной записи,

$$\Delta^* \;=\; XX^t. \tag{14.6}$$

Это и есть теорема Торгерсона, на которой основан его ставший классическим алгоритм многомерного шкалирования.

Доказательство. Очевидно, что без потери общности и без нарушения (14.4) можно считать, что значения $x_{i,j}$ центрированы, т.е.

$$\sum_{i=1}^{K} x_{i,j} \;=\; 0, \quad j = 1,...,q.$$

Из (14.4) немедленно получаем, что

$$D^2_{i,j} \;=\; \sum_{s=1}^{q} x^2_{i,s} + \sum_{s=1}^{q} x^2_{j,s} - 2\sum_{s=1}^{q} x_{i,s}x_{j,s}. \tag{14.7}$$

Вычислим теперь средние по j от обеих частей последнего равенства:

$$D^2_{i,\cdot} \;=\; \sum_{s=1}^{q} x^2_{i,s} + \frac{1}{K}\sum_{j=1}^{K}\sum_{s=1}^{q} x^2_{j,s} - \frac{2}{K}\sum_{j=1}^{K}\sum_{s=1}^{q} x_{i,s}x_{j,s}.$$

Заметим, что последнее слагаемое здесь равно 0 в силу условий центрированности, т.к. может быть переписано в виде

$$-\frac{2}{K}\sum_{s=1}^{q} x_{i,s}\left(\sum_{j=1}^{K} x_{j,s}\right) \;=\; 0.$$

Таким образом, мы получаем

$$D^2_{i,\cdot} \;=\; \sum_{s=1}^{q} x^2_{i,s} \;+\; \sum_{s=1}^{q} x^2_{\cdot,s}, \;\; i = 1,...,K, \tag{14.8}$$

где

$$x^2_{\cdot,s} \;=\; \frac{1}{K}\sum_{j=1}^{K} x_{j,s}.$$

Теперь получим аналогичное выражение для $D_{.,j}$. Действуя в точности так же, но производя усреднение (14.7) по i, получим

$$D^2_{.,j} \;=\; \sum_{s=1}^{q} x^2_{.,s} \;+\; \sum_{s=1}^{q} x^2_{j,s}, \;\; j = 1, ..., K. \tag{14.9}$$

Наконец, усредняя обе части (14.7) одновременно по i, j, получим

$$D_{.,.} \;=\; \frac{1}{K}\sum_{i=1}^{K}\sum_{s=1}^{q} x^2_{i,s} \;+\; \frac{1}{K}\sum_{j=1}^{K}\sum_{s=1}^{q} x^2_{j,s} \;-\; \frac{2}{K^2}\sum_{i=1}^{K}\sum_{j=1}^{K}\sum_{s=1}^{q} x_{i,s}x_{j,s}.$$

Последняя тройная сумма в этой формуле, как и в двух предыдущих случаях, равна нулю в силу центрированности координат. Поэтому

$$D_{.,.} \;=\; 2\sum_{s=1}^{q} x^2_{.,s}. \tag{14.10}$$

Подставим теперь (14.7), (14.8), (14.9) и (14.10) в (14.5). Получим после простого приведения подобных

$$\delta^*_{i,j} \;=\; -\frac{1}{2}\left(-2\sum_{s=1}^{q} x_{i,s}x_{j,s}\right),$$

что и заканчивает доказательство теоремы.

Таким образом, для того, чтобы решить задачу многомерного шкалирования, достаточно найти решение матричного уравнения (14.6), и найденная матрица X, содержащая K строк и q столбцов в своих строках содержит наборы координат точек в q-мерном пространстве, которые нужно сопоставить изучаемым объектам.

14.2.1 Стресс-критерий

Алгоритм Торгерсона, который будет описан ниже, ставит своей задачей поиск таких точек, изображающих наши объекты, для которых геометрическая структура данных наименее искажена в смысле максимизации следующего критерия оптимальности:

$$K_q(X) \;=\; \sqrt{\frac{\sum_{i=1}^{K}\sum_{j=1}^{K}(D^*_{i,j} - D_{i,j})^2}{\sum_{i=1}^{K}\sum_{j=1}^{K} D^2_{i,j}}}. \tag{14.11}$$

Здесь $D^*_{i,j}$ – расстояния между точками, изображающими наши объекты в q-мерном пространстве (строками матрицы X) , а $D_{i,j}$ – истинные расстояния между точками-объектами. Понимать все это легче всего на примере ситуации, когда исходная информация задается таблицей (см. конец предыдущего раздела), а значит, наши объекты изначально отображаются точками в p-мерном пространстве, $p > q$. Если же входная информация задана в другом виде, то "истинные"значения $D_{i,j}$ недоступны измерению, но, организуя некий итеративный процесс, шаг за шагом осуществляющий построение лучшей геометрической конфигурации, мы можем по формуле (14.11) отслеживать изменения, происходящие при смене итераций.

Критерий оптимальности (14.11) принято называть стресс-критерием. Он предложен Дж.Краскалом. В некоторых исследованиях можно встретить разновидность стресс-критерия, задаваемую формулой

$$K_q(X)' \;=\; \sqrt{\frac{\sum_{i=1}^{K}\sum_{j=1}^{K}(D^*_{i,j}-D_{i,j})^2}{\sum_{i=1}^{K}\sum_{j=1}^{K}(D_{i,j}-D_{.,.})^2}},$$

где

$$D_{.,.} \;=\; \frac{1}{K^2}\sum_{i=1}^{K}\sum_{j=1}^{K}D_{i,j}-$$

среднее арифметическое оцениваемых расстояний. Эту формулу, следуя [14], мы будем называть "стресс, формула 2".

14.3 Алгоритм Торгерсона

В этом разделе приведен алгоритм многомерного шкалирования. На входе этого алгоритма – матрица различий $\Delta \;=\; (\delta_{i,j})$. Если вместо различий на начальном этапе мы имели коэффициенты корреляции $r_{i,j} \;=\; \rho(X^{(i)},\, X^{(j)})$, то можно положить

$$\delta_{i,j} \;=\; \sqrt{1-r_{i,j}}, \quad i,j=1,...,K.$$

Пусть также методом двойного центрирования, описанным выше, мы по матрице Δ вычислили матрицу скалярных произведений Δ^* и извлекли из нее q главных компонент $x_{i,j},\; j=1,...,q,\; i=1,...,K$. Зададимся также малым числом ε, которое будет представлять собой требуемую точность вычислений.

Шаг 1. Положим $c = 0$, $\hat{x}_{i,j}^{(c)} = x_{i,j}$, $\delta_{i,j}^{(c)} = \delta_{i,j}$, $St = 0$. Число c – номер "прохода"алгоритма, число St – текущее значение стресс-критерия. К шагу 2.

Шаг 2 (Нормирование). Вычислим

$$\hat{d}_{i,j}^{(c)} = \sqrt{\sum_{s=1}^{q} \left(\hat{x}_{i,s}^{(c)} - \hat{x}_{j,s}^{(c)}\right)^2}; \; Q = \sqrt{\sum_{i=1}^{K-1} \sum_{j=i+1}^{K} (\hat{d}_{i,j}^{(c)})^2},$$

$$d_{i,j}^{(c)} = \hat{d}_{i,j}^{(c)}/Q, \; x_{i,j}^{(c)} = \hat{x}_{i,j}^{(c)}/Q, \quad i,j = 1,...,K,$$

$$S(c) = \sqrt{\sum_{i=1}^{K-1} \sum_{j=i+1}^{K} \left(\delta_{i,j}^{(c)} - d_{i,j}^{(c)}\right)^2}.$$

К шагу 3.

Шаг 3. (Достигнута ли нужная точность?) Если $c > 0$ и $|S(c) - St| < \varepsilon$, то конец алгоритма. Иначе полагаем $St = S(c)$ и переходим к шагу 4.

Шаг 4. (Упорядочивание). Положим $l = 1$. Фактически этот шаг выполняется лишь при $c = 0$. Алгоритм устроен так, что при остальных c наши данные окажутся упорядоченными автоматически. Поэтому если $c > 0$, то можно сразу перейти к шагу 5. Упорядочим пары (i,j) по возрастанию различий $\delta_{i,j}$. Если для каких-нибудь пар

$$\delta_{i,j}^{(c)} = \delta_{pt}^{(c)} \text{ и } d_{i,j}^{(c)} < d_{p,t}^{(c)},$$

то пара (i,j) должна предшествовать (p,t). Если же равны и различия, и расстояния то порядок пар произвольный. Результатом работы этого шага будет таблица из $\mathbf{C}_K^2$ строк и трех столбцов, в первом из них будет обозначение пары, во втором – различия элементов этой пары в порядке возрастания, в третьем – расстояния между ними. Последние располагаются по возрастанию, вообще говоря, только в случае равных различий. Работа алгоритма будет продолжаться до тех пор, пока и столбец расстояний не окажется упорядоченным по возрастанию. При этом некоторые из расстояний придется изменить. Номер "прохода"алгоритма по сформированной сейчас таблице и есть число l. К шагу 5.

Шаг 5. Если соседние в таблице числа $d_{i,j}^{(c)}$ имеют равные величины, объединяем соответствующие им пары индексов в блок. Если равных расстояний нет – каждую пару объявляем самостоятельным блоком. К шагу 6.

Шаг 6. Подсчитаем число блоков. Пусть их $m(l)$, и k-му блоку соответствует расстояние $d(k)$, $k = 1, ..., m(l)$ (напомним, что в пределах одного блока расстояния равны между собой). Положим $k = 1$. К шагу 7.

Шаг 7. Сравниваем k-й и $(k+1)$-й блоки. Если $d(k) < d(k+1)$, то к шагу 8. Иначе сливаем два этих блока в один, присваиваем ему номер $k+1$ и для него определяем

$$d(k+1) \;=\; \frac{1}{2}(d(k) + d(k+1)).$$

К шагу 8.

Шаг 8. Полагаем $k = k+1$. Если $k < m(l)$, то к шагу 7 (продолжаем спуск по таблице), иначе к шагу 9.

Шаг 9. Если на шаге 7 происходило слияние каких-либо блоков, то $l = l+1$ и к шагу 6, иначе к шагу 10.

Шаг 10. (Новые координаты). Пересчитаем

$$\delta_{i,j}^{(c+1)} \;=\; d(k) \text{ при } (i,j) \text{ в } k-\text{м блоке}, \;\; k = 1, ..., m(l),$$

$$\hat{x}_{i,j}^{(c+1)} \;=\; x_{i,j}^{(c)} \;-\; \frac{1}{K}\sum_{k=1}^{K}\left(1 - \frac{\delta_{i,k}^{(c+1)}}{d_{i,k}^{(c)}}\right)(x_{i,k}^{(c)} - x_{j,k}^{(c)})$$

при всех значениях $i, j = 1, ..., K$. Если $d_{i,k}^{(c)} = 0$, то соответствующее слагаемое в последнюю сумму не включаем, или, что то же самое, полагаем это слагаемое равным 0. Увеличиваем c: $c = c+1$. К шагу 2.

По окончании работы алгоритма числа $x_{i,j}^{(c)}$, $i, j = 1, ..., K$ и есть наилучшие оценки координат изображающих изучаемые объекты в q-мерном пространстве с заданной точностью ε. Шаг 10 называют метрическим этапом алгоритма. Формулы пересчета координат, приводимые там, выписаны из условия уменьшения величины стресса, формула 1, самым радикальным образом. Вывод этих формул мы не приводим, но отметим, что при изменении критерия оптимальности нужно в описанном алгоритме заменить только эти формулы. Шаги с 3 по 9 принято называть неметрическим этапом алгоритма Торгерсона, и существуют варианты этого этапа, отличные от описанного.

14.4 Методы шкалирования индивидуальных различий

Предположим, что у нас имеются M экспертов, высказывающих свои отношения о различиях между K различными объектами. Конечно же, можно свести мнения экспертов к одному усредненному мнению и решать задачу многомерного шкалирования для такой ситуации, как мы уже делали выше. Но иногда требуется также наглядно изобразить отличия во мнениях каждого из опрашиваемых экспертов вместе с некоторым "объективным"или средним мнением. В этом случае говорят о шкалировании индивидуальных различий (ШИР).

Предположим, что для каждого из экспертов имеется такая диагональная матрица $W^{(s)}$ с элементами $w_t^{(s)},\ t = 1, ..., K$ по диагонали, что индивидуальная матрица координат отображаемых объектов, $X^{(s)}$, содержащая K строк и q столбцов, вычисляется по формуле

$$X^{(s)} \ = \ W^{(s)}X, \quad s = 1, ..., M. \tag{14.12}$$

Здесь матрица X – матрица "объективных"координат K точек в q-мерном пространстве, отображающих среднее мнение всей группы экспертов в целом. Диагональный элемент $w_t^{(s)}$ матрицы $W^{(s)}$ представляет собой тот вес, который s-й эксперт приписывает t-й координате. Эти элементы называют весами важности. При увеличении $w_t^{(s)}$ различие между i-м и j-м объектами по t-й координате вносит все больший вклад в общую оценку различия этих объектов с точки зрения s-го эксперта. Тем самым, в рамках этой модели мы сводим различия во мнениях экспертов лишь к различным оценкам вкладов координат в общую близость объектов.

По заданным матрицам различий объектов для каждого из экспертов $\Delta^{(s)},\ s = 1, .., M$ требуется восстановить матрицу "объективных"координат X и матрицы весов важности $W^{(s)},\ s = 1, ..., M$, по крайней мере, получить их удовлетворительные оценки. Естественно, что при этом индивидуальные матрицы координат $X^{(s)},\ s = 1, ..., M$ находятся по формуле (14.12), что позволяет изобразить положения всех объектов в q-мерном пространстве по мнению каждого из экспертов (например, точками разных цветов) и рассмотреть положения этих индивидуально шкалированных точек по отношению к "объективным"точкам, построенным по матрице X.

Процесс нахождения $X,\ W^{(s)},\ s = 1, ..., M$ называют подгонкой взве-

шенной евклидовой модели. Такое название эта модель получила потому, что в роли расстояния между i-м и j-м объектами с точки зрения s-го эксперта выступает

$$\delta_{i,j}^{(s)} = \sqrt{\sum_{k=1}^{q} w_k^{(s)}(x_{i,k} - x_{j,k})^2},\ i,j = 1,...,K,\ s = 1,...,M,$$

т.е. к обычному евклидову расстоянию навешиваются веса важности на каждую из координат.

Приведем схему алгоритма шкалирования индивидуальных различий.

На входе алгоритма у нас есть матрицы скалярных произведений $\Delta^*(s)$, $s = 1,...,M$ для каждого эксперта, полученные из индивидуальных матриц различий методом двойного центрирования, ε – малое число, определяющее точность вычислений, q – размерность пространства изображений.

Шаг 0. Положим $p = 0$ (номер текущей итерации). К шагу 1.

Шаг 1. Вычислим

$$\Delta^* = \frac{1}{M}\sum_{s=1}^{M}\Delta^*(s) -$$

среднюю матрицу скалярных произведений. Пусть $\delta_{i,j}(p)$ – элементы этой матрицы. Выделим из нее при помощи одного из алгоритмов предыдущего раздела матрицу $X(p)$ – очередное приближение матрицы "объективных"координат изображений объектов. В этой матрице k строк и q столбцов. К шагу 2.

Шаг 2. Строим матрицу $A(p)$, состоящую из M строк и K^2 столбцов. Каждый столбец ее соответствует паре объектов, каждая строка – эксперту. На s-й ее строке в столбце, соответствующей паре (i,j) стоит элемент $\delta_{i,j}^*(s)$ матрицы $\Delta^*(s)$, $i,j = 1,...,K,\ s = 1,...,M$.

Строим еще одну матрицу $B(p)$, в которой q строк и K^2 столбцов. Каждый столбец соответствует некоторой паре объектов, каждая строка – некоторой координате. На строке t в столбце, соответствующей паре (i,j) расположено число $x_{i,t}(p)x_{j,t}(p)$ (произведение соответствующих элементов матрицы $X(p)$), $i,j = 1,...,K,\ s = 1,...,M$.

Вычислим вспомогательную матрицу, элементы которой будут представлять собой квадраты оценок весов важности:

$$W^2(p) = (B(p)B^t(p))^{-1}B(p)A^t(p).$$

При этом оценки весов важности на p-й итерации имеют вид

$$w_t^{(s)}(p) \;=\; \sqrt{W_{t,s}^2(p)}.$$

К шагу 3.

Шаг 3. Построим здесь вновь две матрицы. Сначала матрицу $C(p)$, содержащую K строк и MK столбцов. Каждая ее строка соответствует какому-то эксперту, а каждый столбец – паре (эксперт, объект). Элемент в i-й строке в столбце, соответствующем паре (s,j), имеет вид $\delta_{i,j}^*(s)$, $i,j = 1,...,K, s = 1,...,M$. Матрица $D(p)$ имеет q строк и MK столбцов. Каждая строка соответствует координате, а каждый столбец – паре (эксперт, объект). Элемент в строке k в столбце, соответствующему s-му эксперту и j-му объекту равен $W_{k,s}^2(p)x_{j,k}(p)$ (произведению соответствующих элементов матриц $W^2(p)$ и $X(p)$).

По этим матрицам вычисляем новое приближение "объективной"матрицы координат

$$X(p+1) \;=\; C(p)D^t(p)(D(p)D^t(p))^{-1}.$$

Вычислим также исправленные скалярные произведения

$$\delta_{i,j}^*(s,p) \;=\; \sum_{k=1}^{q} w_{k,s}^2 x_{i,k}x_{j,k},\; i,j = 1,...,K,$$

сформируем из них новые матрицы $\Delta^*(s),\; s = 1,...,M$ и пусть

$$f \;=\; \sum_{i=1}^{K}\sum_{j=1}^{K}\sum_{s=1}^{M}\left(\delta_{i,j}^*(s) - \delta_{i,j}^*(s,p)\right)^2.$$

К шагу 4.

Шаг 4. Если $f < \varepsilon$, то конец алгоритма. Иначе $p = p+1$ и к шагу 1.

Глава 15

Понятие о временном ряде

15.1 Общие положения

Мы, как всегда, будем иметь дело с некоторым количеством n наблюдений. Только если раньше мы наблюдали случайную величину (или вектор), как правило, в неизменных условиях, то теперь в наши наблюдения вмешивается независимо от нас изменяющийся параметр, который мы будем называть временем. При изменении этого параметра распределение наблюдаемой величины может изменяться. В той теории, которую мы будем строить, предполагается, что состояние некоторой величины измеряется через равные промежутки времени, в результате чего получаем набор значений (выборку) объема $n - u_1, ..., u_n$, при этом время принимало значения $t_1, ..., t_n$ с одним и тем же шагом $h = t_i - t_{i-1},\ i = 2, ..., n$. Обычно принято считать – и мы также для упрощения формул примем эту точку зрения – что $t_i = i$, т.е. измерения происходят в моменты $1, 2, ..., n$. Эту цепочку наблюдений и условимся называть временным рядом.

Каждый может легко представить себе примеры временных рядов: урожайность зерновой культуры на единицу площади, измеряемую год за годом, среднемесячные температуры июля и января, также определяемые ежегодно, количества единиц продукции, производимой цехом за смену, количество населения в стране (периодичность измерения – примерно раз в 10 лет). Нетрудно также придумать примеры, в которых в роли "времени"выступает совсем иной параметр. Так бывает, например, если мы возьмемся измерять засоренность единицы площади поля сор-

няками по мере удаленности от края поля, или будем выяснять процент людей, которым известно, кто такой С.Зубакин по мере удаления места их проживания от Республики Алтай. Заранее договоримся, что во всех этих ситуациях параметр, через равные промежутки изменения которого производятся замеры величины u, мы будем называть временем.

Еще раз подчеркнем, что промежутки времени между соседними моментами измерений заранее задаются и в процессе эксперимента не меняются. Дело в том, что есть много похожих задач, в формулировке которых участвует понятие "время", но, в силу случайного характера его значений в этих задачах, они решаются совсем другими методами – методами, имеющими дело со случайными процессами (процессы восстановления или обслуживания) или методами, относящимися к компетенции теории катастроф. Такими являются, например, задачи изучения периодичности возникновения эпидемий, прогноз наступления того или иного события по изменению неких близких параметров, задачи определения числа появлений какого-то события а течение данного промежутка времени и т.п.

После изучения некоторого количества практических примеров временных рядов становится понятно, что имеет смысл (возможно, достаточно условно, о чем речь несколько ниже) выделить следующие составляющие, определяющие значения наблюдаемой величины u:

1. тренд, или систематическое изменение вместе с изменением времени;

2. случайные колебания относительно небольшой амплитуды вокруг тренда, величина которых в принципе может быть связана со значением тренда;

3. эффект сезонности;

4. чисто случайная или нерегулярная составляющая.

Эффект сезонности понять из перечисленных, пожалуй, проще всего. Название это произошло от того, что вся наша жизнь протекает на фоне сезонных изменений в природе, и говорить о том, что в течение одного и того же срока летом или зимой некий процесс протекает с одинаковой интенсивностью, как правило, нельзя.

Понятие тренда можно представить себе как некое детерминированное изменение, например, среднего значения какой-либо величины со

временем. Так, средняя урожайность пшеницы с гектара с годами увеличивается из-за изобретения новых удобрений и внедрения новых прогрессивных способов хозяйствования и технологий обработки земли. То же самое можно сказать о населении, а вот число гужевых повозок в хозяйствах имеет обратную тенденцию (отрицательный тренд). Иногда тренд может иметь достаточно сложную структуру, например, периодическую. Особенно это касается длительных процессов, в частности, природных. Бывает и так, что то, что мы принимаем за тренд, при "смене масштаба"оказывается лишь проявлением сезонности. Так, сегодня мы говорим об общем потеплении климата и склонны рассматривать это как тренд в изменении среднегодовой температуры, но при переходе к изучению изменения климата в течение геологических эпох мы увидим, что потепления и похолодания на Земле сменяют друг друга довольно регулярно, и с этой точки зрения мы наблюдаем лишь сезонное явление.

Из последнего (довольно длинного) рассуждения должно стать ясно, что некое сходство и взаимное перетекание содержания имеется между всеми перечисленными составляющими временного ряда, и отнесение изменения наблюдаемого значения к одной из этих составляющих довольно условно. Что ж, в задачах статистики явления, сходные этому, встречаются довольно часто, и, как мы знаем, обычно получают разрешение в рамках конкретной задачи, причем весьма субъективным образом. Поэтому, обозначив эти проблемы, больше не будем к ним возвращаться.

15.2 Критерии случайности

В этом разделе мы рассмотрим критерии, позволяющие определить наблюдаемый временной ряд, как имеющий только чисто случайную, нерегулярную составляющую. Естественно, необходимо задаться вопросом об альтернативе высказываемой гипотезе. При ясном понимании, какая именно альтернатива подразумевается – наличие тренда и какого именно вида, наличие сезонности, или то и другое, – можно выбрать наиболее подходящий к ситуации критерий, которых имеется огромное количество.

Здесь мы рассмотрим только критерии, которые

- никак не используют вид альтернативы;

- никак не зависят от распределения случайной составляющей временного ряда;

- требуют относительно небольшого числа вычислений и имеют просто описываемый алгоритм.

Еще раз подчеркнем, что, употребляя слова "случайный временной ряд", мы в этом разделе будем понимать, что наши числа $u_1, ..., u_n$ представляют собой результаты независимых наблюдений над одной и той же случайной величиной – выборку в традиционном значении этого слова. Отсутствие изменения распределения наблюдаемой величины с течением времени, т.е. при переходе от наблюдения к наблюдению, как раз и гарантирует отсутствие любого тренда и сезонности.

15.2.1 Подсчет экстремальных точек

Мы условимся говорить, что в точке k временной ряд $u_j,\ j = 1, ..., n$ имеет пик, если одновременно $u_{k-1} < u_k,\ u_{k+1} < u_k$ и имеет яму, если значение u_k меньше обеих соседних. Будем говорить, что k – экстремальная точка ряда, если в этой точке пик или яма.

Как видно из данных определений, в число экстремальных точек не могут попасть начальная и конечная точки (1 и n). Если в ряду попадается несколько равных значений, причем все они больше (или, соответственно, меньше) их окружающих, то все эти точки мы будем воспринимать как одну экстремальную точку. Интервал между двумя экстремальными точками называют фазой. При этом под словами "длина фазы"условимся понимать количество членов ряда между экстремальными точками. Так, например, если соседние экстремальные значения временного ряда – u_2 и u_5, то длина фазы, заключенной между ними, равна 2.

Целью нашего исследования будет изучать распределение пиков в (чисто) случайном временном ряде. Распределение ям будет, очевидно, таким же. Действительно, если ряд случаен и получен наблюдением над случайной величиной ξ, то при замене ее на $-\xi$ любой пик перейдет в яму и наоборот. Но, по сделанным выше допущениям, мы ищем критерии, не зависящие от распределения ξ!

Наличие экстремального значения определяется сравнением между собой трех последовательных значений временного ряда. Но три раз-

личных фиксированных значения, например, a, b, c всего могут расположиться 6 способами, из которых 2 образуют монотонную цепочку (возрастающую и убывающую), поэтому вероятность того, что в цепочке трех подряд взятых значениях случайного временного ряда будет иметься экстремальная точка, рана $4/6$ или $2/3$.

Для трех последовательных значений $u_i, u_{i+1}, u_{i+2},\ i = 1, ..., (n-2)$ определим

$$X_i \ = \ \begin{cases} 1, \text{ если среди них есть экстремальная точка,} \\ 0 \text{ иначе.} \end{cases}$$

Тогда число экстремальных точек

$$e \ = \ \sum_{i=1}^{n-2} X_i.$$

Нетрудно вычислить,

$$\mathbf{M}e \ = \ \sum_{i=1}^{n-2} \mathbf{M}X_i \ = \ \frac{2(n-2)}{3}.$$

Кроме того,

$$\mathbf{M}e^2 \ = \ \mathbf{M}\left(\sum_{i=1}^{n-2} X_i^2 \ + \ 2 \sum_{i=1}^{n-3} \sum_{j=i+1}^{n-2} X_i X_j \right). \tag{15.1}$$

Заметим, что $X_i^2 = X_i$, а также отметим, что X_i и X_j независимы, когда $j > i + 2$, поэтому

$$\mathbf{M}X_i X_j \ = \ \mathbf{M}X_i \, \mathbf{M}X_j \ = \ \frac{4}{9}, \quad j > i + 2.$$

Пусть $a < b < c < d$. Рассмотрим четыре последовательных члена временного ряда $u_i, u_{i+1}, u_{i+2}, u_{i+3}$. Будем считать, что их величины содержатся в множестве $\{a, b, c, d\}$. Величина $X_i X_{i+1}$ принимает лишь значения 0 или 1, причем равна 1 только если экстремальные точки будут находиться на втором и третьем местах. Нетрудно убедиться (например, полным перебором всех 24 вариантов), что только при 10 перестановках значений a, b, c и d это будет так. Вот эти перестановки:

$$\begin{array}{ccccc} acbd & badc & cadb & dacb & adbc \\ bcad & cbda & dbca & bdac & cdab \end{array}$$

Отсюда следует, что

$$\mathbf{M}X_iX_{i+1} \;=\; \mathbf{P}(X_iX_{i+1}=1) \;=\; \frac{5}{12}.$$

Аналогично, для расчета $\mathbf{M}X_iX_{i+2}$ мы должны перебрать все 120 перестановок пяти последовательных членов нашего временного ряда и выяснить, в скольких из них экстремальные точки находятся на втором и четвертом месте одновременно (только в этом случае наше произведение X_iX_{i+2} отлично от нуля). Таких перестановок оказывается 54. Отсюда

$$\mathbf{M}X_iX_{i+2} \;=\; \frac{54}{120} \;=\; \frac{9}{20}.$$

Перепишем (15.1) с учетом сделанных замечаний в виде

$$\begin{aligned}\mathbf{M}e^2 \quad &= \textstyle\sum_{i=1}^{n-2}\mathbf{M}X_i + 2\sum_{i=1}^{n-3}\mathbf{M}X_iX_{i+1} + \\ &+ 2\textstyle\sum_{i=1}^{n-4}\mathbf{M}X_iX_{i+2} + 2\sum_{i=1}^{n-5}\sum_{j=i+3}^{n-2}\mathbf{M}X_i\mathbf{M}X_j \;= \\ &= \tfrac{2(n-2)}{3} + \tfrac{5(n-3)}{6} + \tfrac{9(n-4)}{10} + \\ &+ \tfrac{4(n-4)(n-5)}{9} \;=\; \tfrac{40n^2-144n+131}{90}.\end{aligned}$$

При этом, конечно же, предполагалось, что $n>5$, и был учтен тот факт, что в двойной сумме осталось $\mathbf{C}_{n-4}^2$ слагаемых. Из последней формулы и рассчитанного ранее среднего значения e выводим

$$\mathbf{D}e \;=\; \frac{16n-29}{90}.$$

Из этого соотношения с привлечением обычной центральной предельной теоремы следует

Теорема 22 *Распределение*

$$\frac{e-\mathbf{M}e}{\sqrt{\mathbf{D}e}} \;=\; \frac{3e-2n+4}{\sqrt{16n-29}}\sqrt{10}$$

сходится к стандартному нормальному при $n\to\infty$.

Из приведенной теоремы следует такой алгоритм проверки гипотезы о случайном характере временного ряда: определим число e экстремальных точек временного ряда по n имеющимся в нашем распоряжении значениям. После этого вычислим

$$t \;=\; \frac{3e-2n+4}{\sqrt{16n-29}}\sqrt{10} \qquad (15.2)$$

и сравним его с двусторонней критической точкой стандартного нормального распределения уровня ϵ (это то же самое, что его квантиль $t_{1-\epsilon/2}$ уровня $1-\epsilon/2$). Если

$$|t| \ < \ t_{1-\epsilon/2},$$

то гипотезу о чисто случайном характере временного ряда можно принять с вероятностью $1-\epsilon$.

15.2.2 Распределение длины фазы

Чтобы обнаружить фазу длины d, необходимо обнаружить цепочку длины $d+3$, имеющую вид

$$u_i < u_{i+1} > u_{i+2} > ... > u_{i+d+1} < u_{i+d+2}$$

(фаза убывания) или такую, в которой все неравенства заменены на противоположные (фаза возрастания). Рассмотрим произвольные фиксированные $d+3$ значения, расположенные в порядке возрастания. Если мы возьмем любые два из этих значений кроме первого и последнего и поставим их на первое и последнее места, то получим фазу возрастания длины d. Очевидно, фаза возрастания требуемой длины, если мы не трогаем крайние элементы, может по заданным значениям быть построена только описанным образом. Следовательно, имеется столько же способов образования фазы возрастания, сколько способов выбора произвольных двух элементов из $d+1$. При этом каждый из двух выбранных элементов может располагаться как в начале,так и в конце строящейся цепочки. Таким образом, всего вариантов получилось $\mathbf{A}^2_{d+1} = d(d+1)$.

Теперь перейдем к подсчету вариантов, разрешающих перестановки крайних членов цепочки. Мы можем переставить первый элемент на последнее место, а любой элемент, кроме второго, на первое. При этом получается еще $d+1$ вариант фазы возрастания длины d. Можно также последний поставить на первое место, а любой, за исключением предпоследнего, на последнее, что дает еще $d+1$ вариант. Но при этом вариант, когда первый становится последним, а последний первым, посчитан дважды, а значит, его нужно вычесть. Таким образом, при заданных $d+3$ значениях мы получили

$$d(d+1) \ + \ 2(d+1) \ - \ 1 \ = \ d^2+3d+1$$

фаз возрастания длины d. Конечно же, можно построить точно такое же количество цепочек, в которых мы обнаружим фазу убывания длины d. Итак, вероятность обнаружить требуемую фазу в цепочке из $d+3$ последовательных значений равна

$$p_{d+3}(d) \;=\; \frac{2(d^2+3d+1)}{(d+3)!}.$$

Аналогично проведенным выше рассуждениям, введем для произвольных $d+3$ последовательных значений индикатор наличия в них фазы длины d:

$$I_i \;=\; \begin{cases} 1, \text{ если среди } d+3 \text{ значений, начиная с } i\text{-го есть} \\ \quad \text{фаза длины } d, \\ 0 \text{ иначе.} \end{cases}$$

Тогда, в силу того, что индикатор принимает только значения 0 или 1, для произвольного i

$$\mathbf{M}I_i \;=\; \mathbf{P}(I_i = 1) \;=\; p_{d+3}(d).$$

Отсюда, т.к. общее число фаз требуемой длины равно сумме всех I_i, $i = 1, ..., (n-d-2)$, получаем, что среднее количество фаз длины d для чисто случайного временного ряда равно

$$N(d) \;=\; \sum_{i=1}^{n-d-2} \mathbf{M}I_i \;=\; \frac{2(n-d-2)(d^2+3d+1)}{(d+3)!}. \qquad (15.3)$$

Исходя из (15.3), а также учитывая тот факт, что если данный участок временного ряда "занят"в формировании фазы длины d, то он не может принимать участие в формировании фаз других длин, можно вывести формулу среднего количества N фаз всех возможных длин:

$$N \;=\; \sum_{d=1}^{n-3} N(d).$$

Достаточно заметить, что

$$(n-d-2)(d^2+3d+1) \;=\; nd(d+3) + n - d(d+2)(d+3) - d(d+2)$$

и расписать формулу для N в виде

$$N \;=\; 2n\left(\sum_{d=1}^{n-3} \frac{1}{(d+3)!} + \sum_{d=1}^{n-3} \frac{1}{(d+1)!} - 2\sum_{d=1}^{n-3} \frac{1}{(d+2)!}\right) \;-$$

$$- 2\left(2\sum_{d=1}^{n-3}\frac{1}{(d+2)!} - \sum_{d=1}^{n-3}\frac{1}{(d+1)!} - \sum_{d=1}^{n-3}\frac{1}{(d+3)!}\right),$$

чтобы, после стандартных сокращений, получить формулу

$$N \;=\; 2\left(\frac{2n-7}{6} + \frac{1}{n!}\right).$$

На практике последним слагаемым обычно пренебрегают в силу его малости. Поэтому справедлива

Теорема 23 *Для (чисто) случайного временного ряда из n членов среднее число фаз длины d задается формулой (15.3), а математическое ожидание числа фаз во всем ряду приближенно равно* $(2n-7)/3$.

К сожалению, распределение числа фаз весьма далеко от нормального, и даже не сходится к нему при увеличении n, поэтому на практике применяют следующий критерий: вычисляют по формуле (15.3) число ожидаемых фаз фиксированной длины, затем по наблюдаемым данным подсчитывают фактическое количество таких фаз и сравнивают два получившихся ряда чисел при помощи одного из критериев. Если значительных различий не наблюдается, гипотезу о случайном характере временного ряда следует принять.

Рассмотрим следующий числовой пример, заимствованный из [10, с.476]. В таблице приведены урожайности ячменя в Англии и Уэльсе с 1884 по 1939 годы в центнерах на акр:

Урожайность ячменя

год	ц с акра	год	ц с акра	год	ц с акра	год	ц с акра
84	15,2	92	16,5	00	14,9	08	15,5
85	16,9	93	13,3	01	14,5	09	17,3
86	15,3	94	16,5	02	16,6	10	15,5
87	14,9	95	15,0	03	15,1	11	15,1
88	15,7	96	15,9	04	14,6	12	14,2
89	15,1	97	15,5	05	16,0	13	15,8
90	16,7	98	16,9	06	16,8	14	15,7
91	16,3	99	16,4	07	16,8	15	14,1

год	ц с акра	год	ц с акра	год	ц с акра
16	14,8	24	15,4	32	16,0
17	14,4	25	15,3	33	16,8
18	15,6	26	16,0	34	16,9
19	13,9	27	16,4	35	16,6
20	14,7	28	17,2	36	16,2
21	14,3	29	17,8	37	14,0
22	14,0	30	14,4	38	18,1
23	14,5	31	15,0	39	17,5

После изучения этой таблицы, видим, что наблюдаемое число фаз заданной длины, а также ожидаемое число их, рассчитанное по формуле (15.3) и теореме, равны

Сравнение количеств фаз

Длина фазы	1	2	3	Всего
Наблюдаемых	23	7	4	34
Вычисленных	21,25	9,17	2,59	33,67

Как видно без дальнейших вычислений, любой разумный критерий сравнения двух выписанных числовых последовательностей укажет на их совпадение. Таким образом, наш временной ряд следует признать чисто случайным.

Приведем здесь также и вывод критерия подсчета числа экстремальных точек. Внимательное изучение данных приводит нас к выводу, что имеется $e = 34$ экстремальных точки. При этом дважды экстремальная точка соседствует с равным ей значением (1906-7 и 1910-11 годах). Тем самым, мы принимаем $n = 54$, уменьшая на два общее число членов наблюдаемого временного ряда. Формула (15.2) дает значение $t = -1,09$, которое меньше по абсолютной вели чине, чем например, $t_{0,95} = 1,64$, что дает основание и при работе с этим критерием принять гипотезу о случайном характере наблюдаемого ряда. Заметим, что если бы мы не уменьшили число n при расчете, формула (15.2) дала бы $t = -1,50$, что все равно меньше любого разумного критического, хотя и несколько хуже значения, получившегося у нас.

15.2.3 Критерии, основанные на ранговой корреляции

В тех критериях, которые были описаны выше, мы основывали свой вывод о случайном характере наблюдаемого временного ряда, изучая только рядом (или, по крайней мере, подряд) расположенные его члены. Имеется веское основание полагать, что возможно улучшение качества принимаемого решения, если мы будем сравнивать все пары значений, в том числе и расположенных довольно далеко друг от друга.

Определим число m пар $(u_i, u_j),\ j > i$ для которых $u_j > u_i$. Поскольку всего пар элементов ряда, у которых второй элемент был получен при наблюдении раньше первого, имеется $\mathbf{C}_n^2 = n(n-1)/2$, то

$$\mathbf{M}m = \frac{1}{4}n(n-1),$$

и если наблюдаемое m оказывается больше своего среднего, следует признать, что временной ряд имеет тенденцию к возрастанию, если же меньше – то к убыванию, что дает повод разумным признать наличие возрастающего (убывающего) тренда и, тем самым, отвергнуть гипотезу о чистой случайности наблюдаемого ряда. Конечно же, при пользовании этим простым эмпирическим критерием речь идет только о значимых отличиях m от теоретического $\mathbf{M}m$.

Предложенному критерию можно придать вполне законченный вид, используя известные способы проверки значимости отличий двух средних, но мы вместо этого отметим, что рассмотренная процедура фактически есть проверка на равенство нулю коэффициента ранговой корреляции. Напомним, что если мы через X_i обозначим ранг u_i $i = 1, ..., n$, т.е. номер его в упорядоченной по возрастанию цепочке элементов наблюдаемого временного ряда, причем наш ряд не содержит равных между собой элементов, то коэффициент ранговой корреляции Спирмена может быть вычислен по формуле

$$r = 1 - \frac{6Q}{n^3 - n},$$

где

$$Q = \sum_{j=1}^{n}(X_j - j)^2 -$$

сумма квадратов разностей рангов. Вычисления по этой формуле могут быть заменены на более удобную в некоторых случаях формулу, если мы

введем следующее обозначение:

$$H_{i,j} \;=\; \begin{cases} 1, \text{ если } u_i > u_j, \\ 0, \text{ если } u_i < u_j. \end{cases} \qquad\qquad i,j = 1,...,n.$$

Лемма 10 *Если X_i – ранги значений $u_i,\ i = 1,...,n$, то*

$$\frac{1}{2}\sum_{i=1}^{n}(X_i - i)^2 \;=\; \sum_{i=1}^{n-1}\sum_{j=i+1}^{n} H_{i,j}(j-i).$$

Доказательство. Будем действовать по индукции. Сначала возьмем $n = 2$. Тогда надо проверить

$$(X_1 - 1)^2 + (X_2 - 2)^2 \;=\; 2H_{1,2}.$$

Если $X_1 = 1,\ X_2 = 2$, то выписанное равенство превращается в 0=0, если же $X_1 = 2,\ X_2 = 1$, то в 2=2. При $n = 2$ этим исчерпаны все возможные случаи.

Пусть нам уже удалось доказать, что при произвольной последовательности рангов ряда длины n утверждение леммы справедливо. Рассмотрим ранги $X_1, ..., X_{n+1}$ для на единицу более длинного ряда и предположим, что $X_{n+1} = k$. Введем

$$Y_j \;=\; \begin{cases} X_j, \text{ если } X_j < k, \\ X_j - 1, \text{ если } X_j > k, \end{cases} \qquad\qquad j = 1,...,n.$$

По индукционному предположению, для $Y_j,\ j = 1,...,n$ утверждение леммы верно, т.е. мы можем ввести обозначение

$$A \;=\; \sum_{j=1}^{n}(Y_j - j)^2 \;=\; 2\sum_{i=1}^{n-1}\sum_{j=i+1}^{n} H^Y_{i,j}(j-i). \tag{15.4}$$

Отметим, что

$$Y_i < Y_j \iff X_i < X_j, \quad i,j \le n,$$

а значит, $H^Y_{i,j} = H_{i,j}$ при этих i, j.

Правая часть доказываемого в лемме равенства может быть с учетом (15.4) записана, как

$$2\sum_{i=1}^{n}\sum_{j=i+1}^{n+1} H_{i,j}(j-i) \;=\; A + 2\sum_{i=1}^{n} H_{i,n+1}(n+1-i) \;=\; A + 2\sum_{i:X_i>k}^{n}(n+1-i). \tag{15.5}$$

Теперь, в силу определения Y_j,

$$\sum_{j=1}^{n+1}(X_j - j)^2 \;=\; \sum_{j:X_j<k}(Y_j - j)^2 + \sum_{j:X_j>k}(Y_j + 1 - j)^2 + (k - n - 1)^2,$$

и, если положить

$$\Delta \;=\; \sum_{j=1}^{n+1}(X_j - j)^2 \;-\; \sum_{j=1}^{n}(Y_j - j)^2,$$

то

$$\begin{aligned}\Delta \;&=\; \textstyle\sum_{j:X_j>k}(Y_j + 1 - j)^2 - \sum_{j:X_j>k}(Y_j - j)^2 + (n - k + 1)^2 \;= \\ &=\; \textstyle\sum_{j:X_j>k}(2Y_j + 1 - 2j) + (n - k + 1)^2 \;= \\ &=\; \textstyle\sum_{j:X_j>k}(2X_j - 2j - 1) + (n - k + 1)^2.\end{aligned}$$

А с учетом совпадения множеств

$$\{j \,:\, X_j > k\} \;=\; \{(k+1), ..., (n+1)\}$$

можно записать

$$\Delta \;=\; 2\sum_{j:X_j>k}(X_j - j) \;-\; (n - k + 1) \;+\; (n - k + 1)^2.$$

Отсюда, привлекая (15.5), видим, что для завершения доказательства достаточно убедиться в справедливости равенства

$$2\sum_{j:X_j>k}(X_j - j) - (n - k + 1) + (n - k + 1)^2 \;=\; 2\sum_{j:X_j>k}(n + 1 - j),$$

или

$$2\sum_{j:X_j>k}(n + 1 - X_j) \;=\; (n - k + 1)(n - k).$$

Но $X_j > k$ пробегает все значения от $k + 1$ до $n + 1$, следовательно,

$$2\sum_{j:X_j>k}(n + 1 - X_j) \;=\; 2\sum_{j=k+1}^{n} 1 \;=\; (n - k)(n - k + 1).$$

Лемма доказана.

Воспользовавшись этой леммой, запишем

$$r = 1 - \frac{12}{n^3 - n} \sum_{i=1}^{n-1} \sum_{j=i+1}^{n} H_{i,j}(j-i). \tag{15.6}$$

Признаком случайности ряда будет близость r к нулю, точнее говоря, незначимость его отличия от нуля. Проверку можно осуществлять методами, описанными в разделе 2.3, и, более точно, по формулам в конце подраздела 3.3.1. К сожалению, нужно отметить, что проверка гипотезы случайности только что описанным методом требует гораздо больших вычислений, чем методы, описанные выше. Но этот метод дает и большую информацию в случае, если гипотеза случайности будет отвергнута.

Обратимся снова к примеру с урожайностью ячменя. Ниже приводится таблица рангов урожайностей по годам, упорядоченных в порядке возрастания.

Сумма квадратов разности рангов $Q = 2616$, коэффициент ранговой корреляции Спирмена $r = -0,41$. Как видим, в этом случае коэффициент получился умеренно отрицательным, что дает, в отличие от всех предыдущих методов, основание заподозрить наличие тренда, направленного в сторону уменьшения. Проверка по критерию, описанному в 3.3.1, дает значение критерия Стьюдента $T = -3,30$, что дает возможность принять гипотезу о чисто случайном характере временного ряда только на уровне 0,999. При меньших доверительных уровнях гипотеза случайности отвергается, например, двусторонняя критическая точка распределения Стьюдента с 54 степенями свободы $t_{0,99} = 2,68$. Это означает, что принять гипотезу случайности в этой ситуации мы можем, только если априори верим в нее очень сильно.

Ранги урожайности ячменя

год	ранг	год	ранг	год	ранг	год	ранг
84	10	92	47	00	12	08	14
85	36	93	18	01	48	09	25
86	39	94	40	02	6	10	27
87	54	95	21	03	20	11	28
88	32	96	37	04	1	12	35
89	29	97	33	05	3	13	5
90	38	98	4	06	42	14	31
91	34	99	17	07	41	15	30

год	ранг	год	ранг	год	ранг
16	13	24	9	32	2
17	22	25	11	33	15
18	43	26	19	34	51
19	49	27	52	35	45
20	53	28	7	36	26
21	8	29	23	37	56
22	16	30	24	38	46
23	44	31	50	39	55

15.2.4 Коррелограмма

При переходе от чисто случайных временных рядов к рядам, несущим на себе все больший отпечаток зависимости между соседними членами, большое значение имеют так называемые коэффициенты сериальной корреляции. Коэффициентом сериальной корреляции с запаздыванием k (или с лагом k) называется число

$$r_k = \frac{\sum_{i=1}^{n-k}(u_i - \bar{u})(u_{i+k} - \bar{u})}{(n-k)S_u^2}, \tag{15.7}$$

где, конечно же,

$$\bar{u} = \frac{1}{n}\sum_{i=1}^{n} u_i, \quad S_u^2 = \frac{1}{n}\sum_{i=1}^{n}(u_i - \bar{u})^2 -$$

выборочные среднее и дисперсия наблюдаемого ряда.

Последовательность $r_1, r_2, ...r_{n-1}$ называют коррелограммой временного ряда, то же слово употребляют и для диаграммы, наглядно показывающей зависимость r_k от k. Ясно, что чем больше k (по отношению к n), тем меньшее число слагаемых содержит сумма в числителе (15.7). Чтобы "уравнять в правах"малые и большие k, предположим, что ряд продолжен циклическим образом, т.е.

$$u_{n+1} = u_1, \ ..., \ u_{n+k} = u_k.$$

Тогда можно определить

$$r_k^{(c)} = \frac{\sum_{i=1}^{n}(u_i - \bar{u})(u_{i+k} - \bar{u})}{\sum_{i=1}^{n}(u_i - \bar{u})^2}.$$

Этот коэффициент называют коэффициентом циклической или круговой сериальной корреляции с лагом k. Если k мало по сравнению с n, то коэффициенты обычной и круговой сериальной корреляции с лагом k практически совпадают.

15.3 Тренд и сезонность

Наличие тренда (в отсутствие сезонных изменений) предполагает, что наблюдаемый временной ряд может быть записан в виде

$$u_t \;=\; f(t) \;+\; \xi_t, \quad t = 1, ..., n,$$

где $f(.)$ – неслучайная функция, определяющая тренд, а $\xi_1, ..., \xi_n$ – последовательность независимых одинаково распределенных случайных величин, на которую можно смотреть как на чисто случайный временной ряд. Тем самым, понятие тренда можно рассматривать, как некую "основную"составляющую, на которую накладываются нерегулярные колебания.

В ряде практических задач вид функции f известен с точностью до параметров, и тогда оценка этих параметров делается при помощи обычных методов регрессионного анализа – об одном особо важном частном случае мы будем говорить подробнее чуть ниже. Иногда же и никаких предположений о виде функции уверенно сделать нельзя. В этом случае принято применять гладкое приближение к истинной функции тренда – об этом мы поговорим в подпункте "Сглаживание".

15.3.1 Полиномиальные тренды

Здесь мы будем предполагать, что тренд является полиномом степени q:

$$f(t) \;=\; a_0 + a_1 t + ... + a_q t^q.$$

Обычно q бывает мало по сравнению с n. Для оценки коэффициентов $a_j,\ j = 0, ..., q$ можно воспользоваться обычными методами регрессионного анализа (см. главу 6). Однако там же доказано, что оценки коэффициентов получаются лучше, если данные, по которым производится их оценка, предварительно ортогонализованы.

Перейдем от $1, t, t^2, ..., t^q$ к ортогональным переменным $\varphi_j(n,t),\ j = 0, ..., n$. Пусть

$$\varphi_j(n,t) \ = \ t^j + C_{j-1}(j,n)t^{j-1} + ... + C_0(j,n),\ j = 1, ..., (n-1).$$

При этом мы дополнительно предположим, что $\varphi_0(n,t) = 1$. Ортогональность $\varphi_j(n,t)$ всем полиномам с меньшими номерами эквивалентна его ортогональности $1, t, ..., t^{j-1}$. Поэтому

$$\sum_{i=1}^{n} \varphi_j(n,t)t^i \ = \ 0,\ i = 0, ..., (j-1),$$

что можно переписать в виде

$$\sum_{s=0}^{j-1} C_s(j,n) \sum_{t=1}^{n} t^{i+s} \ = \ -\sum_{t=1}^{n} t^{i+j}, \quad i = 0, ..., (j-1). \tag{15.8}$$

Формулы (15.8) представляют собой систему уравнений для нахождения $C_s(j,n)$, которая всегда однозначно разрешима. Например, для нахождения $C_0(1,n)$ положим в (15.8) $i = 0,\ j = 1$:

$$C_0(1,n) \sum_{t=1}^{n} 1 \ = \ -\sum_{t=1}^{n} t,$$

откуда

$$C_0(1,n) \ = \ -\frac{n+1}{2}.$$

Аналогично, полагая $j = 2$ и рассматривая $i = 0, 1$, получаем систему

$$\begin{cases} C_0(2,n) \sum_{t=1}^{n} 1 + C_1(2,n) \sum_{t=1}^{n} t \ = \ -\sum_{t=1}^{n} t^2, \\ C_0(2,n) \sum_{t=1}^{n} t + C_1(2,n) \sum_{t=1}^{n} t^2 \ = \ -\sum_{t=1}^{n} t^2, \end{cases}$$

из которой находятся значения $C_0(2,n), C_1(2,n)$.

Вычисляя эти коэффициенты, можно последовательно выписывать ортогональные полиномы

$$\varphi_0(n,t) \ = \ 1,$$

$$\varphi_1(n,t) \ = \ t - \frac{n+1}{2},$$

$$\varphi_2(n,t) \ = \ t^2 - (n+1)t + \frac{n^2+3n+2}{6},$$

и т.д. Здесь полезной бывает таблица сумм степеней натуральных чисел, которая приведена ниже.

Запишем тренд через ортогональные полиномы

$$f(t) \;=\; \sum_{s=0}^{q} \alpha_s \varphi_s(n,t).$$

Все задействованные в наших выкладках коэффициенты связаны формулой

$$a_j \;=\; \alpha_j + \sum_{s=j+1}^{q} C_j(s,n)\alpha_s,\;\; j = 0, ..., q.$$

Согласно (6.9) в условиях ортогональности исходных данных, метод наименьших квадратов дает оценки для α_s

$$\alpha_s^* \;=\; \frac{\sum_{t=1}^{n} u_t \varphi_s(n,t)}{\phi_s^2},\;\; s = 0, ..., q,$$

а несмещенной оценкой для дисперсии чисто случайной составляющей (остаточной дисперсии) будет

$$\sigma_*^2 \;=\; \frac{1}{n-q-1} \sum_{t=1}^{n} \left(u_t - \alpha_0^* - \sum_{i=1}^{q} \alpha_i^* \varphi_i(n,t) \right)^2.$$

Здесь использовано обозначение

$$\phi_j^2 \;=\; \sum_{t=1}^{n} \varphi_j^2(n,t).$$

Осталось только добавить, что формулы для ортогональных полиномов нетрудно найти в специальной литературе, даже не занимаясь специально решением системы (15.8). См., например, [6, с. 376 и далее]. (Следует иметь ввиду, что в упомянутой книге значения t первоначально центрируются числом $(n+1)/2$). Заметим, наконец, что, поскольку постоянные множители не влияют на условие ортогональности, ортогональные полиномы для удобства вычислений (например, чтобы добиться целых, а не дробных коэффициентов), можно умножать на произвольные числа, каждый на свое.

15.3.2 Выбор степени тренда

Если степень полиномиального тренда q известна заранее, то процедура, изложенная выше, дает эффективный способ оценивания тренда. После этого можно, например, заниматься прогнозированием значений временного ряда при $T > n$ по приближенной формуле

$$u_T \approx \sum_{j=1}^{q} \alpha_j^* \varphi_j(n, T),$$

а если необходимо указать интервал, в который с достаточно большой вероятностью попадут значения u_T, то можно воспользоваться стандартной процедурой, например неравенством "трех сигм", т.к. несмещенная оценка для остаточной дисперсии σ_*^2 уже была выписана. В случае же, когда распределение u_t нормально (особенно важный частный случай), можно воспользоваться методикой построения толерантных интервалов для нормального распределения.

Обратимся к задаче определения степени q тренда, если заранее она неизвестна. Отметим, что, чем выше степень, тем точнее мы получим аппроксимацию, но с другой стороны, тренды более высоких степеней обладают меньшей устойчивостью относительно погрешностей измерения и меньшей наглядностью. Поэтому следует выбирать тренд невысоких степеней, если только это возможно.

Из общей теории метода наименьших квадратов нам известно, что оценки α_j^* являются несмещенными оценками коэффициентов α_j, причем они некоррелированы и

$$\mathbf{D}\alpha_j^* = \frac{1}{\phi_j^2} \sum_{t=1}^{n} u_t \varphi_j(n, t).$$

Если u_t имеют нормальные распределения (а параметры этих распределений, конечно же, будут $f(t), \sigma^2$), то и α_j^* также будут иметь нормальные распределения (с математическим ожиданием $f(t) \sum_{t=1}^{n} \varphi_j(n, t)/\phi_j^2$ и дисперсией σ^2/ϕ_j^4) и, как следует из теорем о распределениях статистик, связанных с нормальными совокупностями, величина $(n - q - 1)\sigma_*^2/\sigma^2$ имеет хи-квадрат распределение с $n - q - 1$ степенью свободы.

Отсюда, наилучший критерий для проверки гипотезы о том, что $\alpha_q \neq 0$ отвергает ее, если

$$\frac{|\alpha_q^* \phi_q|}{\sigma_*} < t(\varepsilon), \tag{15.9}$$

где $t(\varepsilon)$ – двусторонняя критическая точка распределения Стьюдента с $n-q-1$ степенями свободы.

Опишем процедуру определения минимальной степени тренда более подробно. Допустим, нам известно из априорных соображений, что степень тренда не может быть ниже m (случай $m=0$ не исключается) и не должна быть выше q. Выскажем $q-m+1$ гипотезу:

$$\begin{aligned}
H_q: &\quad \alpha_q \neq 0,\\
H_{q-1}: &\quad \alpha_q = 0, \alpha_{q-1} \neq 0,\\
&\quad \vdots\\
H_{m+1}: &\quad \alpha_q = 0, \ \ldots \alpha_{m+2} = 0, \ \alpha_{m+1} \neq 0,\\
H_m: &\quad \alpha_q = \ldots = \alpha_{m+1} = 0.
\end{aligned}$$

Условимся, что в данной ситуации мы хотим не выбрать полином более высокой степени, чем он есть на самом деле (ошибка второго рода), и при этом не хотим, чтобы степень этого полинома была бы слишком низкой. Обозначим множество тех наблюдаемых временных рядов, для которых принимается гипотеза H_j, через R_j, $j=m,\ldots,q$. В [12, с.53-55] доказано, что множества R_j можно строить независимо друг от друга, и при этом оптимальность процедуры не нарушается. Поэтому все эти множества будут иметь структуру (15.9): гипотеза H_j принимается, если

$$\frac{|\alpha_j^* \phi_j|}{\sqrt{(n-q-1)\sigma_*^2 + \sum_{j=i+1}^{q} \alpha_j^{*2}\phi_j^2}} > t(\varepsilon_i),$$

где $t(\varepsilon_i)$ – двусторонняя критическая точка распределения Стьюдента с $n-i-1$ степенями свободы.

Если мы хотим добиться уровня ошибки первого рода не выше, чем ε, то можно положить

$$\varepsilon_i = \varepsilon(1-\varepsilon)^{q-i}, \quad i=m,\ldots,q.$$

Таким образом, процедура нахождения оптимальной степени тренда носит последовательный характер: мы должны, начиная с q, проверять соответствующие гипотезы до тех пор, пока какая-то H_j не примется. Тогда j и есть наилучшая степень полинома, аппроксимирующего тренд.

15.3.3 Сглаживание

Иногда, если есть основания считать, что единая формула для тренда для всего рассматриваемого временного ряда не сработает, или для нужд исследования достаточно просто наглядное представление о графике тренда, применяют так называемый метод скользящих средних, суть которого в усреднении (сглаживании) за счет использования некоторого количества подряд расположенных членов ряда. В основе этого метода лежит представление о том, что тренд представляет собой некоторую усредненную неслучайную величину, а случайные составляющие за счет своего нулевого математического ожидания дадут в усреднение весьма незначительный вклад, которым можно пренебречь.

Пусть имеется временной ряд $u_1, ..., u_n$. Оценим значение тренда в точке t при помощи формулы

$$f_t^* \;=\; \sum_{s=-m}^{m} c_s u_{t+s}, \tag{15.10}$$

т.е. как взвешенное среднее с учетом значений ряда, удаленных не более чем на m точек как влево, так и вправо. Сразу договоримся, что мы будем считать всюду ниже, что $c_{-s} = c_s,\ s = 1, ..., m$. Предполагается также, что

$$\sum_{s=-m}^{m} c_s \;=\; 1.$$

Если в действительности было $u_t \;=\; f(t) + \xi_t$, то

$$f_t^* \;=\; \sum_{s=-m}^{m} c_s f(t+s) \;+\; \xi_t^*,$$

где у усредненной случайной компоненты ξ_t^*, как нетрудно понять, по-прежнему нулевое математическое ожидание, а вот дисперсия равна

$$\mathbf{D}\xi_t^* \;=\; \sigma^2 \sum_{s=-m}^{m} c_s^2.$$

Конечно же, коэффициенты c_s подбирают так, что они малы, и выписанная дисперсия существенно меньше, чем σ^2 .

В простейшем случае все коэффициенты могут быть выбраны равными по $1/(2m+1)$, но обычно пытаются подобрать по рассматриваемым $2m+1$ точке полином некоторой заранее заданной степени, график

которого максимально близок к этим точкам. Предположим, что тренд $f(t+s)$ при $s = -m, ..., m$ можно приблизить полиномом

$$f_t(s) \;=\; \sum_{j=0}^{q} a_j(t) s^j.$$

В частности, $f(t) \approx f_t(0) = a_0(t)$. Коэффициенты a_j оценим методом наименьших квадратов. Нормальные уравнения для их определения имеют вид

$$\sum_{j=0}^{q} a_j(t) \sum_{s=-m}^{m} s^{i+j} \;=\; \sum_{s=-m}^{m} s^i u_(t+s), \quad i = 0, ..., q.$$

Заметим, что все суммы, степени в которых нечетны, равны 0. Поэтому, если i четно, то равны нулю коэффициенты при всех нечетных номерах a и наоборот. Поскольку нам достаточно найти $a_0(t)$, воспользуемся этими уравнениями при четных i. Для этого положим

$$p \;=\; \begin{cases} q/2, \text{ если } q \text{ четное}, \\ (q-1)/2, \text{ иначе}. \end{cases}$$

Тогда

$$\sum_{j=0}^{p} a_{2j}(t) \sum_{s=-m}^{m} s^{2i+2j} \;=\; \sum_{s=-m}^{m} s^{2i} u_{t+s}, \quad i = 0, ..., p.$$

Решая это уравнение относительно $a_0(t)$, получим

$$a_0(t) \;=\; \sum_{s=-m}^{m} c_s u_{t+s} \tag{15.11}$$

для некоторых коэффициентов c_s, что и являлось нашей целью.

Возьмем для примера $m = 2, q = 2$, описывая параболой поведение нашего временного ряда в пяти соседних точках. Поскольку q четно, то $a_1 = 0$. Нормальные уравнения имеют вид

$$\begin{cases} 5a_0 + 10a_2 \;=\; \sum_{s=-2}^{2} u_{t+s}, \\ 10a_0 + 34a_2 \;=\; \sum_{s=1}^{2} s^2 (u_{t-s} + u_{t+s}). \end{cases}$$

Отсюда

$$\begin{aligned} a_0 \;&=\; \tfrac{34}{70} \textstyle\sum_{s=-2}^{2} u_{t+s} \;-\; \tfrac{10}{70} \textstyle\sum_{s=1}^{2} s^2 (u_{t-s} + u_{t+s}) \;= \\ &=\; (-3u_{t-2} + 12u_{t-1} + 17u_t + 12u_{t+1} - 3u_{t+2})/35 \end{aligned}$$

Как видим, в этом примере найдены коэффициенты $c_s, s = -2, ..., 2$. Из самого алгоритма видно, что если p не меняется, то не меняются и найденные коэффициенты. Это означает, что одни и те же коэффициенты используются для четного и соседнего нечетного q. Ниже приводится таблица коэффициентов $c_s, s = 0, ..., m$ для $p = 1, 2,\ m \leq 5$. Напомним, что $c_{-s} = c_s$.

Более полные таблицы коэффициентов можно найти в [10].

К методике скользящих средних пришли давно, и совсем не из соображений, имеющих какое-либо отношение к математической статистике, в результате чего практически невозможно описать получающиеся оценки коэффициентов с точки зрения их вероятностных характеристик.

Одна из часто используемых процедур сглаживания использует последовательное, несколько раз примененное усреднение. Такова, например, формула Спенсера для 15 точек. Сначала вычисляем

$$u_t^* \ = \ \frac{1}{4}(-3u_{t-2} + 3u_{t-1} + 4u_t + 3u_{t+1} - 3u_{t+2}),$$

затем усредняется с равными весами 5 последовательных u_t^* и после этого усредняются (снова с равными весами) 4 последовательных члена последнего ряда. Имеется также вариант этой процедуры, использующий 21 последовательную точку.

К сожалению, использование техники сглаживания не позволяет оценить значения тренда в m крайних левых и m крайних правых точках. И если левые – "начальные"точки, возможно и не так важны, но крайние правые точки сильно влияют на точность прогноза значений временного ряда. Поэтому при применении метода скользящих средних имеются специальные методики для восстановления значений тренда в последних точках. Мы не будем здесь на них останавливаться.

15.3.4 Оценка сезонных колебаний

Пусть на временной ряд накладываются регулярные периодические колебания. Не ограничивая общности, условимся считать, что эти колебания обуславливаются сезонными изменениями в природе в течение календарного года. Т.о., мы предполагаем, что

$$u_t \ = \ f(t) \ + \ g(t) \ + \ \xi_t, \quad t = 1, ..., n, \tag{15.12}$$

Таблица 15.1: Суммы степеней натуральных чисел

j	$\sum_{t=1}^{n} t^j$
0	n
1	$n(n+1)/2$
2	$n(n+1)(2n+1)/6$
3	$n^2(n+1)^2/4$
4	$n(n+1)(2n+1)(3n^2+3n-1)/30$
5	$n^2(n+1)^2(2n^2+2n-1)/12$
6	$n(n+1)(2n+1)(3n^4+6n^3-3n+1)/42$
7	$n^2(n+1)^2(3n^4+6n^3-n^2-4n+2)/24$
8	$n(n+1)(2n+1)(5n^6+15n^5+5n^4-15n^3-n^2+9n-3)/90$

Таблица 15.2: Коэффициенты формул сглаживания

$k=1$						
m	c_5	c_4	c_3	c_2	c_1	c_0
2				$-\frac{3}{35}$	$\frac{12}{35}$	$\frac{17}{35}$
3			$-\frac{2}{21}$	$\frac{3}{21}$	$\frac{6}{21}$	$\frac{7}{21}$
4		$-\frac{21}{231}$	$\frac{14}{231}$	$\frac{39}{231}$	$\frac{54}{231}$	$\frac{59}{231}$
5	$-\frac{36}{429}$	$\frac{9}{429}$	$\frac{44}{429}$	$\frac{69}{429}$	$\frac{84}{429}$	$\frac{89}{429}$

$k=2$						
m	c_5	c_4	c_3	c_2	c_1	c_0
3			$\frac{5}{231}$	$-\frac{30}{231}$	$\frac{75}{231}$	$\frac{131}{231}$
4		$\frac{15}{429}$	$-\frac{55}{429}$	$\frac{30}{429}$	$\frac{135}{429}$	$\frac{179}{429}$
5	$\frac{18}{429}$	$-\frac{45}{429}$	$-\frac{10}{429}$	$\frac{60}{429}$	$\frac{120}{429}$	$\frac{143}{429}$

где $f(t)$ – тренд, а $g(t)$ – периодическая функция, имеющая период T (12 для ежемесячных измерений, 365 для ежедневных и т.п.). Будем считать, что n нацело делится на T:

$$n \;=\; hT,$$

а также условимся считать, что $g(t)$ пронормирована таким образом, что

$$\sum_{t=1}^{T} g(t) \;=\; 0.$$

Это в силу периодичности g означает, что для произвольного s

$$\sum_{t=1}^{T} g(t+s) \;=\; 0, \quad s = 0, ..., (n-T).$$

Будем предполагать, что тренд $f(t)$ в отличие от эффекта сезонности $g(t)$ меняется относительно медленно, иначе эффект сезонности однозначно выделить будет невозможно. Получаем задачу одновременной оценки тренда и сезонности. Сначала выделим тренд.

В экономических расчетах T обычно бывает четным. Если $T = 2m$, то для оценки тренда используем скользящее среднее

$$f_t^* \;=\; \frac{1}{T}\left(\frac{1}{2}u_{t-m} + \sum_{s=-m+1}^{m-1} u_{t+s} + \frac{1}{2}u_{t+m}\right),\; t = (m+1), ..., (n-m). \quad (15.13)$$

Тогда в силу условия нормировки g

$$\mathbf{M}f_t^* \;=\; \frac{1}{T}\left(\frac{1}{2}f(t-m) + \sum_{s=-m+1}^{m-1} f(t+s) + \frac{1}{2}f(t+m)\right) \;\approx\; f(t).$$

Поэтому f_t^* является несмещенной оценкой тренда в точке t. Перейдем к новому ряду, вычитая из исходного оценку тренда в каждой точке: $u_t^0 \;=\; u_t - f_t^*,\; t = 1, ..., n.$

Для оценивания сезонности используем соотношение

$$g^*(t) \;=\; \frac{1}{h}\sum_{j=0}^{h-1} u_{t+jT}^0 \;-\; \frac{1}{n}\sum_{s=1}^{n} u_s^0, \quad (15.14)$$

т.е., например, сезонный эффект первого квартала равен разности средних по всем первым кварталам и общего среднего.

Используя формулы (15.13) и (15.14), можно оценить в нашей задаче значения тренда (во всех точках, кроме крайних) и сезонности одновременно. Аналитические аппроксимации сезонных изменений можно получить, проводя аппроксимацию численных значений формулы (15.14) в рамках одного периода комбинациями синуса и косинуса времени (например, методом наименьших квадратов).

Отметим, что аддитивная модель вклада сезонности (15.12) может быть также исследована методом двухфакторного дисперсионного анализа на наличие существенного вклада сезонности в значения временного ряда, а также на взаимодействие сезонности и тренда.

Иногда рассматривают вклад сезонности как мультипликативный, т.е. вместо (15.12) полагают

$$u_{t,j} \;=\; f(t)g(j) + \xi_{t,j}, \quad t = 1, ..., n,\ j = 1, ..., T. \tag{15.15}$$

Простейший способ оценки сезонного эффекта в этой модели состоит в том, чтобы разделить $u_{t,j}$ на его среднее, а затем рассмотреть получающиеся частные, как оценки значений g. Другой способ заключается в логарифмировании исходных данных. Тогда модель (15.15) перейдет в (15.12) относительно логарифмов, и мы придем к уже рассмотренной выше задаче.

Иногда к хорошим результатам приводит итеративный процесс. Сначала исключаем первое приближение тренда, затем оцениваем сезонный фактор, вносим поправки в исходные данные, чтобы исключить его, повторно оцениваем тренд (прибавляя его к уже имевшемуся) и так далее.

Приложение A

Приложение 1. Нормальное распределение

Говорят, что случайная величина ξ имеет нормальное распределение с параметрами $a,\ \sigma^2$, если она является величиной абсолютно непрерывного типа и имеет плотность

$$\varphi_{a,\sigma}(x) \ = \ \frac{1}{\sigma\sqrt{2\pi}} \exp\left\{-\frac{(x-a)^2}{2}\right\}.$$

Это распределение мы будем обозначать $N(a,\sigma^2)$, а соответствующую ему функцию распределения через $\Phi_{a,\sigma}(x)$:

$$\Phi_{a,\sigma}(x) \ = \ \mathbf{P}(\xi < x) \ = \ \int\limits_{-\infty}^{x} \varphi_{a,\sigma}(x)dx.$$

Если $a = 0,\ \sigma = 1$, то такое распределение называется стандартным нормальным, и его плотность и функцию распределения условимся обозначать $\varphi(x), \Phi(x)$ без индексов.

К сожалению, интеграл, выписанный выше, не берется в виде конечной формулы, поэтому при использовании нормального распределения используются таблицы или приближенные формулы для вычислений. Если у вас под рукой имеется компьютер с установленным пакетом Microsoft Office (Excel), то таблицы, приводимые ниже, вам не понадобятся. Для нахождения значения $\Phi(x)$ достаточно вызвать Мастер функций электронных таблиц Excel и найти в разделе Статистические Функции

одну из функций НОРМРАСП, НОРМСТРАСП или НОРМОБР, НОРМСТОБР. Первые две функции позволяют получить значения функции или плотности (в зависимости от вводимого параметра) произвольного и стандартного нормального распределений, а следующие две - найти аргумент функции нормального распределения по заданному ее значению, то есть решить обратную задачу.

От произвольного нормального распределения $N(a, \sigma^2)$ нетрудно перейти к стандартному $N(0, 1)$ путем следующего преобразования

$$\Phi_{a,\sigma}(x) \;=\; \Phi\left(\frac{x-a}{\sigma}\right).$$

Это позволяет ограничиться табулированием стандартного нормального распределения. Более того, стандартное нормальное распределение симметрично, т.е. распределение ξ совпадает с распределением $-\xi$, откуда просто выводится, что для произвольного x справедливо

$$\Phi(x) \;+\; \Phi(-x) \;=\; 1,$$

что ограничивает потребность в таблицах лишь аргументами одного знака.

Ниже приводятся таблицы стандартного нормального распределения. Отметим еще два обстоятельства. Во-первых, функция Φ, как и любая функция распределения, монотонно не убывает, и имеет предельные значения 0 и 1 на бесконечностях разных знаков. Во-вторых, при использовании этих таблиц для расчетов, описанных в других пособиях, следует иметь ввиду, что разные авторы под словами "функция нормального распределения"и близкими к ним иногда понимают нечто иное, чем в нашем курсе. Например, интеграл в определении функции Φ может считаться в пределах от 0 до x, или как-либо еще. Различие в определениях приводит к различным свойствам и иным численным значениям. Будьте внимательны.

Пример на считывание: $\Phi(2,92) = \Phi(2,5 + 0,42) = 0,99825$ (в строке 0,42 и столбце 2,5 написано $0,9^2825$, степень девятки означает ее повторение, целая часть и запятая опущены).

Функция стандартного нормального распределения

$$\Phi(x) \;=\; \int\limits_{-\infty}^{x} e^{-t^2/2} dt$$

x	0.0+	0.5+	1.0+	1.5+	2.0+	2.5+	3.0+	3.5+
0.00	5000	6915	8413	9332	9772	9^2379	9^2865	9^377
0.01	5040	6950	8438	9345	9778	9^2396	9^2869	9^378
0.02	5080	6985	8461	9357	9783	9^2413	9^2874	9^378
0.03	5120	7019	8485	9370	9788	9^2430	9^2878	9^379
0.04	5160	7054	8506	9382	9793	9^2446	9^2882	9^380
0.05	5199	7088	8531	9394	9798	9^2461	9^2886	9^381

Функция стандартного нормального распределения

$$\Phi(x) = \int\limits_{-\infty}^{x} e^{-t^2/2}dt$$

(продолжение)

x	0.0+	0.5+	1.0+	1.5+	2.0+	2.5+	3.0+	3.5+
0.06	5239	7123	8554	9406	9803	9^2477	9^2889	9^381
0.07	5279	7157	8577	9418	9808	9^2492	9^2893	9^382
0.08	5319	7190	8599	9429	9812	9^2506	9^2897	9^383
0.09	5359	7224	8621	9441	9817	9^2520	9^2900	9^383
0.10	5398	7257	8643	9452	9821	9^2534	9^303	9^384
0.11	5438	7291	8665	9463	9826	9^2547	9^306	9^385
0.12	5478	7324	8686	9474	9830	9^2560	9^310	9^385
0.13	5517	7357	8708	9484	9834	9^2572	9^313	9^386
0.14	5557	7389	8729	9495	9838	9^2585	9^316	9^386
0.15	5596	7422	8749	9505	9842	9^2598	9^318	9^387
0.16	5636	7454	8770	9515	9846	9^2609	9^321	9^387
0.17	5675	7486	8790	9525	9850	9^2621	9^324	9^388
0.18	5714	7517	8810	9535	9854	9^2632	9^326	9^388
0.19	5753	7549	8830	9545	9857	9^2643	9^329	9^389
0.20	5793	7580	8849	9554	9861	9^2653	9^331	9^389
0.21	5832	7611	8869	9564	9864	9^2664	9^334	9^390
0.22	5871	7642	8888	9573	9868	9^2674	9^336	9^390
0.23	5910	7673	8907	9582	9871	9^2683	9^338	9^404
0.24	5948	7704	8925	9591	9875	9^2693	9^340	9^408
0.25	5987	7738	8944	9599	9878	9^2702	9^342	9^412
0.26	6026	7764	8962	9608	9881	9^2711	9^344	9^415
0.27	6064	7794	8980	9616	9884	9^2720	9^346	9^418
0.28	6103	7823	8997	9625	9887	9^2728	9^348	9^422
0.29	6141	7852	9015	9633	9890	9^2736	9^350	9^425
0.30	6179	7881	9032	9641	9893	9^2744	9^352	9^428
0.31	6217	7910	9049	9649	9896	9^2752	9^353	9^431
0.32	6255	7939	9066	9656	9898	9^2760	9^355	9^433
0.33	6293	7967	9082	9664	9901	9^2767	9^357	9^436
0.34	6331	7995	9099	9671	9904	9^2774	9^358	9^439
0.35	6368	8023	9115	9678	9906	9^2781	9^360	9^441

Пример на считывание: $\Phi(2,92) = \Phi(2,5+0,42) = 0,99825$ (в строке 0,42 и столбце 2,5 написано $0,9^2825$, степень девятки означает ее повторение,

целая часть и запятая опущены).

Функция стандартного нормального распределения

$$\Phi(x) = \int\limits_{-\infty}^{x} e^{-t^2/2}dt$$

(продолжение)

x	0.0+	0.5+	1.0+	1.5+	2.0+	2.5+	3.0+	3.5+
0.36	6406	8051	9131	9686	9909	9^2788	9^361	9^443
0.37	6443	8078	9147	9693	9911	9^2795	9^362	9^446
0.38	6480	8106	9162	9699	9913	9^2801	9^364	9^448
0.39	6517	8133	9177	9706	9916	9^2807	9^365	9^450
0.40	6554	8159	9192	9713	9918	9^2813	9^366	9^452
0.41	6591	8186	9207	9719	9920	9^2819	9^368	9^454
0.42	6628	8212	9222	9726	9922	9^2825	9^369	9^456
0.43	6664	8238	9236	9732	9925	9^2831	9^370	9^458
0.44	6700	8264	9251	9738	9927	9^2836	9^371	9^459
0.45	6736	8289	9255	9744	9929	9^2841	9^372	9^461
0.46	6772	8315	9279	9750	9931	9^2846	9^373	9^463
0.47	6808	8340	9292	9756	9932	9^2851	9^374	9^464
0.48	6844	8365	9306	9761	9934	9^2856	9^375	9^466
0.49	6879	8389	9319	9767	9936	9^2861	9^376	9^467

Пример на считывание: $\Phi(2,92) = \Phi(2,5+0,42) = 0,99825$ (в строке 0,42 и столбце 2,5 написано $0,9^2825$, степень девятки означает ее повторение, целая часть и запятая опущены).

Таблица заимствована из [10].

Ниже приведена таблица квантилей стандартного нормального распределения, в которой собраны квантили наиболее часто используемых уровней. Напомним, что квантилью уровня α называется решение уравнения $\Phi(x) = \alpha$. При необходимости для нахождения квантилей других уровней используйте предыдущую таблицу.

Квантили стандартного нормального распределения

α	0,0005	0,001	0,005	0,01	0,05	0,1
x	-3,29048	-3,09024	-2,57583	-2,32634	-1,64485	-1,28155
α	0,9	0,95	0,99	0,995	0,999	0.9995
x	1,281551	1,644853	2,326342	2,575835	3,090245	3,29048

Приложение B

Приложение 2. Распределение χ^2

Рассмотрим k независимых стандартных нормальных случайных величин $\xi_1, ..., \xi_k$. Распределение суммы их квадратов называют распределением χ^2_k (с k степенями свободы). Это распределение является одним из так называемых гамма-распределений $\Gamma_{\alpha,\lambda}$ с плотностями

$$p_{\gamma,\lambda}(x) \;=\; \frac{\alpha^\lambda}{\Gamma(\lambda)} x^{\lambda-1} e^{-\alpha x}, x \geq 0.$$

Точнее, $\chi^2_k \;=\; \Gamma_{1/2,\; k/2}$. Это распределение широко применяется в статистических расчетах. Ниже приводится таблица критических точек этого распределения. Напомним, что критической точкой распределения случайной величины ξ с функцией F уровня α называется такое число t, что

$$\alpha \;=\; \mathbf{P}(\xi \geq t) \;=\; 1 - F(t).$$

Очевидно, что для нахождения квантилей распределения следует иметь ввиду, что квантиль уровня β равна критической точке уровня $1 - \beta$. Случайная величина ξ с распределением χ^2_k принимает только положительные значения и в среднем равна k.

Для значений числа степеней свободы k, больших 30, величину $\sqrt{2\xi}$ можно приближенно считать распределенной нормально со средним значением $\sqrt{2k-1}$ и дисперсией, равной 1, откуда

$$\mathbf{P}(\xi < t) \;\approx\; \Phi(\sqrt{2t} - \sqrt{2k-1}), \quad t > 0,\; k > 30,$$

а значит для нахождения критической точки уровня α пользуемся формулой

$$t \;=\; \frac{1}{2}\left(\sqrt{2k-1} \;-\; t_\alpha\right)^2,$$

где t_α - квантиль стандартного нормального распределения уровня α (берем из таблицы первого приложения).

При использовании электронных таблиц Microsoft Excel соответствующие функции в разделе Статистических функций можно найти под названиями ХИ2РАСП и ХИ2ОБР (квантили).

Иногда полезным может оказаться тот факт, что сумма двух независимых случайных величин, имеющих распределения χ^2_k и χ^2_m, имеет распределения χ^2_{k+m}. Это следует непосредственно из определения χ^2.

Традиционным способом, позволяющим найти такие $t^-_\varepsilon,\ t^+_\varepsilon$, что

$$\mathbf{P}(t^-_\varepsilon < \xi < \ t^+_\varepsilon) \;=\; 1-\varepsilon$$

является выбор в качестве t^-_ε квантили уровня $\varepsilon/2$, а в качестве t^+_ε – уровня $1-\varepsilon/2$. Это связано с несимметричностью распределения χ^2.

Квантили распределения χ^2_k

k	0,999	0,9975	0,995	0,99	0,975	0,95	0,9
1	1,6E-6	9,82E-6	3,93E-5	0,0002	0,001	0,004	0,016
2	0,002	0,005	0,010	0,020	0,051	0,103	0,211
3	0,024	0,04	0,072	0,115	0,216	0,352	0,584
4	0,091	0,145	0,207	0,297	0,484	0,711	1,064
5	0,210	0,307	0,412	0,554	0,832	1,145	1,610
6	0,381	0,526	0,676	0,872	1,237	1,635	2,204
7	0,598	0,794	0,989	1,239	1,690	2,167	2,833
8	0,857	1,104	1,344	1,646	2,180	2,733	3,489
9	1,152	1,450	1,735	2,088	2,700	3,325	4,168
10	1,479	1,827	2,156	2,558	3,247	3,940	4,865
11	1,834	2,232	2,603	3,053	3,816	4,575	5,578
12	2,214	2,661	3,074	3,570	4,404	5,226	6,304
13	2,617	3,112	3,565	4,107	5,009	5,892	7,041
14	3,041	3,582	4,075	4,660	5,629	6,571	7,789
15	3,482	4,069	4,601	5,229	6,262	7,261	8,547

k	0,999	0,9975	0,995	0,99	0,975	0,95	0,9
16	3,942	4,573	5,142	5,812	6,908	7,962	9,312
17	4,416	5,092	5,697	6,408	7,564	8,672	10,085
18	4,905	5,623	6,265	7,015	8,231	9,390	10,865
19	5,407	6,167	6,844	7,633	8,906	10,117	11,651
20	5,921	6,723	7,434	8,260	9,591	10,851	12,443
21	6,447	7,289	8,034	8,897	10,283	11,591	13,240
22	6,983	7,865	8,643	9,542	10,982	12,338	14,041
23	7,529	8,450	9,260	10,196	11,688	13,090	14,848
24	8,085	9,044	9,886	10,856	12,401	13,848	15,659
25	8,649	9,646	10,520	11,524	13,120	14,611	16,473
26	9,222	10,256	11,160	12,198	13,844	15,379	17,292
27	9,803	10,873	11,808	12,878	14,573	16,151	18,114
28	10,391	11,497	12,461	13,565	15,308	16,928	18,939
29	10,986	12,128	13,121	14,256	16,047	17,708	19,768
30	11,588	12,764	13,787	14,953	16,791	18,493	20,599

Квантили распределения χ^2_k
(продолжение)

k	0,001	0,0025	0,005	0,01	0,025	0,05	0,1
1	10,827	9,140	7,879	6,635	5,024	3,841	2,705
2	13,815	11,983	10,596	9,210	7,378	5,991	4,605
3	16,266	14,320	12,838	11,345	9,348	7,815	6,251
4	18,466	16,424	14,860	13,277	11,143	9,488	7,779
5	20,515	18,385	16,750	15,086	12,832	11,070	9,236
6	22,457	20,249	18,547	16,812	14,449	12,593	10,645
7	24,321	22,040	20,278	18,475	16,013	14,067	12,017
8	26,124	23,774	21,955	20,090	17,534	15,507	13,361
9	27,877	25,462	23,589	21,666	19,023	16,919	14,684
10	29,588	27,112	25,188	23,209	20,483	18,307	15,987
11	31,263	28,729	26,757	24,725	21,920	19,675	17,275
12	32,909	30,318	28,300	26,217	23,337	21,026	18,549
13	34,527	31,883	29,819	27,688	24,735	22,362	19,812
14	36,124	33,426	31,319	29,141	26,119	23,685	21,064
15	37,698	34,949	32,801	30,578	27,488	24,996	22,307

k	0,001	0,0025	0,005	0,01	0,025	0,05	0,1
16	39,252	36,455	34,267	32,000	28,845	26,296	23,542
17	40,791	37,946	35,718	33,409	30,191	27,587	24,769
18	42,312	39,422	37,156	34,805	31,526	28,869	25,989
19	43,819	40,885	38,582	36,191	32,852	30,143	27,204
20	45,314	42,336	39,997	37,566	34,169	31,410	28,412
21	46,796	43,775	41,401	38,932	35,479	32,670	29,615
22	48,268	45,204	42,796	40,289	36,781	33,924	30,813
23	49,728	46,623	44,181	41,638	38,076	35,172	32,007
24	51,179	48,034	45,558	42,980	39,364	36,415	33,196
25	52,619	49,435	46,928	44,314	40,646	37,652	34,382
26	54,051	50,829	48,290	45,642	41,923	38,885	35,563
27	55,475	52,215	49,645	46,963	43,194	40,113	36,741
28	56,892	53,594	50,993	48,278	44,461	41,337	37,916
29	58,301	54,966	52,335	49,588	45,722	42,557	39,087
30	59,702	56,332	53,672	50,892	46,979	43,773	40,256

Приложение C

Приложение 3. Распределение Стьюдента

Пусть ξ – стандартная нормальная случайная величина, а χ^2_k – случайная величина, имеющая распределение хи-квадрат с k степенями свободы, причем они независимы. Тогда говорят, что случайная величина

$$\tau \;=\; \frac{\xi}{\sqrt{\chi^2_k/k}}$$

имеет распределение Стьюдента с k степенями свободы (T_k).

Распределение Стьюдента симметрично в том же смысле, что и стандартное нормальное, т.е. распределения τ и $-\tau$ совпадают. Отсюда получаем, что если T - функция распределения любого из распределений Стьюдента, то для произвольного x выполнено

$$T(x) \;+\; T(-x) \;=\; 1.$$

Ниже приводится таблица двусторонних критических точек распределения Стьюдента, т.е. решений уравнения

$$\mathbf{P}(|\tau| > t) \;=\; \varepsilon$$

для заданного ε. Из упомянутого выше свойства симметричности следует, что при необходимости найти одностороннюю критическую точку, т. е. решить уравнение

$$\mathbf{P}(\tau > t) \;=\; \delta$$

при заданном δ, следует в предыдущей формуле выбрать $\varepsilon = \delta/2$.

В силу закона больших чисел имеет место сходимость распределения T_k к стандартному нормальному распределению при $k \to \infty$, поэтому при больших k используйте таблицу стандартного нормального распределения.

Таблица составлена при помощи программы Microsoft Excell. Функции, использующиеся при работе с распределением Стьюдента, содержатся в разделе Статистические функции Мастера функций и носят названия СТЬЮДРАСП и СТЬЮДРАСПОБР (квантили). При этом обе функции содержат дополнительные возможности работы как с односторонними, так и с двусторонними критическими точками и квантилями (опция "хвосты").

Двусторонние критические точки распределения Стьюдента T_k

k	0,0005	0,0001	0,005	0,001	0,05	0,01	0,1
1	1273,155	6370,544	127,321	636,578	12,706	63,656	6,314
2	44,703	100,136	14,089	31,600	4,303	9,925	2,920
3	16,326	28,014	7,453	12,924	3,182	5,841	2,350
4	10,305	15,534	5,597	8,610	2,776	4,604	2,132
5	7,976	11,176	4,773	6,868	2,571	4,032	2,015
6	6,788	9,080	4,317	5,959	2,447	3,707	1,943
7	6,081	7,888	4,029	5,408	2,365	3,499	1,894
8	5,617	7,120	3,832	5,041	2,306	3,355	1,859
9	5,291	6,594	3,690	4,781	2,262	3,250	1,833
10	5,049	6,212	3,581	4,587	2,228	3,169	1,812
11	4,863	5,923	3,497	4,437	2,201	3,106	1,796
12	4,716	5,695	3,428	4,318	2,179	3,054	1,782
13	4,597	5,513	3,372	4,221	2,160	3,012	1,771
14	4,499	5,364	3,326	4,140	2,145	2,980	1,761
15	4,417	5,239	3,286	4,073	2,131	2,947	1,753
16	4,346	5,134	3,252	4,015	2,120	2,921	1,746
17	4,286	5,043	3,222	3,965	2,110	2,898	1,741
19	4,187	4,899	3,174	3,883	2,093	2,861	1,729
19	4,187	4,899	3,174	3,883	2,093	2,861	1,729
20	4,146	4,838	3,153	3,849	2,086	2,845	1,725
21	4,109	4,785	3,135	3,819	2,080	2,831	1,721
22	4,077	4,736	3,119	3,792	2,074	2,819	1,717
23	4,047	4,694	3,104	3,768	2,069	2,807	1,714
24	4,021	4,654	3,090	3,745	2,064	2,797	1,711
25	3,996	4,619	3,078	3,725	2,059	2,787	1,708
26	3,970	4,587	3,067	3,707	2,055	2,779	1,706
27	3,954	4,556	3,056	3,689	2,052	2,771	1,703
28	3,935	4,531	3,047	3,674	2,048	2,763	1,701
29	3,918	4,505	3,038	3,659	2,045	2,756	1,699
30	3,902	4,482	3,030	3,646	2,042	2,750	1,697
35	3,836	4,389	2,996	3,591	2,030	2,724	1,689
40	3,788	4,321	2,971	3,551	2,021	2,704	1,684
45	3,752	4,269	2,952	3,520	2,014	2,689	1,679
50	3,723	4,228	2,937	3,496	2,008	2,678	1,676

Приложение D

Приложение 4. Распределение Фишера

Если мы возьмем две независимых случайных величины – ξ, имеющую распределение χ^2_k и η, имеющую распределение χ^2_m, то распределение случайной величины

$$f \;=\; \frac{m\xi}{k\eta}$$

будет называться распределением Фишера с k, m степенями свободы и обозначаться $F_{k,m}$.

Как ясно из определения, случайная величина f принимает лишь положительные значения, и значит, ее распределение несимметрично. Однако, определенная симметрия все же имеется. Так, величина $1/f$ имеет распределение Фишера с m, k степенями свободы. Отсюда нетрудно получить, что для произвольного положительного t имеет место соотношение

$$\mathbf{P}\left(\frac{1}{f} \geq t\right) \;=\; \mathbf{P}\left(f < \frac{1}{t}\right) \;=\; 1 - \mathbf{P}\left(f \geq \frac{1}{t}\right),$$

а значит критическая точка распределения $F_{k,m}$ уровня ε – это то же самое, что критическая точка распределения $F_{m,k}$ уровня $1 - \varepsilon$.

В силу изложенных причин можно табулировать либо значения при малых и больших значениях ε, но ограничиться, например, случаем, когда число степеней свободы числителя меньше, чем число степеней свободы знаменателя, либо брать только малые значения уровня критической точки, но предусмотреть любые сочетания чисел степеней свободы. Ни-

же был принят второй подход. По столбцам расположены количества степеней свободы числителя (ξ), а по строкам – знаменателя (η).

Как уже было сказано в главе 1, таблицы распределения Фишера представляют собой трехвходовые таблицы, поэтому достаточно подробное их воспроизведение заняло бы слишком много места. Мы ограничились случаями $\varepsilon = 0,05$ (пятипроцентная критическая точка) и $\varepsilon = 0,01$ (однопроцентная критическая точка). В случае необходимости получить критические точки других уровней отсылаем читателя к электронной таблице Microsoft Excel, где вычисления, связанные с распределением Фишера, содержатся, как и в предыдущих приложениях, в разделе Статистические функции Мастера функций. Они называются FРАСП и FРАСПОБР.

Отметим,наконец, следующие связи между функцией распределения $F_{k,m}$ и функцией распределения Φ стандартного нормального распределения, представляющиеся интересными (см. [6]):

$$\lim_{m\to\infty} F_{1,m}(x) \;=\; 2\Phi(\sqrt{x}) - 1,$$

$$\lim_{k\to\infty} F_{k,1}(x) \;=\; 2(1 - \Phi(1/\sqrt{x})).$$

Критические точки распределения Фишера $F_{k,m}$ уровня $\varepsilon = 0{,}05$

(по горизонтали указаны k, по вертикали m)

	1	2	3	4	5
1	161,446	199,499	215,707	224,583	230,160
2	18,513	19,000	19,164	19,247	19,296
3	10,128	9,552	9,277	9,117	9,013
4	7,709	6,944	6,591	6,388	6,256
5	6,608	5,786	5,409	5,192	5,050
6	5,987	5,143	4,757	4,534	4,387
7	5,591	4,737	4,347	4,120	3,972
8	5,318	4,459	4,066	3,838	3,688
9	5,117	4,256	3,863	3,633	3,482
10	4,965	4,103	3,708	3,478	3,326
11	4,844	3,982	3,587	3,357	3,204
12	4,747	3,885	3,490	3,259	3,106
13	4,667	3,806	3,411	3,179	3,025
14	4,600	3,739	3,344	3,112	2,958
15	4,543	3,682	3,287	3,056	2,901
16	4,494	3,634	3,239	3,007	2,852
17	4,451	3,592	3,197	2,965	2,810
18	4,414	3,555	3,160	2,928	2,773
19	4,381	3,522	3,127	2,895	2,740
20	4,351	3,493	3,098	2,866	2,711
21	4,325	3,467	3,072	2,840	2,685
22	4,301	3,443	3,049	2,817	2,661
23	4,279	3,422	3,028	2,796	2,640
24	4,260	3,403	3,009	2,776	2,621
25	4,242	3,385	2,991	2,759	2,603
26	4,225	3,369	2,975	2,743	2,587
27	4,210	3,354	2,960	2,728	2,572
28	4,196	3,340	2,947	2,714	2,558
29	4,183	3,328	2,934	2,701	2,545
30	4,171	3,316	2,922	2,690	2,534
34	4,130	3,276	2,883	2,650	2,494
40	4,085	3,232	2,839	2,606	2,449
50	4,034	3,183	2,790	2,557	2,400
100	3,936	3,087	2,696	2,463	2,305

Критические точки распределения Фишера $F_{k,m}$ уровня $\varepsilon = 0{,}05$ – продолжение
(по горизонтали указаны k, по вертикали m)

	6	7	12	24	50
1	233,988	236,767	243,905	249,052	251,774
2	19,329	19,353	19,412	19,454	19,476
3	8,941	8,887	8,745	8,638	8,581
4	6,163	6,094	5,912	5,774	5,699
5	4,950	4,876	4,678	4,527	4,444
6	4,284	4,207	4,000	3,841	3,754
7	3,866	3,787	3,575	3,410	3,319
8	3,581	3,500	3,284	3,115	3,020
9	3,374	3,293	3,073	2,900	2,803
10	3,217	3,135	2,913	2,737	2,637
11	3,095	3,012	2,788	2,609	2,507
12	2,996	2,913	2,687	2,505	2,401
13	2,915	2,832	2,604	2,420	2,314
14	2,848	2,764	2,534	2,349	2,241
15	2,790	2,707	2,475	2,288	2,178
16	2,741	2,657	2,425	2,235	2,124
17	2,699	2,614	2,381	2,190	2,077
18	2,661	2,577	2,342	2,150	2,035
19	2,628	2,544	2,308	2,114	1,999
20	2,599	2,514	2,278	2,082	1,966
21	2,573	2,488	2,250	2,054	1,936
22	2,549	2,464	2,226	2,028	1,909
23	2,528	2,442	2,204	2,005	1,885
24	2,508	2,423	2,183	1,984	1,863
25	2,490	2,405	2,165	1,964	1,842
26	2,474	2,388	2,148	1,946	1,823
27	2,459	2,373	2,132	1,930	1,806
28	2,445	2,359	2,118	1,915	1,790
29	2,432	2,346	2,104	1,901	1,775
30	2,421	2,334	2,092	1,887	1,761
34	2,380	2,294	2,050	1,843	1,713
40	2,336	2,249	2,003	1,793	1,660
50	2,286	2,199	1,952	1,737	1,599
100	2,191	2,103	1,850	1,627	1,477

Критические точки распределения Фишера $F_{k,m}$ уровня $\varepsilon = 0{,}01$

(по горизонтали указаны k, по вертикали m)

	1	2	3	4	5
1	4052,185	4999,340	5403,534	5624,257	5763,955
2	98,502	99,000	99,164	99,251	99,302
3	34,116	30,816	29,457	28,710	28,237
4	21,198	18,000	16,694	15,977	15,522
5	16,258	13,274	12,060	11,392	10,967
6	13,745	10,925	9,780	9,148	8,746
7	12,246	9,547	8,451	7,847	7,460
8	11,259	8,649	7,591	7,006	6,632
9	10,562	8,022	6,992	6,422	6,057
10	10,044	7,559	6,552	5,994	5,636
11	9,646	7,206	6,217	5,668	5,316
12	9,330	6,927	5,953	5,412	5,064
13	9,074	6,701	5,739	5,205	4,862
14	8,862	6,515	5,564	5,035	4,695
15	8,683	6,359	5,417	4,893	4,556
16	8,531	6,226	5,292	4,773	4,437
17	8,400	6,112	5,185	4,669	4,336
18	8,285	6,013	5,092	4,579	4,248
19	8,185	5,926	5,010	4,500	4,171
20	8,096	5,849	4,938	4,431	4,103
21	8,017	5,780	4,874	4,369	4,042
22	7,945	5,719	4,817	4,313	3,988
23	7,881	5,664	4,765	4,264	3,939
24	7,823	5,614	4,718	4,218	3,895
25	7,770	5,568	4,675	4,177	3,855
26	7,721	5,526	4,637	4,140	3,818
27	7,677	5,488	4,601	4,106	3,785
28	7,636	5,453	4,568	4,074	3,754
29	7,598	5,420	4,538	4,045	3,725
30	7,562	5,390	4,510	4,018	3,699
34	7,444	5,289	4,416	3,927	3,611
40	7,314	5,178	4,313	3,828	3,514
50	7,171	5,057	4,199	3,720	3,408
100	6,895	4,824	3,984	3,513	3,206

Критические точки распределения Фишера $F_{k,m}$ уровня $\varepsilon = 0{,}01$ – продолжение
(по горизонтали указаны k, по вертикали m)

	6	7	12	24	50
1	5858,950	5928,334	6106,682	6234,273	6302,260
2	99,331	99,357	99,419	99,455	99,477
3	27,911	27,671	27,052	26,597	26,354
4	15,207	14,976	14,374	13,929	13,690
5	10,672	10,456	9,888	9,466	9,238
6	8,466	8,260	7,718	7,313	7,091
7	7,191	6,993	6,469	6,074	5,858
8	6,371	6,178	5,667	5,279	5,065
9	5,802	5,613	5,111	4,729	4,517
10	5,386	5,200	4,706	4,327	4,115
11	5,069	4,886	4,397	4,021	3,810
12	4,821	4,640	4,155	3,780	3,569
13	4,620	4,441	3,960	3,587	3,375
14	4,456	4,278	3,800	3,427	3,215
15	4,318	4,142	3,666	3,294	3,081
16	4,202	4,026	3,553	3,181	2,967
17	4,101	3,927	3,455	3,083	2,869
18	4,015	3,841	3,371	2,999	2,784
19	3,939	3,765	3,297	2,925	2,709
20	3,871	3,699	3,231	2,859	2,643
21	3,812	3,640	3,173	2,801	2,584
22	3,758	3,587	3,121	2,749	2,531
23	3,710	3,539	3,074	2,702	2,483
24	3,667	3,496	3,032	2,659	2,440
25	3,627	3,457	2,993	2,620	2,400
26	3,591	3,421	2,958	2,585	2,364
27	3,558	3,388	2,926	2,552	2,330
28	3,528	3,358	2,896	2,522	2,300
29	3,499	3,330	2,868	2,495	2,271
30	3,473	3,305	2,843	2,469	2,245
34	3,386	3,218	2,758	2,383	2,156
40	3,291	3,124	2,665	2,288	2,058
50	3,186	3,020	2,563	2,183	1,949
100	2,988	2,823	2,368	1,983	1,735

Литература

[1] П.Мюллер, П.Нойман, Р.Шторм - Таблицы по математической статистике. - М.:Финансы и статистика, 1987. - 278 с.

[2] С.А.Айвазян, В.М.Бухштабер, И.С.Енюков, Л.Д.Мешалкин Прикладная статистика: Классификация и снижение размерности. - М.:Финансы и статистика, 1989. - 607 с.

[3] В.С.Дронов - Основы математики (избранные главы). - Барнаул: Изд-во АГУ, 1998 - 95 с.

[4] А.А.Боровков Математическая статистика - М.:Наука, 1984. - 472 с.

[5] М.Кендалл, А.Стьюарт - Статистические выводы и связи - М.:Наука, 1973. - 900 с.

[6] Л.Н.Болшев, Н.В.Смирнов - Таблицы математической статистики - М.:Наука, 1983. - 416 с.

[7] Д.В.Гаскаров, В.И.Шаповалов - Малая выборка - М.: Статистика, 1978. - 248 с.

[8] С.А.Айвазян, И.С.Енюков, Л.Д.Мешалкин - Прикладная статистика: Исследование зависимостей. - М.: Финансы и статистика, 1985. - 487 с.

[9] Г.Шеффе - Дисперсионный анализ. - М.:Физматгиз, 1963. - 626 с.

[10] М.Кендалл, А.Стьюарт - Многомерный статистический анализ и временные ряды. - М.:Наука, 1976, - 736 с.

[11] С.А.Айвазян, И.С.Енюков, Л.Д.Мешалкин - Прикладная статистика: Основы моделирования и первичная обработка данных. - М.: Финансы и статистика, 1983.- 472 с.

[12] Т.Андерсон - Статистический анализ временных рядов. - М.:Мир, 1976. - 755 с.

[13] Б.Болч, К.Дж.Хуань - Многомерные статистические методы для экономики. - М.:Статистика, 1979.- 379 с.

[14] М.Дейвисон - Многомерное шкалирование: методы наглядного представления данных. - М.:Финансы и статистика, 1988.- 254 с.

Оглавление

1 Предварительные сведения **3**
1.1 Анализ и алгебра 4
1.2 Теория вероятностей 8
1.3 Математическая статистика 10

2 Многомерные распределения **13**
2.1 Случайные векторы 13
2.2 Независимость 16
2.3 Числовые характеристики 18
2.4 Нормальное распределение в многомерном случае 20
2.5 Корреляционная теория 23

3 Группировка и цензурирование **29**
3.1 Одномерная группировка 29
3.2 Одномерное цензурирование 31
3.3 Таблицы сопряженности 32
3.3.1 Гипотеза независимости 33
3.3.2 Гипотеза однородности 34
3.3.3 Поле корреляции 35
3.4 Многомерная группировка 36
3.5 Многомерное цензурирование 37

4 Нечисловые данные **39**
4.1 Вводные замечания 39
4.2 Шкалы сравнений 40
4.3 Экспертные оценки 42
4.4 Группы экспертов 43

5 Доверительные множества **49**
5.1 Доверительные интервалы 49
5.2 Доверительные множества 52
5.2.1 Многомерный параметр 52
5.2.2 Многомерная выборка 53
5.3 Толерантные множества 54
5.4 Малая выборка 58

6 Регрессионный анализ **59**
6.1 Постановка задачи 59
6.2 Поиск ОМНК 62
6.3 Ограничения 65
6.4 Матрица плана 67
6.5 Статистический прогноз 69

7 Дисперсионный анализ **73**
7.1 Вводные замечания 73
7.1.1 Нормальность 74
7.1.2 Однородность дисперсий 76
7.2 Один фактор 77
7.3 Два фактора 81
7.4 Общий случай 85

8 Снижение размерности **87**
8.1 Зачем нужна классификация 87
8.2 Модель и примеры 89
8.2.1 Метод главных компонент 90
8.2.2 Экстремальная группировка признаков 90
8.2.3 Многомерное шкалирование 91
8.2.4 Отбор показателей для дискриминантного анализа . 92
8.2.5 Отбор показателей в модели регрессии 93

9 Дискриминантный анализ **95**
9.1 Применимость модели 95
9.2 Линейное прогностическое правило 100
9.3 Практические рекомендации 103
9.4 Один пример 105
9.5 Более двух классов 109

9.6 Проверка качества дискриминации 111

10 Эвристические методы **113**
10.1 Экстремальная группировка 113
10.1.1 Критерий квадратов 114
10.1.2 Критерий модулей 117
10.2 Метод плеяд 119

11 Метод главных компонент **121**
11.1 Постановка задачи 121
11.2 Вычисление главных компонент 124
11.3 Пример 127
11.4 Свойства главных компонент 130
11.4.1 Самовоспроизводимость 130
11.4.2 Геометрические свойства 132

12 Факторный анализ **135**
12.1 Постановка задачи 135
12.1.1 Связь с главными компонентами 135
12.1.2 Однозначность решения 137
12.2 Математическая модель 139
12.2.1 Условия на A^tA 141
12.2.2 Условия на матрицу нагрузок. Центроидный метод . 142
12.3 Латентные факторы 144
12.3.1 Метод Бартлетта 145
12.3.2 Метод Томсона 145
12.4 Пример 146

13 Оцифровка **153**
13.1 Анализ соответствий 153
13.1.1 Расстояние хи-квадрат 155
13.1.2 Оцифровка для задач дискриминантного анализа . . 159
13.2 Более двух переменных 162
13.2.1 Использование бинарной матрицы данных в качестве матрицы соответствий 163
13.2.2 Максимальные корреляции 165
13.3 Размерность 167
13.4 Пример 169

13.5 Случай смешанных данных 174

14 Многомерное шкалирование **179**
14.1 Вводные замечания 180
14.2 Модель Торгерсона 182
14.2.1 Стресс-критерий 185
14.3 Алгоритм Торгерсона 186
14.4 Индивидуальные различия 189

15 Временные ряды **193**
15.1 Общие положения 193
15.2 Критерии случайности 195
15.2.1 Пики и ямы 196
15.2.2 Распределение длины фазы 199
15.2.3 Критерии, основанные на ранговой корреляции . . . 203
15.2.4 Коррелограмма 207
15.3 Тренд и сезонность 208
15.3.1 Полиномиальные тренды 208
15.3.2 Выбор степени тренда 211
15.3.3 Сглаживание 213
15.3.4 Оценка сезонных колебаний 215

A Нормальное распределение **219**

B Распределение χ^2 **225**

C Распределение Стьюдента **231**

D Распределение Фишера **235**